Summary of References Related to Convolutional Neural Networks

Miquel Perello, E-mail: miquel.perellonieto@aalto.fi

February 9, 2015

Contents

1 Introduction

2 Missing year

3 1700

3.1 An essay concerning human understanding [159]

3.1.1 Original Abstract

Many of the earliest books, particularly those dating back to the 1900s and before, are now extremely scarce and increasingly expensive. Pomona Press are republishing these classic works in affordable, high quality, modern editions, using the original text and artwork.

4 1749

4.1 Observations on man, his frame, his duty, and his expectations [95]

4.1.1 Original Abstract

None

5 1873

5.1 Mind and body. The theories of their relation [16]

5.1.1 Original Abstract

None

5.1.2 Main points

6 1890

6.1 The principles of psychology [264]

6.1.1 Original Abstract

vol. 1

7 1909

7.1 Histologie du systeme nerveux de l'homme & des vertebres [34]

7.1.1 Original Abstract

Translation of Textura del sistema nervioso del hombre y de los vertebrados; Microfilmed for preservation; t. 1. Généralités, moelle, ganglions rachidiens, bulbe
protubérance.– t. 2. Cervelet, cerveau moyen, rétine, couche optique, corps strié, écorce cérébrale générale
régionale, grand sympathique

7.1.2 Main points

8 1921

8.1 A history of the association psychology [109]

8.1.1 Original Abstract

None

9 1943

9.1 A logical calculus of the ideas immanent in nervous activity [166]

9.1.1 Original Abstract

Because of the "all-or-none" character of nervous activity, neural events and the relations among them can be treated by means of propositional logic. It is found that the behavior of every net can be described in these terms, with the addition of more complicated logical means for nets containing circles; and that for any logical expression satisfying certain conditions, one can find a net behaving in the fashion it describes. It is shown that many particular choices among possible neurophysiological assumptions are equivalent, in the sense that for every net behaving under one assumption, there exists another net which behaves under the other and gives the same results, although perhaps not in the same time. Various applications of the calculus are discussed.

10 1945

10.1 First Draft of a Report on the EDVAC [254]

10.1.1 Original Abstract

None

11 1947

11.1 On a test of whether one of two random variables is stochastically larger than the other [162]

11.1.1 Original Abstract

Let xx and yy be two random variables with continuous cumulative distribution functions ff and gg. A statistic UU depending on the relative ranks of the xx's and yy's is proposed for testing the hypothesis f=gf = g. Wilcoxon proposed an equivalent test in the Biometrics Bulletin, December, 1945, but

gave only a few points of the distribution of his statistic. Under the hypothesis f=gf = g the probability of obtaining a given UU in a sample of nx′sn x's and my′sm y's is the solution of a certain recurrence relation involving nn and mm. Using this recurrence relation tables have been computed giving the probability of UU for samples up to n=m=8n = m = 8. At this point the distribution is almost normal. From the recurrence relation explicit expressions for the mean, variance, and fourth moment are obtained. The 2rth moment is shown to have a certain form which enabled us to prove that the limit distribution is normal if m,nm, n go to infinity in any arbitrary manner. The test is shown to be consistent with respect to the class of alternatives f(x)>g(x)f(x) > g(x) for every xx.

11.1.2 Main points

12 1948

12.1 Cybernetics or Control and Communication in the Animal and the Machine [262]

12.1.1 Original Abstract

"It appers impossible for anyone seriously interested in our civilization to ignore this book. It is a 'must' book forthose in every branch of science... in addition, economists, politicians, statesmen, and businessmen cannot afford to overlook cybernetics and its tremendous, even terrifying implications. "It is a beautifully written book, lucid, direct, and despite its complexity, as readable by the layman as the trained scientist." – John B. Thurston

13 1949

13.1 The Orgamization of Behavior a Neuropsychological Theory [98]

13.1.1 Original Abstract

None

14 1953

14.1 Equation of State Calculations by Fast Computing Machines [169]

14.1.1 Original Abstract

A general method, suitable for fast computing machines, for investigating such properties as equations of state for substances consisting of interacting individual molecules is described. The method consists of a modified Monte Carlo integration over configuration space. Results for the two [U+2010] dimensional rigid [U+2010] sphere system have been obtained on the Los Alamos MANIAC and are presented here. These results are compared to the free volume equation of state and to a four [U+2010] term virial coefficient expansion.

15 1954

15.1 Theory of neural-analog reinforcement systems and its application to the brain model problem [175]

15.1.1 Original Abstract

None

15.2 Communication theory and cybernetics [76]

15.2.1 Original Abstract

In the case of band-limited signals, the sampling theorem permits us to replace analytic operations with algebraic operations. We are then able to discuss problems of measurement of information, coding, and transmission over noisy channels in terms of discrete samples, rather than continuous time functions. The design of the optimum linear filter reduces from a very difficult analysis problem involving spec- trum factorizatlon to a straightforward problem of solving a set of simultaneous linear equations. Unless we are interested in the most economical implementation, it is not even necessary to solve the equations. since a synthesis procedure involving only simple functions of the correlation functions is available. When extended to the general nonlinear case, the design is still specified by a set of simultaneous algebraic

equations, but the labor of solution grows very rapidly. It is proposed to short circuit this labor by building a learning filter which in effect designs itself. A training period in which the adjust- ments are automatically optimized precedes the use period. By modi- fying the training program, it is possible that the filter could be taught to recognize specific signals, including, perhaps, certain speech sounds.

15.3 Simulation of self-organizing systems by digital computer [63]

15.3.1 Original Abstract

A general discussion of ideas and definitions relating to self-organizing systems and their synthesis is given, together with remarks concerning their simulation by digital computer. Synthesis and simulation of an actual system is then described. This system, initially randomly organized within wide limits, organizes itself to perform a simple prescribed task.

16 1955

16.1 Memory: The Analogy with Ferromagnetic Hysteresis [46]

16.1.1 Original Abstract

None

17 1956

17.1 Electrical simulation of some nervous system functional activities. [244]

17.1.1 Original Abstract

None

17.2 Temporal and spatial patterns in a conditional probability machine [249]

17.2.1 Original Abstract

None

17.3 Conditional probability machines and conditional reflexes [248]

17.3.1 Original Abstract

None

17.4 Probabilistic logics and the synthesis of reliable organisms from unreliable components [255]

17.4.1 Original Abstract

None

17.5 Tests on a cell assembly theory of the action of the brain, using a large digital computer [207]

17.5.1 Original Abstract

Theories by D.O. Hebb and P.M. Milner on how the brain works were tested by simulating neuron nets on the IBM Type 704 Electronic Calculator. The formation of cell assemblies from an unorganized net of neurons was demonstrated, as well as a plausible mechanism for short-term memory and the phenomena of growth and fractionation of cell assemblies. The cell assemblies do not yet act just as the theory requires, but changes in the theory and the simulation offer promise for further experimentation.

18 1957

18.1 The Perceptron, a Perceiving and Recognizing Automaton [210]

18.1.1 Original Abstract

None

18.1.2 Main points

19 1958

19.1 The perceptron: a probabilistic model for information storage and organization in the brain. [208]

19.1.1 Original Abstract

To answer the questions of how information about the physical world is sensed, in what form is information remembered, and how does information retained in memory influence recognition and behavior, a theory is developed for a hypothetical nervous system called a perceptron. The theory serves as a bridge between biophysics and psychology. It is possible to predict learning curves from neurological variables and vice versa. The quantitative statistical approach is fruitful in the understanding of the organization of cognitive systems. 18 references.

19.1.2 Main points

19.2 The perceptron: a probabilistic model for information storage and organization in the brain. [208]

19.2.1 Original Abstract

To answer the questions of how information about the physical world is sensed, in what form is information remembered, and how does information retained in memory influence recognition and behavior, a theory is developed for a hypothetical nervous system called a perceptron. The theory serves as a bridge between biophysics and psychology. It is possible to predict learning

curves from neurological variables and vice versa. The quantitative statistical approach is fruitful in the understanding of the organization of cognitive systems. 18 references.

19.2.2 Main points

19.3 The perceptron: a probabilistic model for information storage and organization in the brain. [208]

19.3.1 Original Abstract

To answer the questions of how information about the physical world is sensed, in what form is information remembered, and how does information retained in memory influence recognition and behavior, a theory is developed for a hypothetical nervous system called a perceptron. The theory serves as a bridge between biophysics and psychology. It is possible to predict learning curves from neurological variables and vice versa. The quantitative statistical approach is fruitful in the understanding of the organization of cognitive systems. 18 references.

20 1960

20.1 An Adaptive "ADALINE" Neuron Using Chemical "Memistors" [261]

20.1.1 Original Abstract

None

20.2 Design for a Brain: The Origin of Adaptive Behavior [13]

20.2.1 Original Abstract

None

20.2.2 Main points

21 1961

21.1 Principles of neurodynamics. perceptrons and the theory of brain mechanisms [209]

21.1.1 Original Abstract

Part I attempts to review the background, basic sources of data, concepts, and methodology to be employed in the study of perceptrons. In Chapter 2, a brief review of the main alternative approaches to the development of brain models is presented. Chapter 3 considers the physiological and psychological criteria for a suitable model, and attempts to evaluate the empirical evidence which is available on several important issues. Chapter 4 contains basic definitions and some of the notation to be used in later sections are presented. Parts II and III are devoted to a summary of the established theoretical results obtained to date. Part II (Chapters 5 through 14) deals with the theory of three-layer series-coupled perceptrons, on which most work has been done to date. Part III (Chapters 15 through 20) deals with the theory of multi-layer and cross-coupled perceptrons. Part IV is concerned with more speculative models and problems for future analysis. Of necessity, the final chapters become increasingly heuristic in character, as the theory of perceptrons is not yet complete, and new possibilities are continually coming to light.

22 1962

22.1 On convergence proofs on perceptrons [190]

22.1.1 Original Abstract

None

22.2 Receptive fields, binocular interaction and functional architecture in the cat's visual cortex [112]

22.2.1 Original Abstract

What chiefly distinguishes cerebral cortex from other parts of the central nervous system is the great diversity of its cell types and interconnexions. It would be astonishing if such a structure did not profoundly modify the response patterns of fibres coming into it. In the cat's visual cortex, the receptive field arrangements of single cells suggest that there is indeed a degree of complexity far exceeding anything yet seen at lower levels in the visual system. In a previous paper we described receptive fields of single cortical cells, observing responses to spots of light shone on one or both retinas (Hubel Wiesel, 1959). In the present work this method is used to examine receptive fields of a more complex type (Part I) and to make additional observations on binocular interaction (Part II). This approach is necessary in order to understand the behaviour of individual cells, but it fails to deal with the problem of the relationship of one cell to its neighbours. In the past, the technique of recording evoked slow waves has been used with great success in studies of individual cells, but it fails to deal with the problem of the relationship of one cell to its neighbours. In the past, the technique of recording evoked slow waves has been used with great success in studies of functional anatomy. It was employed by Talbot Marshall (1941) and by Thompson, Woolsey Talbot (1950) for mapping out the visual cortex in the rabbit, cat, and monkey. Daniel Whitteiidge (1959) have recently extended this work in the primate. Most of our present knowledge of retinotopic projections, binocular overlap, and the second visual area is based on these investigations. Yet the method of evoked potentials is valuable mainly for detecting behaviour common to large populations of neighbouring cells; it cannot differentiate functionally between areas of cortex smaller than about 1 mm 2. To overcome this difficulty a method has in recent years been developed for studying cells separately or in small groups during long micro-electrode penetrations through nervous tissue. Responses are correlated with cell location by reconstructing the electrode tracks from histological material. These techniques have been applied to the somatic sensory cortex of the cat and monkey in a remarkable series of studies by Mountcastle (1957) and Powell

Mountcastle (1959). Their results show that the approach is a powerful one, capable of revealing systems of organization not hinted at by the known morphology. In Part III of the present paper we use this method in studying the functional architecture of the visual cortex. It helped us attempt to explain on anatomical grounds how cortical receptive fields are built up.

23 1965

23.1 Learning machines: foundations of trainable pattern-classifying systems [186]

23.1.1 Original Abstract

None

24 1966

24.1 Theory of self-reproducing automata [183]

24.1.1 Original Abstract

None

25 1967

25.1 A Theory of Adaptive Pattern Classifiers [7]

25.1.1 Original Abstract

This paper describes error-correction adjustment procedures for determining the weight vector of linear pattern classifiers under general pattern distribution. It is mainly aimed at clarifying theoretically the performance of adaptive pattern classifiers. In the case where the loss depends on the distance between a pattern vector and a decision boundary and where the average risk function is unimodal, it is proved that, by the procedures proposed here, the weight vector converges to the optimal one even under nonseparable pattern distributions. The speed and the accuracy of convergence are analyzed, and it is

shown that there is an important tradeoff between speed and accuracy of convergence. Dynamical behaviors, when the probability distributions of patterns are changing, are also shown. The theory is generalized and made applicable to the case with general discriminant functions, including piecewise-linear discriminant functions.

25.1.2 Main points

26 1968

26.1 Receptive fields and functional architecture of monkey striate cortex [113]

26.1.1 Original Abstract

1. The striate cortex was studied in lightly anaesthetized macaque and spider monkeys by recording extracellularly from single units and stimulating the retinas with spots or patterns of light. Most cells can be categorized as simple, complex, or hypercomplex, with response properties very similar to those previously described in the cat. On the average, however, receptive fields are smaller, and there is a greater sensitivity to changes in stimulus orientation. A small proportion of the cells are colour coded.2. Evidence is presented for at least two independent systems of columns extending vertically from surface to white matter. Columns of the first type contain cells with common receptive-field orientations. They are similar to the orientation columns described in the cat, but are probably smaller in cross-sectional area. In the second system cells are aggregated into columns according to eye preference. The ocular dominance columns are larger than the orientation columns, and the two sets of boundaries seem to be independent.3. There is a tendency for cells to be grouped according to symmetry of responses to movement; in some regions the cells respond equally well to the two opposite directions of movement of a line, but other regions contain a mixture of cells favouring one direction and cells favouring the other.4. A horizontal organization corresponding to the cortical layering can also be discerned. The upper layers (II and the upper two-thirds of III) contain complex and hypercomplex cells, but simple cells are virtually absent. The cells are mostly binocularly driven. Simple cells are found deep in layer III, and in IV A and IV B. In layer IV B they form a large proportion of the population, whereas complex cells are rare. In layers

IV A and IV B one finds units lacking orientation specificity; it is not clear whether these are cell bodies or axons of geniculate cells. In layer IV most cells are driven by one eye only; this layer consists of a mosaic with cells of some regions responding to one eye only, those of other regions responding to the other eye. Layers V and VI contain mostly complex and hypercomplex cells, binocularly driven.5. The cortex is seen as a system organized vertically and horizontally in entirely different ways. In the vertical system (in which cells lying along a vertical line in the cortex have common features) stimulus dimensions such as retinal position, line orientation, ocular dominance, and perhaps directionality of movement, are mapped in sets of superimposed but independent mosaics. The horizontal system segregates cells in layers by hierarchical orders, the lowest orders (simple cells monocularly driven) located in and near layer IV, the higher orders in the upper and lower layers.

26.1.2 Main points

27 1969

27.1 Perceptrons [174]

27.1.1 Original Abstract

Perceptrons: an introduction to computational geometry is a book written by Marvin Minsky and Seymour Papert and published in 1969. An edition with handwritten corrections and additions was released in the early 1970s. An expanded edition was further published in 1987, containing a chapter dedicated to counter the criticisms made of it in the 1980s.

27.2 Non-Holographic Associative Memory [266]

27.2.1 Original Abstract

The features of a hologram that commend it as a model of associative memory can be improved on by other devices.

27.3 Non-Holographic Associative Memory [267]

27.3.1 Original Abstract

The features of a hologram that commend it as a model of associative memory can be improved on by other devices.

28 1971

28.1 On the uniform convergence of relative frequencies of events to their probabilities [250]

28.1.1 Original Abstract

None

29 1972

29.1 A simple neural network generating an interactive memory [8]

29.1.1 Original Abstract

A model of a neural system where a group of neurons projects to another group of neurons is discussed. We assume that a trace is the simultaneous pattern of individual activities shown by a group of neurons. We assume synaptic interactions add linearly and that synaptic weights (quantitative measure of degree of coupling between two cells) can be coded in a simple but optimal way where changes in synaptic weight are proportional to the product of pre-and postsynaptic activity at a given time. Then it is shown that this simple system is capable of "memory" in the sense that it can (1) recognize a previously presented trace and (2) if two traces have been associated in the past (that is, if trace f^- was impressed on the first group of neurons and trace $\bar{g}$ was impressed on the second group of neurons and synaptic weights coupling the two groups changed according to the above rule) presentation of f^- to the first group of neurons gives rise to f^- plus a calculable amount of noise at the second set of neurons. This kind of memory is called an "interactive memory" since distinct stored traces interact in storage. It is shown that this model can

effectively perform many functions. Quantitative expressions are derived for the average signal to noise ratio for recognition and one type of association. The selectivity of the system is discussed. References to physiological data are made where appropriate. A sketch of a model of mammalian cerebral cortex which generates an interactive memory is presented and briefly discussed. We identify a trace with the activity of groups of cortical pyramidal cells. Then it is argued that certain plausible assumptions about the properties of the synapses coupling groups of pyramidal cells lead to the generation of an interactive memory.

29.1.2 Main points

29.2 Characteristics of Random Nets of Analog Neuron-Like Elements [6]

29.2.1 Original Abstract

The dynamic behavior of randomly connected analog neuron-like elements that process pulse-frequency modulated signals is investigated from the macroscopic point of view. By extracting two statistical parameters, the macroscopic state equations are derived in terms of these parameters under some hypotheses on the stochastics of microscopic states. It is shown that a random net of statistically symmetric structure is monostable or bistable, and the stability criteria are explicitly given. Random nets consisting of many different classes of elements are also analyzed. Special attention is paid to nets of randomly connected excitatory and inhibitory elements. It is shown that a stable oscillation exists in such a netÂ¿in contrast with the fact that no stable oscillations exist in a net of statistically symmetric structure even if negative as well as positive synaptic weights are permitted at a time. The results are checked by computer-simulated experiments.

29.2.2 Main points

29.3 Correlation matrix memories [132]

29.3.1 Original Abstract

A new model for associative memory, based on a correlation matrix, is suggested. In this model information is accumulated on memory elements as

products of component data. Denoting a key vector by q(p), and the data associated with it by another vector x(p), the pairs (q(p), x(p)) are memorized in the form of a matrix see the Equation in PDF File where c is a constant. A randomly selected subset of the elements of Mxq can also be used for memorizing. The recalling of a particular datum x(r) is made by a transformation x(r)=Mxqq(r). This model is failure tolerant and facilitates associative search of information; these are properties that are usually assigned to holographic memories. Two classes of memories are discussed: a complete correlation matrix memory (CCMM), and randomly organized incomplete correlation matrix memories (ICMM). The data recalled from the latter are stochastic variables but the fidelity of recall is shown to have a deterministic limit if the number of memory elements grows without limits. A special case of correlation matrix memories is the auto-associative memory in which any part of the memorized information can be used as a key. The memories are selective with respect to accumulated data. The ICMM exhibits adaptive improvement under certain circumstances. It is also suggested that correlation matrix memories could be applied for the classification of data.

29.4 Automata Studies: Annals of Mathematics Studies. Number 34 [223]

29.4.1 Original Abstract

None

30 1973

30.1 Self-organization of orientation sensitive cells in the striate cortex [252]

30.1.1 Original Abstract

A nerve net model for the visual cortex of higher vertebrates is presented. A simple learning procedure is shown to be sufficient for the organization of some essential functional properties of single units. The rather special assumptions usually made in the literature regarding preorganization of the visual cortex are thereby avoided. The model consists of 338 neurones forming a sheet analogous to the cortex. The neurones are connected randomly to a

"retina" of 19 cells. Nine different stimuli in the form of light bars were applied. The afferent connections were modified according to a mechanism of synaptic training. After twenty presentations of all the stimuli individual cortical neurones became sensitive to only one orientation. Neurones with the same or similar orientation sensitivity tended to appear in clusters, which are analogous to cortical columns. The system was shown to be insensitive to a background of disturbing input excitations during learning. After learning it was able to repair small defects introduced into the wiring and was relatively insensitive to stimuli not used during training.

31 1974

31.1 Beyond regression: new tools for prediction and analysis in the behavioral sciences [259]

31.1.1 Original Abstract

None

32 1975

32.1 A statistical theory of short and long term memory [157]

32.1.1 Original Abstract

We present a theory of short, intermediate and long term memory of a neural network incorporating the known statistical nature of chemical transmission at the synapses. Correlated pre- and post-synaptic facilitation (related to Hebb's Hypothesis) on three time scales are crucial to the model. Considerable facilitation is needed on a short time scale both for establishing short term memory (active persistent firing pattern for the order of a sec) and the recall of intermediate and long term memory (latent capability for a pattern to be re-excited). Longer lasting residual facilitation and plastic changes (of the same nature as the short term changes) provide the mechanism for imprinting of the intermediate and long term memory. We discuss several interesting

features of our theory: nonlocal memory storage, large storage capacity, access of memory, single memory mechanism, robustness of the network and statistical reliability, and usefulness of statistical fluctuations.

33 1976

33.1 A mechanism for producing continuous neural mappings: ocularity dominance stripes and ordered retino-tectal projections [253]

33.1.1 Original Abstract

None

33.2 Luminance and opponent-color contributions to visual detection and adaptation and to temporal and spatial integration [130]

33.2.1 Original Abstract

We show how the processes of visual detection and of temporal and spatial summation may be analyzed in terms of parallel luminance (achromatic) and opponent-color systems; a test flash is detected if it exceeds the threshold of either system. The spectral sensitivity of the luminance system may be determined by a flicker method, and has a single broad peak near 555 nm; the spectral sensitivity of the opponent-color system corresponds to the color recognition threshold, and has three peaks at about 440, 530, and 600 nm (on a white background). The temporal and spatial integration of the opponent-color system are generally greater than for the luminance system; further, a white background selectively depresses the sensitivity of the luminance system relative to the opponent-color system. Thus relatively large (1°) and long (200 msec) spectral test flashes on a white background are detected by the opponent-color system except near 570 nm; the contribution of the luminance system becomes more prominent if the size or duration of the test flash is reduced, or if the white background is extinguished. The present analysis is discussed in relation to Stiles' model of independent mechanisms.

34 1980

34.1 Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position [74]

34.1.1 Original Abstract

A neural network model for a mechanism of visual pattern recognition is proposed in this paper. The network is self-organized by "learning without a teacher", and acquires an ability to recognize stimulus patterns based on the geometrical similarity (Gestalt) of their shapes without affected by their positions. This network is given a nickname "neocognitron". After completion of self-organization, the network has a structure similar to the hierarchy model of the visual nervous system proposed by Hubel and Wiesel. The network consists of an input layer (photoreceptor array) followed by a cascade connection of a number of modular structures, each of which is composed of two layers of cells connected in a cascade. The first layer of each module consists of "S-cells', which show charac- teristics similar to simple cells or lower order hyper- complex cells, and the second layer consists of "C-cells" similar to complex cells or higher order hypercomplex cells. The afferent synapses to each S-cell have plasticity and are modifiable. The network has an ability of unsupervised learning: We do not need any "teacher" during the process of self- organization, and it is only needed to present a set of stimulus patterns repeatedly to the input layer of the network. The network has been simulated on a digital computer. After repetitive presentation of a set of stimulus patterns, each stimulus pattern has become to elicit an output only from one of the C-cells of the last layer, and conversely, this C-cell has become selectively responsive only to that stimulus pattern. That is, none of the C-cells of the last layer responds to more than one stimulus pattern. The response of the C-cells of the last layer is not affected by the pattern's position at all. Neither is it affected by a small change in shape nor in size of the stimulus pattern. 1.

34.1.2 Main points

- Reiteration of self-organized by "learning without a teacher"
- Similar structure to the hierarchy model of the visual nervous system

proposed by Hubel and Wiesel.

- Network structure:
  - Input layer (photoreceptor array)
  - Cascade of modules each one with :
    - S-cells: in the first layer Simple cells or lower order hypercomplex cells
    - C-cells: in the second layer Complex cells or higher order hypercomplex cells
- Hubel and Wiesel : the neural network in the visual cortex has a hierarchy structure:
  - LGB (Lageral Geniculate Body)
  - Simple cells
  - Complex cells
  - Lower order hypercomplex cells
  - Higher order hypercomplex cells
- a cell in a higher stage generally has tendency to respond selectively to a more complicated feature of the stimulus pattern
- we extend the hierarchy model of Hubel and Wiesel, and **hypothesize** the existance of a similar hierarchy structure even in hte stages higher than hypercomplex cells.
- In the last module, the receptive field of each C-cell becomes so large as to cover the whole area of input layer U_0, and each C-plane is so determined as to have only one C-cell
- The output of an S-cell in the k_l-th S-plane in the l-th module is described below

35 1981

35.1 A. I. [24]

35.1.1 Original Abstract

ABSTRACT: PROFILE of Marvin Minsky, professor at MIT, who works in artificial intelligence. He was born in N.Y. in 1927. The term "artificial intelligence" is usually attributed to John McCarthy, a former colleague of Minsky's. He coined the phrase in the mid-fifties to describe the ability of certain machines to do things that people call intelligent. In 1958 McCarthy & Minsky created the Artificial Intelligence Group at MIT, & it soon became one of the most distinguished scientific enterprises in the world. Today about a hundred people work in the lab & it gets some $2.5 million a year from various government agencies. In 1968 $ *arithmetic problems N in short, to make them intelligent. At present, debate rages about what* ı *artificia* $ *thus the attempts to make machines that play games,* ı *understand* ȷ *newspaper accounts, &* $ *can recogniz*

35.1.2 Main points

p69 For a while, I studied topology, and then I ran into a young graduate student in physics named Dean Edmonds, who was a whiz at electronics. We began to build vacuum-tube circuits that did all sorts of things.quot; As an undergraduate, Minsky had begun to imagine building an electronic machine that could learn. He had become fascinated by a paper that had been written, in 1943, by Warren S. McCulloch, a neurophysiologist, and Walter Pitts, a mathematical prodigy. In this paper, McCulloch and Pitts created an abstract model of the brain cells—the neurons—and showed how they might be connected to carry out mental processes such as learning. Minsky now thought that the time might be ripe to try to create such a machine. "I told Edmonds that I thought it might be too hard to build,quot; he said. "The one I then envisioned would have needed a lot of memory circuits. There would be electronic neurons connected by synapses that would determine when the neurons fired. The synapses would have various probabilities for conducting. But to reinforce 'success' one would have to have a way of changing these probabilities. There would have to be loops and cycles in the circuits so that the machine could remember traces of its past and adjust its behavior. I thought that if I could ever build such a machine I might get it to learn to run mazes through its electronics— like rats or something. I didn't think that it would be very intelligent. I thought it would work pretty well with about forty

neurons. Edmonds and I worked out some circuits so that —in principle, at least—we could realize each of these neurons with just six vacuum tubes and a motor.quot; Minsky told George Miller, at Harvard, about the prospective design. "He said, 'Why don't we just try it?' " Minsky recalled. "He had a lot of faith in me, which I appreciated. Somehow, he managed to get a couple of thousand dollars from the Office of Naval Research, and in the summer of 1951 Dean Edmonds and I went up to Harvard and built our machine. It had three hundred tubes and a lot of motors. It needed some automatic electric clutches, which we machined ourselves. The memory of the machine as stored in the positions of its control knobs—forty of them—and when the machine was learning it used the clutches to adjust its own knobs. We used a surplus gyropilot from a B-24 bomber to move the clutches.quot; Minsky's machine was certainly one of the first electronic learning machines, and perhaps the very first one. In addition to its neurons and synapses and its internal memory loops, many of the networks were wired at random, so that it was impossible to predict what it would do. A "ratquot; would be created at some point in the network and would then set out to learn a path to some specified end point. First, it would proceed randomly, and then correct choices would be reinforced by making it easier for the machine to make this choice again—to increase the probability of its doing so. There was an arrangement of lights that allowed observers to follow the progress of the rat—or rats. "It turned out that because of an electronic accident in our design we could put two or three rats in the same maze and follow them all,quot; Minsky told me. "The rats actually interacted with one another. If one of them found a good path, the others would tend to follow it. We sort of quit science for a while to watch the machine. We were amazed that it could have several activities going on at once in its little nervous system. Because of the random wiring, it had a sort of fail-safe characteristic. If one of the neurons wasn't working, it wouldn't make much of a difference —and, with nearly three hundred tubes and the thousands of connections we had soldered, there would usually be something wrong somewhere. In those days, even a radio set with twenty tubes tended to fail a lot. I don't think we ever debugged our machine completely, but that didn't matter. By having this crazy random design, it was almost sure to work, no matter how you built it.quot; Minsky went on, "My Harvard machine was basically Skinnerian, although Skinner, with whom I talked a great deal while I was building it, was never much interested in it. The unrewarded behavior of my machine was more or less random. This limited its learning ability. It could never

formulate a plan. The next idea I had, which I worked on for my doctoral thesis, was to give the network a second memory, which remembered after a response what the stimulus had been. This enabled one to bring in the idea of prediction. If the machine or animal is confronted with a new situation, it can search its memory to see what would happen if it reacted in certain ways. If, say, there was an unpleasant association with a certain stimulus, then the machine could choose a different response. I had the naive idea that if one could build a big enough network, with enough memory loops, it might get lucky and acquire the ability to envision things in its head. This became a field of study later. It was called self-organizing random networks. Even today, I still get letters from young students who say, 'Why are you people trying to program intelligence? Why don't you try to find a way to build a nervous system that will just spontaneously create it?' Finally, I decided that either this was a bad idea or it would take thousands or millions of neurons to make it work, and I couldn't afford to try to build a machine like that.quot; I asked Minsky why it had not occurred to him to use a computer to simulate his machine. By this time, the first electronic digital computer— named ENIAC, for "electronic numerical integrator and calculatorquot;—had been built, at the University of Pennsylvania's Moore School of Electrical Engineering; and the mathematician John von Neumann was completing work on a computer, the prototype of many present-day computers, at the Institute for Advanced Study. "I knew a little bit about computers,quot; Minsky answered. "At Harvard, I had even taken a course with Howard Aikenquot;—one of the first computer designers. "Aiken had built an electromechanical machine in the early forties. It had only about a hundred memory registers, and even von Neumann's machine had only a thousand. On the one hand, I was afraid of the complexity of these machines. On the other hand, I thought that they weren't big enough to do anything interesting in the way of learning. In any case, I did my thesis on ideas about how the nervous system might learn.

36 1982

36.1 Learning-logic [194]

36.1.1 Original Abstract

None

37 1983

37.1 Optimization by Simulated Annealing [131]

37.1.1 Original Abstract

None

37.2 Neocognitron: A neural network model for a mechanism of visual pattern recognition [73]

37.2.1 Original Abstract

A recognition with a large-scale network is simulated on a PDP-11/34 minicomputer and is shown to have a great capability for visual pattern recognition. The model consists of nine layers of cells. The authors demonstrate that the model can be trained to recognize handwritten Arabic numerals even with considerable deformations in shape. A learning-with-a-teacher process is used for the reinforcement of the modifiable synapses in the new large-scale model, instead of the learning-without-a-teacher process applied to a previous model. The authors focus on the mechanism for pattern recognition rather than that for self-organization.

37.3 Neuronlike adaptive elements that can solve difficult learning control problems [17]

37.3.1 Original Abstract

It is shown how a system consisting of two neuronlike adaptive elements can solve a difficult learning control problem. The task is to balance a pole that is hinged to a movable cart by applying forces to the cart's base. It is argued that the learning problems faced by adaptive elements that are components of adaptive networks are at least as difficult as this version of the pole-balancing problem. The learning system consists of a single associative search element (ASE) and a single adaptive critic element (ACE). In the course of learning to balance the pole, the ASE constructs associations between input and output by searching under the influence of reinforcement feedback, and the ACE constructs a more informative evaluation function than reinforcement feedback alone can provide. The differences between this approach and other attempts

to solve problems using neurolike elements are discussed, as is the relation of this work to classical and instrumental conditioning in animal learning studies and its possible implications for research in the neurosciences.

37.3.2 Main points

38 1985

38.1 Une procédure d'apprentissage pour réseau a seuil asymmetrique (a Learning Scheme for Asymmetric Threshold Networks) [148]

38.1.1 Original Abstract

None

38.2 A Learning Algorithm for Boltzmann Machines* [3]

38.2.1 Original Abstract

The computational power of massively parallel networks of simple processing elements resides in the communication bandwidth provided by the hardware connections between elements. These connections can allow a significant fraction of the knowledge of the system to be applied to an instance of a problem in a very short time. One kind of computation for which massively parallel networks appear to be well suited is large constraint satisfaction searches, but to use the connections efficiently two conditions must be met: First, a search technique that is suitable for parallel networks must be found. Second, there must be some way of choosing internal representations which allow the preexisting hardware connections to be used efficiently for encoding the constraints in the domain being searched. We describe a general parallel search method, based on statistical mechanics, and we show how it leads to a general learning rule for modifying the connection strengths so as to incorporate knowledge about a task domain in an efficient way. We describe some simple examples in which the learning algorithm creates internal representations that are demonstrably the most efficient way of using the preexisting connectivity structure.

38.2.2 Main points

39 1986

39.1 Learning Process in an Asymmetric Threshold Network [149]

39.1.1 Original Abstract

Threshold functions and related operators are widely used as basic elements of adaptive and associative networks [Nakano 72, Amari 72, Hopfield 82]. There exist numerous learning rules for finding a set of weights to achieve a particular correspondence between input-output pairs. But early works in the field have shown that the number of threshold functions (or linearly separable functions) in N binary variables is small compared to the number of all possible boolean mappings in N variables, especially if N is large. This problem is one of the main limitations of most neural networks models where the state is fully specified by the environment during learning: they can only learn linearly separable functions of their inputs. Moreover, a learning procedure which requires the outside world to specify the state of every neuron during the learning session can hardly be considered as a general learning rule because in real-world conditions, only a partial information on the "ideal" network state for each task is available from the environment. It is possible to use a set of so-called "hidden units" [Hinton,Sejnowski,Ackley. 84], without direct interaction with the environment, which can compute intermediate predicates. Unfortunately, the global response depends on the output of a particular hidden unit in a highly non-linear way, moreover the nature of this dependence is influenced by the states of the other cells.

40 1987

40.1 Parallel networks that learn to pronounce English text [219]

40.1.1 Original Abstract

None

40.2 Intelligence: The Eye, the Brain, and the Computer [65]

40.2.1 Original Abstract

This book treats the question of how far we have come in understanding intelligence and in duplicating it mechanically. The major facets of intelligence– reasoning, vision, language and learning are discussed as an approach to contrasting biological intelligence with current computer realizations.

40.3 Highly parallel, hierarchical, recognition cone perceptual structures [247]

40.3.1 Original Abstract

None

41 1988

41.1 Neurocomputing: foundations of research [9]

41.1.1 Original Abstract

None

41.1.2 Main points

41.2 Radial basis functions, multi-variable functional interpolation and adaptive networks [31]

41.2.1 Original Abstract

The relationship between 'learning' in adaptive layered networks and the fitting of data with high dimensional surfaces is discussed. This leads naturally to a picture of 'generalization in terms of interpolation between known data points and suggests a rational approach to the theory of such networks. A class of adaptive networks is identified which makes the interpolation scheme explicit. This class has the property that learning is equivalent to the solution of a set of linear equations. These networks thus represent nonlinear relationships while having a guaranteed learning rule. Great Britain.

41.2.2 Main points

41.3 Self-organisation in a perceptual network [156]

41.3.1 Original Abstract

The emergence of a feature-analyzing function from the development rules of simple, multilayered networks is explored. It is shown that even a single developing cell of a layered network exhibits a remarkable set of optimization properties that are closely related to issues in statistics, theoretical physics, adaptive signal processing, the formation of knowledge representation in artificial intelligence, and information theory. The network studied is based on the visual system. These results are used to infer an information-theoretic principle that can be applied to the network as a whole, rather than a single cell. The organizing principle proposed is that the network connections develop in such a way as to maximize the amount of information that is preserved when signals are transformed at each processing stage, subject to certain constraints. The operation of this principle is illustrated for some simple cases.

41.4 A Combined Corner and Edge Detector [94]

41.4.1 Original Abstract

None

41.4.2 Main points

41.5 A theoretical framework for back-propagation [50]

41.5.1 Original Abstract

None

42 1989

42.1 A learning algorithm for continually running fully recurrent neural networks [265]

42.1.1 Original Abstract

The exact form of a gradient-following learning algorithm for completely recurrent networks running in continually sampled time is derived and used as the basis for practical algorithms for temporal supervised learning tasks. These algorithms have (1) the advantage that they do not require a precisely defined training interval, operating while the network runs; and (2) the disadvantage that they require nonlocal communication in the network being trained and are computationally expensive. These algorithms allow networks having recurrent connections to learn complex tasks that require the retention of information over time periods having either fixed or indefinite length.

42.1.2 Main points

42.2 A learning algorithm for continually running fully recurrent neural networks [265]

42.2.1 Original Abstract

The exact form of a gradient-following learning algorithm for completely recurrent networks running in continually sampled time is derived and used as the basis for practical algorithms for temporal supervised learning tasks. These algorithms have (1) the advantage that they do not require a precisely defined training interval, operating while the network runs; and (2) the disadvantage that they require nonlocal communication in the network being trained and are computationally expensive. These algorithms allow networks having recurrent connections to learn complex tasks that require the retention of information over time periods having either fixed or indefinite length.

42.2.2 Main points

42.3 Connectionism: Past, present, and future [200]

42.3.1 Original Abstract

Research efforts to study computation and cognitive modeling on neurally-inspired mechanisms have come to be called Connectionism. Rather than being brand new, it is actually the rebirth of a research programme which thrived from the 40s through the 60s and then was severely retrenched in the 70s. Connectionism is often posed as a paradigmatic competitor to the Symbolic Processing tradition of Artificial Intelligence (Dreyfus & Dreyfus, 1988), and, indeed, the counterpoint in the timing of their intellectual and commercial fortunes may lead one to believe that research in cognition is merely a zero-sum game. This paper surveys the history of the field, often in relation to AI, discusses its current successes and failures, and makes some predictions for where it might lead in the future.

42.4 Neurocomputing [99]

42.4.1 Original Abstract

Exploring many aspects of neurocomputers, this book gives an overview of the network theory behind them, including a background review, basic concepts, associative networks, mapping networks, spatiotemporal networks, and adaptive resonance networks.

42.5 Multilayer feedforward networks are universal approximators [108]

42.5.1 Original Abstract

This paper rigorously establishes that standard multilayer feedforward networks with as few as one hidden layer using arbitrary squashing functions are capable of approximating any Borel measurable function from one finite dimensional space to another to any desired degree of accuracy, provided sufficiently many hidden units are available. In this sense, multilayer feedforward networks are a class of universal approximators.

42.5.2 Main points

42.6 A learning algorithm for continually running fully recurrent neural networks [265]

42.6.1 Original Abstract

The exact form of a gradient-following learning algorithm for completely recurrent networks running in continually sampled time is derived and used as the basis for practical algorithms for temporal supervised learning tasks. These algorithms have (1) the advantage that they do not require a precisely defined training interval, operating while the network runs; and (2) the disadvantage that they require nonlocal communication in the network being trained and are computationally expensive. These algorithms allow networks having recurrent connections to learn complex tasks that require the retention of information over time periods having either fixed or indefinite length.

42.7 Backpropagation applied to handwritten zip code recognition [146]

42.7.1 Original Abstract

The ability of learning networks to generalize can be greatly enhanced by providing constraints from the task domain. This paper demonstrates how such constraints can be integrated into a backpropagation network through the architecture of the network. This approach has been successfully applied to the recognition of handwritten zip code digits provided by the U.S. Postal Service. A single network learns the entire recognition operation, going from the normalized image of the character to the final classification.

42.8 Generalization and network design strategies [144]

42.8.1 Original Abstract

An interestmg property of connectiomst systems is their ability tolearn from examples. Although most recent work in the field concentrateson reducing learning times, the most important feature of a learning ma-chine is its generalization performance. It is usually accepted that goodgeneralization performance on real-world problems cannot be achievedunless some a pnon knowl-

edge about the task is butlt Into the system.Back-propagation networks provide a way of specifymg such knowledgeby imposing constraints both on the architecture of the network and onits weights. In general, such constramts can be considered as particulartransformations of the parameter spaceBuilding a constramed network for image recogmtton appears to be afeasible task. We descnbe a small handwritten digit recogmtion problemand show that, even though the problem is linearly separable, single layernetworks exhibit poor generalizatton performance. Multtlayer constrainednetworks perform very well on this task when orgamzed in a hierarchicalstructure with shift invariant feature detectors.These results confirm the idea that minimizing the number of freeparameters in the network enhances generalization.

43 1990

43.1 Handwritten digit recognition with a back-propagation network [150]

43.1.1 Original Abstract

None

44 1992

44.1 Artificial Neural Networks: Concepts and Control Applications [251]

44.1.1 Original Abstract

None

44.2 A training algorithm for optimal margin classifiers [30]

44.2.1 Original Abstract

A training algorithm that maximizes the margin between the training patterns and the decision boundary is presented. The technique is applicable to a wide variety of classifiaction functions, including Perceptrons, polynomials,

and Radial Basis Functions. The effective number of parameters is adjusted automatically to match the complexity of the problem. The solution is expressed as a linear combination of supporting patterns. These are the subset of training patterns that are closest to the decision boundary. Bounds on the generalization performance based on the leave-one-out method and the VC-dimension are given. Experimental results on optical character recognition problems demonstrate the good generalization obtained when compared with other learning algorithms. 1 INTRODUCTION Good generalization performance of pattern classifiers is achieved when the capacity of the classification function is matched to the size of the training set. Classifiers with a large numb...

44.2.2 Main points

44.3 Connectionist learning of belief networks [181]

44.3.1 Original Abstract

Connectionist learning procedures are presented for "sigmoid" and "noisy-OR" varieties of probabilistic belief networks. These networks have previously been seen primarily as a means of representing knowledge derived from experts. Here it is shown that the "Gibbs sampling" simulation procedure for such networks can support maximum-likelihood learning from empirical data through local gradient ascent. This learning procedure resembles that used for "Boltzmann machines", and like it, allows the use of "hidden" variables to model correlations between visible variables. Due to the directed nature of the connections in a belief network, however, the "negative phase" of Boltzmann machine learning is unnecessary. Experimental results show that, as a result, learning in a sigmoid belief network can be faster than in a Boltzmann machine. These networks have other advantages over Boltzmann machines in pattern classification and decision making applications, are naturally applicable to unsupervised learning problems, and provide a link between work on connectionist learning and work on the representation of expert knowledge.

45 1993

45.1 AI: The tumultuous history of the search for artificial intelligence [47]

45.1.1 Original Abstract

None

45.2 Mining association rules between sets of items in large databases [5]

45.2.1 Original Abstract

We are given a large database of customer transactions. Each transaction consists of items purchased by a customer in a visit. We present an efficient algorithm that generates all significant association rules between items in the database. The algorithm incorporates buffer management and novel estimation and pruning techniques. We also present results of applying this algorithm to sales data obtained from a large retailing company, which shows the effectiveness of the algorithm.

45.2.2 Main points

46 1994

46.1 Neural Network Modeling: Statistical Mechanics and Cybernetic Perspectives [182]

46.1.1 Original Abstract

Neural Network Modeling offers a cohesive approach to the statistical mechanics and principles of cybernetics as a basis for neural network modeling. It brings together neurobiologists and the engineers who design intelligent automata to understand the physics of collective behavior pertinent to neural elements and the self-control aspects of neurocybernetics. The theoretical perspectives and explanatory projections portray the most current information in the field, some of which counters certain conventional concepts in the visualization of neuronal interactions.

46.2 Neuro-vision systems: A tutorial. [89]

46.2.1 Original Abstract

None

46.3 Neural Networks and Related Methods for Classification [205]

46.3.1 Original Abstract

Feed-forward neural networks are now widely used in classification problems, whereas nonlinear methods of discrimination developed in the statistical field are much less widely known. A general framework for classification is set up within which methods from statistics, neural networks, pattern recognition and machine learning can be compared. Neural networks emerge as one of a class of flexible non-linear regression methods which can be used to classify via regression. Many interesting issues remain, including parameter estimation, the assessment of the classifiers and in algorithm development.

46.4 Neural networks: a comprehensive foundation [96]

46.4.1 Original Abstract

None

47 1995

47.1 An information-maximization approach to blind separation and blind deconvolution [20]

47.1.1 Original Abstract

We derive a new self-organizing learning algorithm that maximizes the information transferred in a network of nonlinear units. The algorithm does not assume any knowledge of the input distributions, and is defined here for the zero-noise limit. Under these conditions, information maximization has extra properties not found in the linear case (Linsker 1989). The nonlinearities in the transfer function are able to pick up higher-order moments of the

input distributions and perform something akin to true redundancy reduction between units in the output representation. This enables the network to separate statistically independent components in the inputs: a higher-order generalization of principal components analysis. We apply the network to the source separation (or cocktail party) problem, successfully separating unknown mixtures of up to 10 speakers. We also show that a variant on the network architecture is able to perform blind deconvolution (cancellation of unknown echoes and reverberation in a speech signal). Finally, we derive dependencies of information transfer on time delays. We suggest that information maximization provides a unifying framework for problems in "blind" signal processing.

47.1.2 Main points

47.2 Principles of digital image synthesis: Vol. 1 [81]

47.2.1 Original Abstract

None

47.3 Survey and critique of techniques for extracting rules from trained artificial neural networks [10]

47.3.1 Original Abstract

It is becoming increasingly apparent that, without some form of explanation capability, the full potential of trained artificial neural networks (ANNs) may not be realised. This survey gives an overview of techniques developed to redress this situation. Specifically, the survey focuses on mechanisms, procedures, and algorithms designed to insert knowledge into ANNs (knowledge initialisation), extract rules from trained ANNs (rule extraction), and utilise ANNs to refine existing rule bases (rule refinement). The survey also introduces a new taxonomy for classifying the various techniques, discusses their modus operandi, and delineates criteria for evaluating their efficacy.

47.3.2 Main points

47.4 The "wake-sleep" algorithm for unsupervised neural networks. [100]

47.4.1 Original Abstract

An unsupervised learning algorithm for a multilayer network of stochastic neurons is described. Bottom-up "recognition" connections convert the input into representations in successive hidden layers, and top-down "generative" connections reconstruct the representation in one layer from the representation in the layer above. In the "wake" phase, neurons are driven by recognition connections, and generative connections are adapted to increase the probability that they would reconstruct the correct activity vector in the layer below. In the "sleep" phase, neurons are driven by generative connections, and recognition connections are adapted to increase the probability that they would produce the correct activity vector in the layer above.

47.4.2 Main points

47.5 Convolutional networks for images, speech, and time series [145]

47.5.1 Original Abstract

INTRODUCTION The ability of multilayer back-propagation networks to learn complex, high-dimensional, nonlinear mappings from large collections of examples makes them obvious candidates for image recognition or speech recognition tasks (see PATTERN RECOGNITION AND NEURAL NETWORKS). In the traditional model of pattern recognition, a hand-designed feature extractor gathers relevant information from the input and eliminates irrelevant variabilities. A trainable classifier then categorizes the resulting feature vectors (or strings of symbols) into classes. In this scheme, standard, fully-connected multilayer networks can be used as classifiers. A potentially more interesting scheme is to eliminate the feature extractor, feeding the network with "raw" inputs (e.g. normalized images), and to rely on backpropagation to turn the first few layers into an appropriate feature extractor. While this can be done with an ordinary fully connected feed-forward network with some success for tasks

48 1996

48.1 On Alan Turing's anticipation of connectionism [45]

48.1.1 Original Abstract

It is not widely realised that Turing was probably the first person to consider building computing machines out of simple, neuron-like elements connected together into networks in a largely random manner. Turing called his networks 'unorganised machines'. By the application of what he described as 'appropriate interference, mimicking education' an unorganised machine can be trained to perform any task that a Turing machine can carry out, provided the number of 'neurons' is sufficient. Turing proposed simulating both the behaviour of the network and the training process by means of a computer program. We outline Turing's connectionist project of 1948.

48.2 Mean Field Theory for Sigmoid Belief Networks [213]

48.2.1 Original Abstract

We develop a mean field theory for sigmoid belief networks based on ideas from statistical mechanics. Our mean field theory provides a tractable approximation to the true probability distribution in these networks; it also yields a lower bound on the likelihood of evidence. We demonstrate the utility of this framework on a benchmark problem in statistical pattern recognition—the classification of handwritten digits.

48.2.2 Main points

Comment: See http://www.jair.org/ for any accompanying files

48.3 Affine / photometric invariants for planar intensity patterns [87]

48.3.1 Original Abstract

The paper contributes to the viewpoint invariant recognition of planar patterns, especially labels and signs under affine deformations. By their nature, the information of such 'eye-catchers' is not contained in the outline or frame — they often are affinely equivalent like parallelograms and ellipses — but in the intensity content within. Moment invariants are well suited for their recognition. They need a closed bounding contour, but this is comparatively easy to provide for the simple shapes considered. On the other hand, they characterize the intensity patterns without the need for error prone feature extraction. This paper uses moments as the basic features, but extends the literature in two respects: (1) deliberate mixes of different types of moments to keep the order of the moments (and hence also the sensitivity to noise) low and yet have a sufficiently large number to safeguard discriminant power; and (2) invariance with respect to photometric changes is incorporated in order to find the simplest moment invariants that can cope with changing lighting conditions which can hardly be avoided when changing viewpoint. The paper gives complete classifications of such affine / photometric moment invariants. Experiments are described that illustrate the use of some of them.

48.4 Pattern Recognition and Neural Networks [206]

48.4.1 Original Abstract

This 1996 book is a reliable account of the statistical framework for pattern recognition and machine learning. With unparalleled coverage and a wealth of case-studies this book gives valuable insight into both the theory and the enormously diverse applications (which can be found in remote sensing, astrophysics, engineering and medicine, for example). So that readers can develop their skills and understanding, many of the real data sets used in the book are available from the author's website: www.stats.ox.ac.uk/ ripley/PRbook/. For the same reason, many examples are included to illustrate real problems in pattern recognition. Unifying principles are highlighted, and the author gives an overview of the state of the subject, making the book valuable to experienced researchers in statistics, machine learning/artificial intelligence and engineering. The clear writing style means that the book is also a superb

introduction for non-specialists.

49 1997

49.1 Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference [196]

49.1.1 Original Abstract

Probabilistic Reasoning in Intelligent Systems is a complete and accessible account of the theoretical foundations and computational methods that underlie plausible reasoning under uncertainty. The author provides a coherent explication of probability as a language for reasoning with partial belief and offers a unifying perspective on other AI approaches to uncertainty, such as the Dempster-Shafer formalism, truth maintenance systems, and nonmonotonic logic.The author distinguishes syntactic and semantic approaches to uncertainty–and offers techniques, based on belief networks, that provide a mechanism for making semantics-based systems operational. Specifically, network-propagation techniques serve as a mechanism for combining the theoretical coherence of probability theory with modern demands of reasoning-systems technology: modular declarative inputs, conceptually meaningful inferences, and parallel distributed computation. Application areas include diagnosis, forecasting, image interpretation, multi-sensor fusion, decision support systems, plan recognition, planning, speech recognition–in short, almost every task requiring that conclusions be drawn from uncertain clues and incomplete information.Probabilistic Reasoning in Intelligent Systems will be of special interest to scholars and researchers in AI, decision theory, statistics, logic, philosophy, cognitive psychology, and the management sciences. Professionals in the areas of knowledge-based systems, operations research, engineering, and statistics will find theoretical and computational tools of immediate practical use. The book can also be used as an excellent text for graduate-level courses in AI, operations research, or applied probability.

49.2 Elements of artificial neural networks [167]

49.2.1 Original Abstract

None

49.3 Bain on neural networks [263]

49.3.1 Original Abstract

In his book Mind and body (1873), Bain set out an account in which he related the processes of associative memory to the distribution of activity in neural groupings–or neural networks as they are now termed. In the course of this account, Bain anticipated certain aspects of connectionist ideas that are normally attributed to 20th-century authors–most notably Hebb (1949). In this paper we reproduce Bain's arguments relating neural activity to the workings of associative memory which include an early version of the principles enshrined in Hebb's neurophysiological postulate. Nonetheless, despite their prescience, these specific contributions to the connectionist case have been almost entirely ignored. Eventually, Bain came to doubt the practicality of his own arguments and, in so doing, he seems to have ensured that his ideas concerning neural groupings exerted little or no influence on the subsequent course of theorizing in this area.

49.4 Bidirectional recurrent neural networks [217]

49.4.1 Original Abstract

In the first part of this paper, a regular recurrent neural network (RNN) is extended to a bidirectional recurrent neural network (BRNN). The BRNN can be trained without the limitation of using input information just up to a preset future frame. This is accomplished by training it simultaneously in positive and negative time direction. Structure and training procedure of the proposed network are explained. In regression and classification experiments on artificial data, the proposed structure gives better results than other approaches. For real data, classification experiments for phonemes from the TIMIT database show the same tendency. In the second part of this paper, it is shown how the proposed bidirectional structure can be easily modified to allow efficient estimation of the conditional posterior probability of complete symbol sequences without making any explicit assumption about the shape of the distribution. For this part, experiments on real data are reported

49.5 Introduction to multi-layer feed-forward neural networks [238]

49.5.1 Original Abstract

None

49.6 Neural Networks for Pattern Recognition. [135]

49.6.1 Original Abstract

None

50 1998

50.1 Reinforcement Learning: An Introduction [237]

50.1.1 Original Abstract

Reinforcement learning, one of the most active research areas in artificial-intelligence, is a computational approach to learning whereby an agent tries to maximize the totalamount of reward it receives when interacting with a complex, uncertain environment. InReinforcement Learning, Richard Sutton and Andrew Barto provide a clear andsimple account of the key ideas and algorithms of reinforcement learning. Their discussion rangesfrom the history of the field's intellectual foundations to the most recent developments andapplications. The only necessary mathematical background is familiarity with elementary concepts ofprobability.The book is divided into three parts. Part I defines thereinforcement learning problem in terms of Markov decision processes. Part II provides basicsolution methods: dynamic programming, Monte Carlo methods, and temporal-difference learning. PartIII presents a unified view of the solution methods and incorporates artificial neural networks,eligibility traces, and planning; the two final chapters present case studies and consider thefuture of reinforcement learning.

50.2 Neural Networks: An Introductory Guide for Social Scientists [77]

50.2.1 Original Abstract

This book provides the first accessible introduction to neural network analysis as a methodological strategy for social scientists. The author details numerous studies and examples which illustrate the advantages of neural network analysis over other quantitative and modelling methods in widespread use. Methods are presented in an accessible style for readers who do not have a background in computer science. The book provides a history of neural network methods, a substantial review of the literature, detailed applications, coverage of the most common alternative models and examples of two leading software packages for neural network analysis.

50.3 Feature detection with automatic scale selection [155]

50.3.1 Original Abstract

The fact that objects in the world appear in different ways depending on the scale of observation has important implications if one aims at describing them. It shows that the notion of scale is of utmost importance when processing unknown measurement data by automatic methods. In their seminal works, Witkin (1983) and Koenderink (1984) proposed to approach this problem by representing image structures at different scales in a so-called scale-space representation. Traditional scale-space theory building on this work, however, does not address the problem of how to select local appropriate scales for further analysis. This article proposes a systematic methodology for dealing with this problem. A framework is presented for generating hypotheses about interesting scale levels in image data, based on a general principle stating that local extrema over scales of different combinations of -normalized derivatives are likely candidates to correspond to interesting structures. Specifically, it is shown how this idea can be used as a major mechanism in algorithms for automatic scale selection, which adapt the local scales of processing to the local image structure. Support for the proposed approach is given in terms of a general theoretical investigation of the behaviour of the scale selection method under rescalings of the input pattern and by integration with different types of early visual modules, including experiments

on real-world and synthetic data. Support is also given by a detailed analysis of how different types of feature detectors perform when integrated with a scale selection mechanism and then applied to characteristic model patterns. Specifically, it is described in detail how the proposed methodology applies to the problems of blob detection, junction detection, edge detection, ridge detection and local frequency estimation. In many computer vision applications, the poor performance of the low-level vision modules constitutes a major bottleneck. It is argued that the inclusion of mechanisms for automatic scale selection is essential if we are to construct vision systems to automatically analyse complex unknown environments.

50.4 Gradient-based learning applied to document recognition [147]

50.4.1 Original Abstract

Multilayer neural networks trained with the back-propagation algorithm constitute the best example of a successful gradient based learning technique. Given an appropriate network architecture, gradient-based learning algorithms can be used to synthesize a complex decision surface that can classify high-dimensional patterns, such as handwritten characters, with minimal preprocessing. This paper reviews various methods applied to handwritten character recognition and compares them on a standard handwritten digit recognition task. Convolutional neural networks, which are specifically designed to deal with the variability of 2D shapes, are shown to outperform all other techniques. Real-life document recognition systems are composed of multiple modules including field extraction, segmentation recognition, and language modeling. A new learning paradigm, called graph transformer networks (GTN), allows such multimodule systems to be trained globally using gradient-based methods so as to minimize an overall performance measure. Two systems for online handwriting recognition are described. Experiments demonstrate the advantage of global training, and the flexibility of graph transformer networks. A graph transformer network for reading a bank cheque is also described. It uses convolutional neural network character recognizers combined with global training techniques to provide record accuracy on business and personal cheques. It is deployed commercially and reads several million cheques per day

50.4.2 Main points

- *LeNet-5*
- *Clarification: In this paper "stride" is not mentioned, but as Krizhevsky2012 et.al. started using it, new implementations of CNN need to define its value.*
- *Conv: Convolutional layer*
- *Subs: Subsampling layer (summed * coefficient + bias)*
- *Full: Fully connected network*
- *ERBF: Euclidian Radial Basis Function units*
  - *input 32x32 pixel image (original images are 28x28)*
  - *Conv1 :*
    - *6@28x28 filter 5x5*
    - *stride 1*
    - *Connections* $= 5 * 5 * 28 * 28 * 6 + 6 * 28 * 28 = 122,304$
    - *Train. param.* $= 5 * 5 * 6 + 6 = 156$
  - *Subs2 :*
    - *6@14x14 range 2x2*
    - *stride 2*
    - *Connections* $= 6 * 28 * 28 + 6 * 14 * 14 = 5,880$
    - *Train. param. = coefficient + bias* $= 6 + 6 = 156$
  - *Conv3 :*
    - *16@10x10 filter 5x5*
    - *stride 1*
    - *Connections* $= 6 * 5 * 5 * 10 * 10 * 10 + 10 * 10 * 16 = 151,600$
    - *Train. param.* $= 5*5*3*6+5*5*4*9+5*5*6*1+16 = 1,516$
    - *Note:*
    - *This layer is not completly connected, see table 1 for specific connections*
    - *Expected Connections* $= 6 * 5 * 5 * 10 * 10 * 16 + 10 * 10 * 16 = 241,600$

    - *Expected train. param* $= 5 * 5 * 16 * 6 + 16 = 2416$
- *Subs4 :*
    - *16@5x5 range 2x2*
    - *stride 2*
    - *Connections* $= 16 * 10 * 10 + 16 * 5 * 5 = 2,000$
    - *Train. param. = coefficient + bias* $= 16 + 16 = 32$
- *Conv5 :*
    - *120@1x1 filter 5x5*
    - *stride 0*
    - *Connections and train. param.* $= 16*5*5*120+120 = 48,120$
- *Full6 : 84 Atanh(Sa)*
    - *Connections and train. param.* $= 120 * 84 + 84 = 10,164$
- *ERBF7 : 10*
    - *Connections and train. param.* $= 84 * 10 = 840$

50.5 Locating facial region of a head-and-shoulders color image [38]

50.5.1 Original Abstract

This paper addresses our proposed method to automatically locate the person's face from a given image that consists of a head-and-shoulders view of the person and a complex background scene. The method involves a fast, simple and yet robust algorithm that exploits the spatial distribution characteristics of human skin color. It first uses the chrominance component of the input image to detect pixels with skin color appearance. Then, bused on the spatial distribution of the detected skin-color pixels and their corresponding luminance values, the algorithm employs some regularization processes to reinforce regions of skin-color pixels that are more likely to belong to the facial regions and eliminate those that are not. The performance of the face localization algorithm is illustrated by some simulation results carried out on various head-and-shoulders test images

50.5.2 Main points

51 1999

51.1 Alan Turingś forgotten ideas in Computer Science [43]

51.1.1 Original Abstract

None

51.2 Text categorisation: A survey [1]

51.2.1 Original Abstract

None

51.2.2 Main points

51.3 Face segmentation using skin-color map in videophone applications [39]

51.3.1 Original Abstract

This paper addresses our proposed method to automatically segment out a person's face from a given image that consists of a head-and-shoulders view of the person and a complex background scene. The method involves a fast, reliable, and effective algorithm that exploits the spatial distribution characteristics of human skin color. A universal skin-color map is derived and used on the chrominance component of the input image to detect pixels with skin-color appearance. Then, based on the spatial distribution of the detected skin-color pixels and their corresponding luminance values, the algorithm employs a set of novel regularization processes to reinforce regions of skin-color pixels that are more likely to belong to the facial regions and eliminate those that are not. The performance of the face-segmentation algorithm is illustrated by some simulation results carried out on various head-and-shoulders test images. The use of face segmentation for video coding in applications such as videotelephony is then presented. We explain how the face-segmentation results can be used to improve the perceptual quality of a videophone sequence encoded by the H.261-compliant coder

51.3.2 Main points

51.4 Efficient mining of emerging patterns: Discovering trends and differences [56]

51.4.1 Original Abstract

None

52 2000

52.1 Principles of Neurocomputing for Science and Engineering [90]

52.1.1 Original Abstract

From the Publisher:This exciting new text covers artificial neural networks,but more specifically,neurocomputing. Neurocomputing is concerned with processing information,which involves a learning process within an artificial neural network architecture. This neural architecture responds to inputs according to a defined learning rule and then the trained network can be used to perform certain tasks depending on the application. Neurocomputing can play an important role in solving certain problems such as pattern recognition,optimization,event classification,control and identification of nonlinear systems,and statistical analysis. "Principles of Neurocomputing for Science and Engineering," unlike other neural networks texts,is written specifically for scientists and engineers who want to apply neural networks to solve complex problems. For each neurocomputing concept,a solid mathematical foundation is presented along with illustrative examples to accompany that particular architecture and associated training algorithm. The book is primarily intended for graduate-level neural networks courses,but in some instances may be used at the undergraduate level. The book includes many detailed examples and an extensive set of end-of-chapter problems.

52.2 Principles of Neurocomputing for Science and Engineering [91]

52.2.1 Original Abstract

From the Publisher:This exciting new text covers artificial neural networks,but more specifically,neurocomputing. Neurocomputing is concerned with processing information,which involves a learning process within an artificial neural network architecture. This neural architecture responds to inputs according to a defined learning rule and then the trained network can be used to perform certain tasks depending on the application. Neurocomputing can play an important role in solving certain problems such as pattern recognition,optimization,event classification,control and identification of nonlinear systems,and statistical analysis. "Principles of Neurocomputing for Science and Engineering," unlike other neural networks texts,is written specifically for scientists and engineers who want to apply neural networks to solve complex problems. For each neurocomputing concept,a solid mathematical foundation is presented along with illustrative examples to accompany that particular architecture and associated training algorithm. The book is primarily intended for graduate-level neural networks courses,but in some instances may be used at the undergraduate level. The book includes many detailed examples and an extensive set of end-of-chapter problems.

52.3 Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces [114]

52.3.1 Original Abstract

Olshausen and Field (1996) applied the principle of independence maximization by sparse coding to extract features from natural images. This leads to the emergence of oriented linear filters that have simultaneous localization in space and in frequency, thus resembling Gabor functions and simple cell receptive fields. In this article, we show that the same principle of independence maximization can explain the emergence of phase- and shift-invariant features, similar to those found in complex cells. This new kind of emergence is obtained by maximizing the independence between norms of projections on linear subspaces (instead of the independence of simple linear filter outputs).

The norms of the projections on such "independent feature subspaces" then indicate the values of invariant features.

52.4 Independent component analysis applied to feature extraction from colour and stereo images. [110]

52.4.1 Original Abstract

Previous work has shown that independent component analysis (ICA) applied to feature extraction from natural image data yields features resembling Gabor functions and simple-cell receptive fields. This article considers the effects of including chromatic and stereo information. The inclusion of colour leads to features divided into separate red/green, blue/yellow, and bright/dark channels. Stereo image data, on the other hand, leads to binocular receptive fields which are tuned to various disparities. The similarities between these results and the observed properties of simple cells in the primary visual cortex are further evidence for the hypothesis that visual cortical neurons perform some type of redundancy reduction, which was one of the original motivations for ICA in the first place. In addition, ICA provides a principled method for feature extraction from colour and stereo images; such features could be used in image processing operations such as denoising and compression, as well as in pattern recognition.

52.4.2 Main points

52.5 Independent component analysis: algorithms and applications. [115]

52.5.1 Original Abstract

A fundamental problem in neural network research, as well as in many other disciplines, is finding a suitable representation of multivariate data, i.e. random vectors. For reasons of computational and conceptual simplicity, the representation is often sought as a linear transformation of the original data. In other words, each component of the representation is a linear combination of the original variables. Well-known linear transformation methods include principal component analysis, factor analysis, and projection pursuit. Independent component analysis (ICA) is a recently developed method in which the goal is to find a linear representation of non-Gaussian data so that the

components are statistically independent, or as independent as possible. Such a representation seems to capture the essential structure of the data in many applications, including feature extraction and signal separation. In this paper, we present the basic theory and applications of ICA, and our recent work on the subject.

52.6 Fast and inexpensive color image segmentation for interactive robots [32]

52.6.1 Original Abstract

Vision systems employing region segmentation by color are crucial in real-time mobile robot applications. With careful attention to algorithm efficiency, fast color image segmentation can be accomplished using commodity image capture and CPU hardware. This paper describes a system capable of tracking several hundred regions of up to 32 colors at 30 Hz on general purpose commodity hardware. The software system consists of: a novel implementation of a threshold classifier, a merging system to form regions through connected components, a separation and sorting system that gathers various region features, and a top down merging heuristic to approximate perceptual grouping. A key to the efficiency of our approach is a new method for accomplishing color space thresholding that enables a pixel to be classified into one or more, up to 32 colors, using only two logical AND operations. The algorithms and representations are described, as well as descriptions of three applications in which it has been used

52.6.2 Main points

52.7 A Bayesian approach to skin color classification in YCbCr color space [37]

52.7.1 Original Abstract

This paper addresses an image classification technique that uses the Bayes decision rule for minimum cost to classify pixels into skin color and non-skin color. Color statistics are collected from YCbCr color space. The Bayesian approach to skin color classification is discussed along with an overview of YCbCr color space. Experimental results demonstrate that this approach can

achieve good classification outcomes, and it is robust against different skin colors

52.7.2 Main points

53 2001

53.1 Saliency, Scale and Image Description [123]

53.1.1 Original Abstract

Many computer vision problems can be considered to consist of two main tasks: the extraction of image content descriptions and their subsequent matching. The appropriate choice of type and level of description is of course task dependent, yet it is generally accepted that the low-level or so called early vision layers in the Human Visual System are context independent. This paper concentrates on the use of low-level approaches for solving computer vision problems and discusses three inter-related aspects of this: saliency; scale selection and content description. In contrast to many previous approaches which separate these tasks, we argue that these three aspects are intrinsically related. Based on this observation, a multiscale algorithm for the selection of salient regions of an image is introduced and its application to matching type problems such as tracking, object recognition and image retrieval is demonstrated.

53.2 The elements of statistical learning [69]

53.2.1 Original Abstract

None

54 2002

54.1 Computer vision: a modern approach [67]

54.1.1 Original Abstract

None

54.2 Why color management? [129]

54.2.1 Original Abstract

It seems that everywhere you look there is some article or discussion about color management. Why all the fuss? Do I need to management my colors? We have been creating colored artifacts for a very long time and I don't think we have needed color management. So why now? Most of these discussions also refer to the ICC. What is that? These and other questions will be answered in a straightforward manner in plain English. Adobe Systems has pioneered the use of desktop computers for color work, and the author has helped Adobe pick its way down conflicting color paths with confusing road signs over the last 10 years.

55 2003

55.1 Neural networks in computer intelligence [70]

55.1.1 Original Abstract

None

55.2 Models of distributed associative memory networks in the brain $*$ [230]

55.2.1 Original Abstract

None

55.3 Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis [224]

55.3.1 Original Abstract

Neural networks are a powerful technology forclassification of visual inputs arising from documents.However, there is a confusing plethora of different neuralnetwork methods that are used in the literature and inindustry. This paper describes a set of concrete bestpractices that document analysis researchers can use toget good results with neural networks. The mostimportant

practice is getting a training set as large aspossible: we expand the training set by adding a newform of distorted data. The next most important practiceis that convolutional neural networks are better suited forvisual document tasks than fully connected networks. Wepropose that a simple "do-it-yourself" implementation ofconvolution with a flexible architecture is suitable formany visual document problems. This simpleconvolutional neural network does not require complexmethods, such as momentum, weight decay, structure-dependent learning rates, averaging layers, tangent prop,or even finely-tuning the architecture. The end result is avery simple yet general architecture which can yieldstate-of-the-art performance for document analysis. Weillustrate our claims on the MNIST set of English digitimages.

55.3.2 Main points

- *Get a training set as large as possible*
- *No need of complex methods, such as momentum, weight decay, structure-dependent learning rates, averaging layers, tangent prop, or even finely-tuning the architecture*
- *Increment dataset by:*
  - *Affine transformations: translations, scaling, homothety, similarity transformation, reflection, rotation, shear mapping, and compositions.*
  - *Elastic distortions*
- *In this paper the authors justify the use of elastic deformations on MNIST data corresponding to uncontrolled oscillations of the hand muscles, dampened by inertia.*
- *They get the best results on MNIST to date with CNN, affine and elastic transformations of the dataset (0.4% error).*

56 2004

56.1 Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication [?]

56.1.1 Original Abstract

We present a method for learning nonlinear systems, echo state networks (ESNs). ESNs employ artificial recurrent neural networks in a way that has recently been proposed independently as a learning mechanism in biological brains. The learning method is computationally efficient and easy to use. On a benchmark task of predicting a chaotic time series, accuracy is improved by a factor of 2400 over previous techniques. The potential for engineering applications is illustrated by equalizing a communication channel, where the signal error rate is improved by two orders of magnitude.

56.1.2 Main points

56.2 Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication [117]

56.2.1 Original Abstract

We present a method for learning nonlinear systems, echo state networks (ESNs). ESNs employ artificial recurrent neural networks in a way that has recently been proposed independently as a learning mechanism in biological brains. The learning method is computationally efficient and easy to use. On a benchmark task of predicting a chaotic time series, accuracy is improved by a factor of 2400 over previous techniques. The potential for engineering applications is illustrated by equalizing a communication channel, where the signal error rate is improved by two orders of magnitude.

56.2.2 Main points

56.3 Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication [?]

56.3.1 Original Abstract

We present a method for learning nonlinear systems, echo state networks (ESNs). ESNs employ artificial recurrent neural networks in a way that has recently been proposed independently as a learning mechanism in biological brains. The learning method is computationally efficient and easy to use. On a benchmark task of predicting a chaotic time series, accuracy is improved by a factor of 2400 over previous techniques. The potential for engineering applications is illustrated by equalizing a communication channel, where the signal error rate is improved by two orders of magnitude.

56.3.2 Main points

56.4 Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication [117]

56.4.1 Original Abstract

We present a method for learning nonlinear systems, echo state networks (ESNs). ESNs employ artificial recurrent neural networks in a way that has recently been proposed independently as a learning mechanism in biological brains. The learning method is computationally efficient and easy to use. On a benchmark task of predicting a chaotic time series, accuracy is improved by a factor of 2400 over previous techniques. The potential for engineering applications is illustrated by equalizing a communication channel, where the signal error rate is improved by two orders of magnitude.

56.4.2 Main points

56.5 Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication [?]

56.5.1 Original Abstract

We present a method for learning nonlinear systems, echo state networks (ESNs). ESNs employ artificial recurrent neural networks in a way that has recently been proposed independently as a learning mechanism in biological brains. The learning method is computationally efficient and easy to use. On a benchmark task of predicting a chaotic time series, accuracy is improved by a factor of 2400 over previous techniques. The potential for engineering applications is illustrated by equalizing a communication channel, where the signal error rate is improved by two orders of magnitude.

56.5.2 Main points

56.6 Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication [117]

56.6.1 Original Abstract

We present a method for learning nonlinear systems, echo state networks (ESNs). ESNs employ artificial recurrent neural networks in a way that has recently been proposed independently as a learning mechanism in biological brains. The learning method is computationally efficient and easy to use. On a benchmark task of predicting a chaotic time series, accuracy is improved by a factor of 2400 over previous techniques. The potential for engineering applications is illustrated by equalizing a communication channel, where the signal error rate is improved by two orders of magnitude.

56.6.2 Main points

56.7 PCA-SIFT: a more distinctive representation for local image descriptors [126]

56.7.1 Original Abstract

Stable local feature detection and representation is a fundamental component of many image registration and object recognition algorithms. Mikolajczyk and Schmid (June 2003) recently evaluated a variety of approaches and identified the SIFT [D. G. Lowe, 1999] algorithm as being the most resistant to common image deformations. This paper examines (and improves upon) the local image descriptor used by SIFT. Like SIFT, our descriptors encode the salient aspects of the image gradient in the feature point's neighborhood; however, instead of using SIFT's smoothed weighted histograms, we apply principal components analysis (PCA) to the normalized gradient patch. Our experiments demonstrate that the PCA-based local descriptors are more distinctive, more robust to image deformations, and more compact than the standard SIFT representation. We also present results showing that using these descriptors in an image retrieval application results in increased accuracy and faster matching.

56.8 Robust wide-baseline stereo from maximally stable extremal regions [165]

56.8.1 Original Abstract

A new set of image elements that are put into correspondence, the so called extremal regions, is introduced. Extremal regions possess highly desirable properties: the set is closed under (1) continuous (and thus projective) transformation of image coordinates and (2) monotonic transformation of image intensities. An efficient (near linear complexity) and practically fast detection algorithm (near frame rate) is presented for an affinely invariant stable subset of extremal regions, the maximally stable extremal regions (MSER).

A new robust similarity measure for establishing tentative correspondences is proposed. The robustness ensures that invariants from multiple measurement regions (regions obtained by invariant constructions from extremal regions), some that are significantly larger (and hence discriminative) than the

MSERs, may be used to establish tentative correspondences.

The high utility of MSERs, multiple measurement regions and the robust metric is demonstrated in wide-baseline experiments on image pairs from both indoor and outdoor scenes. Significant change of scale (3.5×), illumination conditions, out-of-plane rotation, occlusion, locally anisotropic scale change and 3D translation of the viewpoint are all present in the test problems. Good estimates of epipolar geometry (average distance from corresponding points to the epipolar line below 0.09 of the inter-pixel distance) are obtained. The wide-baseline stereo problem, i.e. the problem of establishing correspondences between a pair of images taken from different viewpoints is studied.

A new set of image elements that are put into correspondence, the so called extremal regions, is introduced. Extremal regions possess highly desirable properties: the set is closed under (1) continuous (and thus projective) transformation of image coordinates and (2) monotonic transformation of image intensities. An efficient (near linear complexity) and practically fast detection algorithm (near frame rate) is presented for an affinely invariant stable subset of extremal regions, the maximally stable extremal regions (MSER).

A new robust similarity measure for establishing tentative correspondences is proposed. The robustness ensures that invariants from multiple measurement regions (regions obtained by invariant constructions from extremal regions), some that are significantly larger (and hence discriminative) than the MSERs, may be used to establish tentative correspondences.

The high utility of MSERs, multiple measurement regions and the robust metric is demonstrated in wide-baseline experiments on image pairs from both indoor and outdoor scenes. Significant change of scale (3.5×), illumination conditions, out-of-plane rotation, occlusion, locally anisotropic scale change and 3D translation of the viewpoint are all present in the test problems. Good estimates of epipolar geometry (average distance from corresponding points to the epipolar line below 0.09 of the inter-pixel distance) are obtained.

56.8.2 Main points

56.9 Scale & affine invariant interest point detectors [173]

56.9.1 Original Abstract

In this paper we propose a novel approach for detecting interest points invariant to scale and affine transformations. Our scale and affine invariant detectors are based on the following recent results: (1) Interest points extracted with the Harris detector can be adapted to affine transformations and give repeatable results (geometrically stable). (2) The characteristic scale of a local structure is indicated by a local extremum over scale of normalized derivatives (the Laplacian). (3) The affine shape of a point neighborhood is estimated based on the second moment matrix. Our scale invariant detector computes a multi-scale representation for the Harris interest point detector and then selects points at which a local measure (the Laplacian) is maximal over scales. This provides a set of distinctive points which are invariant to scale, rotation and translation as well as robust to illumination changes and limited changes of viewpoint. The characteristic scale determines a scale invariant region for each point. We extend the scale invariant detector to affine invariance by estimating the affine shape of a point neighborhood. An iterative algorithm modifies location, scale and neighborhood of each point and converges to affine invariant points. This method can deal with significant affine transformations including large scale changes. The characteristic scale and the affine shape of neighborhood determine an affine invariant region for each point. We present a comparative evaluation of different detectors and show that our approach provides better results than existing methods. The performance of our detector is also confirmed by excellent matching results; the image is described by a set of scale/affine invariant descriptors computed on the regions associated with our points.

56.9.2 Main points

56.10 Visual categorization with bags of keypoints [48]

56.10.1 Original Abstract

None

56.11 Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication [117]

56.11.1 Original Abstract

We present a method for learning nonlinear systems, echo state networks (ESNs). ESNs employ artificial recurrent neural networks in a way that has recently been proposed independently as a learning mechanism in biological brains. The learning method is computationally efficient and easy to use. On a benchmark task of predicting a chaotic time series, accuracy is improved by a factor of 2400 over previous techniques. The potential for engineering applications is illustrated by equalizing a communication channel, where the signal error rate is improved by two orders of magnitude.

56.12 Recognizing human actions: a local SVM approach [216]

56.12.1 Original Abstract

Local space-time features capture local events in video and can be adapted to the size, the frequency and the velocity of moving patterns. In this paper, we demonstrate how such features can be used for recognizing complex motion patterns. We construct video representations in terms of local space-time features and integrate such representations with SVM classification schemes for recognition. For the purpose of evaluation we introduce a new video database containing 2391 sequences of six human actions performed by 25 people in four different scenarios. The presented results of action recognition justify the proposed method and demonstrate its advantage compared to other relative approaches for action recognition.

56.13 Distinctive Image Features from Scale-Invariant Keypoints [161]

56.13.1 Original Abstract

None

56.13.2 Main points

56.14 Gaussian processes for machine learning. [218]

56.14.1 Original Abstract

Gaussian processes (GPs) are natural generalisations of multivariate Gaussian random variables to infinite (countably or continuous) index sets. GPs have been applied in a large number of fields to a diverse range of ends, and very many deep theoretical analyses of various properties are available. This paper gives an introduction to Gaussian processes on a fairly elementary level with special emphasis on characteristics relevant in machine learning. It draws explicit connections to branches such as spline smoothing models and support vector machines in which similar ideas have been investigated. Gaussian process models are routinely used to solve hard machine learning problems. They are attractive because of their flexible non-parametric nature and computational simplicity. Treated within a Bayesian framework, very powerful statistical methods can be implemented which offer valid estimates of uncertainties in our predictions and generic model selection procedures cast as nonlinear optimization problems. Their main drawback of heavy computational scaling has recently been alleviated by the introduction of generic sparse approximations.13,78,31 The mathematical literature on GPs is large and often uses deep concepts which are not required to fully understand most machine learning applications. In this tutorial paper, we aim to present characteristics of GPs relevant to machine learning and to show up precise connections to other "kernel machines" popular in the community. Our focus is on a simple presentation, but references to more detailed sources are provided.

57 2005

57.1 Computers and Commerce: A Study of Technology and Management at Eckert-Mauchly Computer Company, Engineering Research Associates, and Remington Rand, 1946 – 1957 [188]

57.1.1 Original Abstract

Between 1946 and 1957 computing went from a preliminary, developmental stage to more widespread use accompanied by the beginnings of the digital computer industry. During this crucial decade, spurred by rapid technological advances, the computer enterprise became a major phenomenon. In Computers and Commerce, Arthur Norberg explores the importance of these years in the history of computing by focusing on technical developments and business strategies at two important firms, both established in 1946, Engineering Research Associates (ERA) and Eckert-Mauchly Computer Company (EMCC), from their early activities through their acquisition by Remington Rand.Both ERA and EMCC had their roots in World War II, and in postwar years both firms received major funding from the United States government. Norberg analyzes the interaction between the two companies and the government and examines the impact of this institutional context on technological innovation. He assesses the technical contributions of such key company figures as J. Presper Eckert, John Mauchly, Grace Hopper, and William Norris, analyzing the importance of engineering knowledge in converting theoretical designs into workable machines. Norberg looks at the two firms' operations after 1951 as independent subsidiaries of Remington Rand, and documents the management problems that began after Remington Rand merged with Sperry Gyroscope to form Sperry Rand in 1955.

57.2 A sparse texture representation using local affine regions [138]

57.2.1 Original Abstract

This paper introduces a texture representation suitable for recognizing images of textured surfaces under a wide range of transformations, including viewpoint changes and nonrigid deformations. At the feature extraction stage,

a sparse set of affine Harris and Laplacian regions is found in the image. Each of these regions can be thought of as a texture element having a characteristic elliptic shape and a distinctive appearance pattern. This pattern is captured in an affine-invariant fashion via a process of shape normalization followed by the computation of two novel descriptors, the spin image and the RIFT descriptor. When affine invariance is not required, the original elliptical shape serves as an additional discriminative feature for texture recognition. The proposed approach is evaluated in retrieval and classification tasks using the entire Brodatz database and a publicly available collection of 1,000 photographs of textured surfaces taken from different viewpoints.

57.3 A performance evaluation of local descriptors [171]

57.3.1 Original Abstract

In this paper, we compare the performance of descriptors computed for local interest regions, as, for example, extracted by the Harris-Affine detector [Mikolajczyk, K and Schmid, C, 2004]. Many different descriptors have been proposed in the literature. It is unclear which descriptors are more appropriate and how their performance depends on the interest region detector. The descriptors should be distinctive and at the same time robust to changes in viewing conditions as well as to errors of the detector. Our evaluation uses as criterion recall with respect to precision and is carried out for different image transformations. We compare shape context [Belongie, S, et al., April 2002], steerable filters [Freeman, W and Adelson, E, Setp. 1991], PCA-SIFT [Ke, Y and Sukthankar, R, 2004], differential invariants [Koenderink, J and van Doorn, A, 1987], spin images [Lazebnik, S, et al., 2003], SIFT [Lowe, D. G., 1999], complex filters [Schaffalitzky, F and Zisserman, A, 2002], moment invariants [Van Gool, L, et al., 1996], and cross-correlation for different types of interest regions. We also propose an extension of the SIFT descriptor and show that it outperforms the original method. Furthermore, we observe that the ranking of the descriptors is mostly independent of the interest region detector and that the SIFT-based descriptors perform best. Moments and steerable filters show the best performance among the low dimensional descriptors.

57.3.2 Main points

57.4 A comparison of affine region detectors [172]

57.4.1 Original Abstract

The paper gives a snapshot of the state of the art in affine covariant region detectors, and compares their performance on a set of test images under varying imaging conditions. Six types of detectors are included: detectors based on affine normalization around Harris (Mikolajczyk and Schmid, 2002; Schaffalitzky and Zisserman, 2002) and Hessian points (Mikolajczyk and Schmid, 2002), a detector of 'maximally stable extremal regions', proposed by Matas et al. (2002); an edge-based region detector (Tuytelaars and Van Gool, 1999) and a detector based on intensity extrema (Tuytelaars and Van Gool, 2000), and a detector of 'salient regions', proposed by Kadir, Zisserman and Brady (2004). The performance is measured against changes in viewpoint, scale, illumination, defocus and image compression. The objective of this paper is also to establish a reference test set of images and performance software, so that future detectors can be evaluated in the same framework.

57.5 Learning a similarity metric discriminatively, with application to face verification [41]

57.5.1 Original Abstract

We present a method for training a similarity metric from data. The method can be used for recognition or verification applications where the number of categories is very large and not known during training, and where the number of training samples for a single category is very small. The idea is to learn a function that maps input patterns into a target space such that the L1 norm in the target space approximates the "semantic" distance in the input space. The method is applied to a face verification task. The learning process minimizes a discriminative loss function that drives the similarity metric to be small for pairs of faces from the same person, and large for pairs from different persons. The mapping from raw to the target space is a convolutional network whose architecture is designed for robustness to geometric distortions. The system is tested on the Purdue/AR face database which has a very high degree of variability in the pose, lighting, expression, position, and artificial occlusions such as dark glasses and obscuring scarves.

57.6 Local features for object class recognition [170]

57.6.1 Original Abstract

In this paper, we compare the performance of local detectors and descriptors in the context of object class recognition. Recently, many detectors/descriptors have been evaluated in the context of matching as well as invariance to viewpoint changes (Mikolajczyk and Schmid, 2004). However, it is unclear if these results can be generalized to categorization problems, which require different properties of features. We evaluate 5 state-of-the-art scale invariant region detectors and 5 descriptors. Local features are computed for 20 object classes and clustered using hierarchical agglomerative clustering. We measure the quality of appearance clusters and location distributions using entropy as well as precision. We also measure how the clusters generalize from training set to novel test data. Our results indicate that attended SIFT descriptors (Mikolajczyk and Schmid, 2005) computed on Hessian-Laplace regions perform best. Second score is obtained by salient regions (Kadir and Brady, 2001). The results also show that these two detectors provide complementary features. The new detectors/descriptors significantly improve the performance of a state-of-the art recognition approach (Leibe, et al., 2005) in pedestrian detection task

57.6.2 Main points

57.7 Rank, trace-norm and max-norm [232]

57.7.1 Original Abstract

We study the rank, trace-norm and max-norm as complexity measures of matrices, focusing on the problem of fitting a matrix with matrices having low complexity. We present generalization error bounds for predicting unobserved entries that are based on these measures. We also consider the possible relations between these measures. We show gaps between them, and bounds on the extent of such gaps.

57.8 On contrastive divergence learning [36]

57.8.1 Original Abstract

Maximum-likelihood(ML) learning of Markov random fields is challenging because it requires estimates of averages that have anexponential number of terms. Markov chain Monte Carlo methods typically take a long time to converge on unbiased estimates, but Hinton (2002) showed that if the Markovchain is only run for a few steps, the learning can still work well and it approximately minimizes a different function called "contrastive divergence"(CD). CD learning has been successfully applied to various types of random fields. Here, we study the properties of CD learning and show that it provides biased estimates in general, but that the bias is typically very small. Fast CD learning can therefore be used to get close to an ML solution and slow ML learning can then be used to fine-tune the CD solution.

57.8.2 Main points

cited: 193 (01/06/2014)

57.9 Toward automatic phenotyping of developing embryos from videos. [187]

57.9.1 Original Abstract

We describe a trainable system for analyzing videos of developing C. elegans embryos. The system automatically detects, segments, and locates cells and nuclei in microscopic images. The system was designed as the central component of a fully automated phenotyping system. The system contains three modules 1) a convolutional network trained to classify each pixel into five categories: cell wall, cytoplasm, nucleus membrane, nucleus, outside medium; 2) an energy-based model, which cleans up the output of the convolutional network by learning local consistency constraints that must be satisfied by label images; 3) a set of elastic models of the embryo at various stages of development that are matched to the label images.

57.9.2 Main points

57.10 Object Recognition with Features Inspired by Visual Cortex [221]

57.10.1 Original Abstract

We introduce a novel set of features for robust object recognition. Each element of this set is a complex feature obtained by combining position- and scale-tolerant edge-detectors over neighboring positions and multiple orientations. Our system's architecture is motivated by a quantitative model of visual cortex. We show that our approach exhibits excellent recognition performance and outperforms several state-of-the-art systems on a variety of image datasets including many different object categories. We also demonstrate that our system is able to learn from very few examples. The performance of the approach constitutes a suggestive plausibility proof for a class of feedforward models of object recognition in cortex.

57.11 Skin segmentation using color pixel classification: analysis and comparison [199]

57.11.1 Original Abstract

This work presents a study of three important issues of the color pixel classification approach to skin segmentation: color representation, color quantization, and classification algorithm. Our analysis of several representative color spaces using the Bayesian classifier with the histogram technique shows that skin segmentation based on color pixel classification is largely unaffected by the choice of the color space. However, segmentation performance degrades when only chrominance channels are used in classification. Furthermore, we find that color quantization can be as low as 64 bins per channel, although higher histogram sizes give better segmentation performance. The Bayesian classifier with the histogram technique and the multilayer perceptron classifier are found to perform better compared to other tested classifiers, including three piecewise linear classifiers, three unimodal Gaussian classifiers, and a Gaussian mixture classifier.

57.11.2 Main points

57.12 Histograms of oriented gradients for human detection [51]

57.12.1 Original Abstract

We study the question of feature sets for robust visual object recognition; adopting linear SVM based human detection as a test case. After reviewing existing edge and gradient based descriptors, we show experimentally that grids of histograms of oriented gradient (HOG) descriptors significantly outperform existing feature sets for human detection. We study the influence of each stage of the computation on performance, concluding that fine-scale gradients, fine orientation binning, relatively coarse spatial binning, and high-quality local contrast normalization in overlapping descriptor blocks are all important for good results. The new approach gives near-perfect separation on the original MIT pedestrian database, so we introduce a more challenging dataset containing over 1800 annotated human images with a large range of pose variations and backgrounds.

57.12.2 Main points

58 2006

58.1 The legacy of John von Neumann [82]

58.1.1 Original Abstract

The ideas of John von Neumann have had a profound influence on modern mathematics and science. One of the great thinkers of our century, von Neumann initiated major branches of mathematics–from operator algebras to game theory to scientific computing–and had a fundamental impact on such areas as self-adjoint operators, ergodic theory and the foundations of quantum mechanics, and numerical analysis and the design of the modern computer. This volume contains the proceedings of an AMS Symposium in Pure Mathematics, held at Hofstra University, in May 1988. The symposium brought together some of the foremost researchers in the wide range of areas in which von Neumann worked. These articles illustrate the sweep of von Neumann's ideas and thinking and document their influence on contemporary

mathematics. In addition, some of those who knew von Neumann when he was alive have presented here personal reminiscences about him. This book is directed to those interested in operator theory, game theory, ergodic theory, and scientific computing, as well as to historians of mathematics and others having an interest in the contemporary history of the mathematical sciences. This book will give readers an appreciation for the workings of the mind of one of the mathematical giants of our time.

58.2 Philosophy of Psychology and Cognitive Science: A Volume of the Handbook of the Philosophy of Science Series [75]

58.2.1 Original Abstract

ELSEVIER SALE - ALL SCIENCE AND TECHNOLOGY BOOKS 50

58.3 Mind as machine: A history of cognitive science [27]

58.3.1 Original Abstract

The development of cognitive science is one of the most remarkable and fascinating intellectual achievements of the modern era. The quest to understand the mind is as old as recorded human thought; but the progress of modern science has offered new methods and techniques which have revolutionized this enquiry. Oxford University Press now presents a masterful history of cognitive science, told by one of its most eminent practitioners. Cognitive science is the project of understanding the mind by modeling its workings. Psychology is its heart, but it draws together various adjoining fields of research, including artificial intelligence; neuroscientific study of the brain; philosophical investigation of mind, language, logic, and understanding; computational work on logic and reasoning; linguistic research on grammar, semantics, and communication; and anthropological explorations of human similarities and differences. Each discipline, in its own way, asks what the mind is, what it does, how it works, how it developed - how it is even possible. The key distinguishing characteristic of cognitive science, Boden suggests, compared with older ways of thinking about the mind, is the notion of understanding the mind as a kind of machine. She traces the origins of cognitive science back to

Descartes's revolutionary ideas, and follows the story through the eighteenth and nineteenth centuries, when the pioneers of psychology and computing appear. Then she guides the reader through the complex interlinked paths along which the study of the mind developed in the twentieth century. Cognitive science, in Boden's broad conception, covers a wide range of aspects of mind: not just 'cognition' in the sense of knowledge or reasoning, but emotion, personality, social communication, and even action. In each area of investigation, Boden introduces the key ideas and the people who developed them. No one else could tell this story as Boden can: she has been an active participant in cognitive science since the 1960s, and has known many of the key figures personally. Her narrative is written in a lively, swift-moving style, enriched by the personal touch of someone who knows the story at first hand. Her history looks forward as well as back: it is her conviction that cognitive science today–and tomorrow–cannot be properly understood without a historical perspective. Mind as Machine will be a rich resource for anyone working on the mind, in any academic discipline, who wants to know how our understanding of our mental activities and capacities has developed.

58.3.2 Main points

58.4 Pattern recognition and machine learning. [26]

58.4.1 Original Abstract

None

58.4.2 Main points

58.5 A convolutional neural network approach for objective video quality assessment [35]

58.5.1 Original Abstract

This paper describes an application of neural networks in the field of objective measurement method designed to automatically assess the perceived quality of digital videos. This challenging issue aims to emulate human judgment and to replace very complex and time consuming subjective quality assessment. Several metrics have been proposed in literature to tackle this issue. They are based on a general framework that combines different stages, each of them

addressing complex problems. The ambition of this paper is not to present a global perfect quality metric but rather to focus on an original way to use neural networks in such a framework in the context of reduced reference (RR) quality metric. Especially, we point out the interest of such a tool for combining features and pooling them in order to compute quality scores. The proposed approach solves some problems inherent to objective metrics that should predict subjective quality score obtained using the single stimulus continuous quality evaluation (SSCQE) method. This latter has been adopted by video quality expert group (VQEG) in its recently finalized reduced referenced and no reference (RRNR-TV) test plan. The originality of such approach compared to previous attempts to use neural networks for quality assessment, relies on the use of a convolutional neural network (CNN) that allows a continuous time scoring of the video. Objective features are extracted on a frame-by-frame basis on both the reference and the distorted sequences; they are derived from a perceptual-based representation and integrated along the temporal axis using a time-delay neural network (TDNN). Experiments conducted on different MPEG-2 videos, with bit rates ranging 2-6 Mb/s, show the effectiveness of the proposed approach to get a plausible model of temporal pooling from the human vision system (HVS) point of view. More specifically, a linear correlation criteria, between objective and subjective scoring, up to 0.92 has been obtained on a- - set of typical TV videos

58.6 Extreme learning machine: Theory and applications [111]

58.6.1 Original Abstract

None

58.7 A fast learning algorithm for deep belief nets [101]

58.7.1 Original Abstract

None

58.7.2 Main points

58.8 Reducing the dimensionality of data with neural networks [103]

58.8.1 Original Abstract

High-dimensional data can be converted to low-dimensional codes by training a multilayer neural network with a small central layer to reconstruct high-dimensional input vectors. Gradient descent can be used for fine-tuning the weights in such "autoencoder" networks, but this works well only if the initial weights are close to a good solution. We describe an effective way of initializing the weights that allows deep autoencoder networks to learn low-dimensional codes that work much better than principal components analysis as a tool to reduce the dimensionality of data.

58.9 A fast learning algorithm for deep belief nets [102]

58.9.1 Original Abstract

We show how to use "complementary priors" to eliminate the explaining-away effects thatmake inference difficult in densely connected belief nets that have many hidden layers. Using complementary priors, we derive a fast, greedy algorithm that can learn deep, directed belief networks one layer at a time, provided the top two layers form an undirected associative memory. The fast, greedy algorithm is used to initialize a slower learning procedure that fine-tunes the weights using a contrastive version of thewake-sleep algorithm. After fine-tuning, a networkwith three hidden layers forms a very good generative model of the joint distribution of handwritten digit images and their labels. This generative model gives better digit classification than the best discriminative learning algorithms. The low-dimensional manifolds on which the digits lie are modeled by long ravines in the free-energy landscape of the top-level associative memory, and it is easy to explore these ravines by using the directed connections to displaywhat the associativememory has in mind.

58.9.2 Main points

58.10 Surf: Speeded up robust features [19]

58.10.1 Original Abstract

None

59 2007

59.1 The mathematical biophysics of Nicolas Rashevsky [49]

59.1.1 Original Abstract

N. Rashevsky (1899–1972) was one of the pioneers in the application of mathematics to biology. With the slogan: mathematical biophysics : biology :: mathematical physics : physics, he proposed the creation of a quantitative theoretical biology. Here, we will give a brief biography, and consider Rashevsky's contributions to mathematical biology including neural nets and relational biology. We conclude that Rashevsky was an important figure in the introduction of quantitative models and methods into biology.

59.2 Classifier fusion for SVM-based multimedia semantic indexing [14]

59.2.1 Original Abstract

None

59.2.2 Main points

59.3 Local features and kernels for classification of texture and object categories: A comprehensive study [273]

59.3.1 Original Abstract

Recently, methods based on local image features have shown promise for texture and object recognition tasks. This paper presents a large-scale eval-

uation of an approach that represents images as distributions (signatures or histograms) of features extracted from a sparse set of keypoint locations and learns a Support Vector Machine classifier with kernels based on two effective measures for comparing distributions, the Earth Mover's Distance and the 2 distance. We first evaluate the performance of our approach with different keypoint detectors and descriptors, as well as different kernels and classifiers. We then conduct a comparative evaluation with several state-of-the-art recognition methods on four texture and five object databases. On most of these databases, our implementation exceeds the best reported results and achieves comparable performance on the rest. Finally, we investigate the influence of background correlations on recognition performance via extensive tests on the PASCAL database, for which ground-truth object localization information is available. Our experiments demonstrate that image representations based on distributions of local features are surprisingly effective for classification of texture and object images under challenging real-world conditions, including significant intra-class variations and substantial background clutter.

59.4 Human action recognition using a modified convolutional neural network [128]

59.4.1 Original Abstract

In this paper, a human action recognition method using a hybrid neural network is presented. The method consists of three stages: preprocessing, feature extraction, and pattern classification. For feature extraction, we propose a modified convolutional neural network (CNN) which has a three-dimensional receptive field. The CNN generates a set of feature maps from the action descriptors which are derived from a spatiotemporal volume. A weighted fuzzy min-max (WFMM) neural network is used for the pattern classification stage. We introduce a feature selection technique using the WFMM model to reduce the dimensionality of the feature space. Two kinds of relevance factors between features and pattern classes are defined to analyze the salient features.

59.5 Scaling learning algorithms towards AI [22]

59.5.1 Original Abstract

One long-term goal of machine learning research is to produce methods thatare applicable to highly complex tasks, such as perception (vision, audition), reasoning, intelligent control, and other artificially intelligent behaviors. We argue that in order to progress toward this goal, the Machine Learning community mustendeavor to discover algorithms that can learn highly complex functions, with minimal need for prior knowledge, and with minimal human intervention. We present mathematical and empirical evidence suggesting that many popular approaches to non-parametric learning, particularly kernel methods, are fundamentally limited in their ability to learn complex high-dimensional functions. Our analysis focuses on two problems. First, kernel machines are shallow architectures, in which one large layer of simple template matchers is followed by a single layer of trainable coefficients. We argue that shallow architectures can be very inefficient in terms of required number of computational elements and examples. Second, we analyze a limitation of kernel machines with a local kernel, linked to the curse of dimensionality, that applies to supervised, unsupervised (manifold learning) and semi-supervised kernel machines. Using empirical results on invariant image recognition tasks, kernel methods are compared with deep architectures, in which lower-level features or concepts are progressively combined into more abstract and higher-level representations. We argue that deep architectures have the potential to generalize in non-local ways, i.e., beyond immediate neighbors, and that this is crucial in order to make progress on the kind of complex tasks required for artificial intelligence

59.5.2 Main points

<m:note/>

59.6 An empirical evaluation of deep architectures on problems with many factors of variation [137]

59.6.1 Original Abstract

Recently, several learning algorithms relying on models with deep architectures have been proposed. Though they have demonstrated impressive perfor-

mance, to date, they have only been evaluated on relatively simple problems such as digit recognition in a controlled environment, for which many machine learning algorithms already report reasonable results. Here, we present a series of experiments which indicate that these models show promise in solving harder learning problems that exhibit many factors of variation. These models are compared with well-established algorithms such as Support Vector Machines and single hidden-layer feed-forward neural networks.

59.6.2 Main points

<m:note/>

59.7 Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition [201]

59.7.1 Original Abstract

We present an unsupervised method for learning a hierarchy of sparse feature detectors that are invariant to small shifts and distortions. The resulting feature extractor consists of multiple convolution filters, followed by a feature-pooling layer that computes the max of each filter output within adjacent windows, and a point-wise sigmoid non-linearity. A second level of larger and more invariant features is obtained by training the same algorithm on patches of features from the first level. Training a supervised classifier on these features yields 0.64

59.8 Robust object recognition with cortex-like mechanisms. [222]

59.8.1 Original Abstract

We introduce a new general framework for the recognition of complex visual scenes, which is motivated by biology: We describe a hierarchical system that closely follows the organization of visual cortex and builds an increasingly complex and invariant feature representation by alternating between a template matching and a maximum pooling operation. We demonstrate the strength of the approach on a range of recognition tasks: From invariant

single object recognition in clutter to multiclass categorization problems and complex scene understanding tasks that rely on the recognition of both shape-based as well as texture-based objects. Given the biological constraints that the system had to satisfy, the approach performs surprisingly well: It has the capability of learning from only a few training examples and competes with state-of-the-art systems. We also discuss the existence of a universal, redundant dictionary of features that could handle the recognition of most object categories. In addition to its relevance for computer vision, the success of this approach suggests a plausibility proof for a class of feedforward models of object recognition in cortex.

59.9 To recognize shapes, first learn to generate images [105]

59.9.1 Original Abstract

The uniformity of the cortical architecture and the ability of functions to move to different areas of cortex following early damage strongly suggest that there is a single basic learning algorithm for extracting underlying structure from richly structured, high-dimensional sensory data. There have been many attempts to design such an algorithm, but until recently they all suffered from serious computational weaknesses. This chapter describes several of the proposed algorithms and shows how they can be combined to produce hybrid methods that work efficiently in networks with many layers and millions of adaptive connections.

59.9.2 Main points

60 2008

60.1 Connectionism: A Hands-on Approach [53]

60.1.1 Original Abstract

"Connectionism" is a "hands on" introduction to connectionist modeling through practical exercises in different types of connectionist architectures. explores three different types of connectionist architectures - distributed associative memory, perceptron, and multilayer perceptron provides a brief overview

of each architecture, a detailed introduction on how to use a program to explore this network, and a series of practical exercises that are designed to highlight the advantages, and disadvantages, of each accompanied by a website at http: //www.bcp.psych.ualberta.ca/ mike/Book3/ that includes practice exercises and software, as well as the files and blank exercise sheets required for performing the exercises designed to be used as a stand-alone volume or alongside "Minds and Machines: Connectionism and Psychological Modeling" (by Michael R.W. Dawson, Blackwell 2004)

60.1.2 Main points

60.2 The matrix cookbook [198]

60.2.1 Original Abstract

None

60.3 Learning realistic human actions from movies [136]

60.3.1 Original Abstract

The aim of this paper is to address recognition of natural human actions in diverse and realistic video settings. This challenging but important subject has mostly been ignored in the past due to several problems one of which is the lack of realistic and annotated video datasets. Our first contribution is to address this limitation and to investigate the use of movie scripts for automatic annotation of human actions in videos. We evaluate alternative methods for action retrieval from scripts and show benefits of a text-based classifier. Using the retrieved action samples for visual learning, we next turn to the problem of action classification in video. We present a new method for video classification that builds upon and extends several recent ideas including local space-time features, space-time pyramids and multi-channel non-linear SVMs. The method is shown to improve state-of-the-art results on the standard KTH action dataset by achieving 91.8

60.4 Representational power of restricted boltzmann machines and deep belief networks. [143]

60.4.1 Original Abstract

Deep belief networks (DBN) are generative neural network models with many layers of hidden explanatory factors, recently introduced by Hinton, Osindero, and Teh (2006) along with a greedy layer-wise unsupervised learning algorithm. The building block of a DBN is a probabilistic model called a restricted Boltzmann machine (RBM), used to represent one layer of the model. Restricted Boltzmann machines are interesting because inference is easy in them and because they have been successfully used as building blocks for training deeper models. We first prove that adding hidden units yields strictly improved modeling power, while a second theorem shows that RBMs are universal approximators of discrete distributions. We then study the question of whether DBNs with more layers are strictly more powerful in terms of representational power. This suggests a new and less greedy criterion for training RBMs within DBNs.

60.4.2 Main points

<m:note/>

60.5 Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words [185]

60.5.1 Original Abstract

We present a novel unsupervised learning method for human action categories. A video sequence is represented as a collection of spatial-temporal words by extracting space-time interest points. The algorithm automatically learns the probability distributions of the spatial-temporal words and the intermediate topics corresponding to human action categories. This is achieved by using latent topic models such as the probabilistic Latent Semantic Analysis (pLSA) model and Latent Dirichlet Allocation (LDA). Our approach can handle noisy feature points arisen from dynamic background and moving cameras due to the application of the probabilistic models. Given a novel video sequence, the algorithm can categorize and localize the human action(s) contained in the video. We test our algorithm on three challenging datasets:

the KTH human motion dataset, the Weizmann human action dataset, and a recent dataset of figure skating actions. Our results reflect the promise of such a simple approach. In addition, our algorithm can recognize and localize multiple actions in long and complex video sequences containing multiple motions.

60.6 Action snippets: How many frames does human action recognition require? [214]

60.6.1 Original Abstract

Visual recognition of human actions in video clips has been an active field of research in recent years. However, most published methods either analyse an entire video and assign it a single action label, or use relatively large look-ahead to classify each frame. Contrary to these strategies, human vision proves that simple actions can be recognised almost instantaneously. In this paper, we present a system for action recognition from very short sequences (ldquosnippetsrdquo) of 1-10 frames, and systematically evaluate it on standard data sets. It turns out that even local shape and optic flow for a single frame are enough to achieve ap90

60.7 Deep learning via semi-supervised embedding [260]

60.7.1 Original Abstract

We show how nonlinear semi-supervised embedding algorithms popular for use with "shallow" learning techniques such as kernel methods can be easily applied to deep multi-layer architectures, either as a regularizer at the output layer, or on each layer of the architecture. Compared to standard supervised backpropagation this can give significant gains. This trick provides a simple alternative to existing approaches to semi-supervised deep learning whilst yielding competitive error rates compared to those methods, and existing shallow semi-supervised techniques.

60.7.2 Main points

60.8 Speeded-up robust features (SURF) [18]

60.8.1 Original Abstract

This article presents a novel scale- and rotation-invariant detector and descriptor, coined SURF (Speeded-Up Robust Features). SURF approximates or even outperforms previously proposed schemes with respect to repeatability, distinctiveness, and robustness, yet can be computed and compared much faster. This is achieved by relying on integral images for image convolutions; by building on the strengths of the leading existing detectors and descriptors (specifically, using a Hessian matrix-based measure for the detector, and a distribution-based descriptor); and by simplifying these methods to the essential. This leads to a combination of novel detection, description, and matching steps. The paper encompasses a detailed description of the detector and descriptor and then explores the effects of the most important parameters. We conclude the article with SURF's application to two challenging, yet converse goals: camera calibration as a special case of image registration, and object recognition. Our experiments underline SURF's usefulness in a broad range of topics in computer vision.

60.8.2 Main points

61 2009

61.1 Evaluation of local spatio-temporal features for action recognition [258]

61.1.1 Original Abstract

Local space-time features have recently become a popular video representation for action recognition. Several methods for feature localization and description have been proposed in the literature and promising recognition results were demonstrated for a number of action classes. The comparison of existing methods, however, is often limited given the different experimental settings used. The purpose of this paper is to evaluate and compare previously proposed space-time features in a common experimental setup. In particular, we consider four different feature detectors and six local feature descriptors

and use a standard bag-of-features SVM approach for action recognition. We investigate the performance of these methods on a total of 25 action classes distributed over three datasets with varying difficulty. Among interesting conclusions, we demonstrate that regular sampling of space-time features consistently outperforms all tested space-time interest point detectors for human actions in realistic settings. We also demonstrate a consistent ranking for the majority of methods over different datasets and discuss their advantages and limitations.

61.1.2 Main points

- Detectors
  - Harris3D
  - Cuboid
  - Hessian
  - Dense sampling
- Descriptors
  - HOG/HOF
  - HOG3D
  - ESURF (extended SURF)
- Datasets
  - KTH actions
    - 6 human action classes
    - walking, jogging, running, boxing, waving and clapping
    - 25 subjects
    - 4 scenarios
    - 2391 video samples
    - `http://www.nada.kth.se/cvap/actions/`
  - UCF sport actions
    - 10 human action classes

    - winging, diving, kicking, weight-lifting, horse-riding, running, skateboarding, swinging, golf swinging and walking
    - 150 video samples
    - `http://crcv.ucf.edu/data/UCF_Sports_Action.php`
  - Hollywood2 actions
    - 12 action classes
    - answering the hone, driving car, eating, fighting, geting out of the car, hand shaking, hugging, kissing, running, sitting down, sitting up, and standing up.
    - 69 Hollywood movies
    - 1707 video samples
    - `http://www.di.ens.fr/~laptev/actions/hollywood2/`

61.2 Evaluation of local spatio-temporal features for action recognition [?]

61.2.1 Original Abstract

Local space-time features have recently become a popular video representation for action recognition. Several methods for feature localization and description have been proposed in the literature and promising recognition results were demonstrated for a number of action classes. The comparison of existing methods, however, is often limited given the different experimental settings used. The purpose of this paper is to evaluate and compare previously proposed space-time features in a common experimental setup. In particular, we consider four different feature detectors and six local feature descriptors and use a standard bag-of-features SVM approach for action recognition. We investigate the performance of these methods on a total of 25 action classes distributed over three datasets with varying difficulty. Among interesting conclusions, we demonstrate that regular sampling of space-time features consistently outperforms all tested space-time interest point detectors for human actions in realistic settings. We also demonstrate a consistent ranking for the majority of methods over different datasets and discuss their advantages and limitations.

61.2.2 Main points

<m:note/>

61.3 Actions in context [163]

61.3.1 Original Abstract

This paper exploits the context of natural dynamic scenes for human action recognition in video. Human actions are frequently constrained by the purpose and the physical properties of scenes and demonstrate high correlation with particular scene classes. For example, eating often happens in a kitchen while running is more common outdoors. The contribution of this paper is three-fold: (a) we automatically discover relevant scene classes and their correlation with human actions, (b) we show how to learn selected scene classes from video without manual supervision and (c) we develop a joint framework for action and scene recognition and demonstrate improved recognition of both in natural video. We use movie scripts as a means of automatic supervision for training. For selected action classes we identify correlated scene classes in text and then retrieve video samples of actions and scenes for training using script-to-video alignment. Our visual models for scenes and actions are formulated within the bag-of-features framework and are combined in a joint scene-action SVM-based classifier. We report experimental results and validate the method on a new large dataset with twelve action classes and ten scene classes acquired from 69 movies.

61.4 Computational Intelligence: The Legacy of Alan Turing and John von Neumann [177]

61.4.1 Original Abstract

In this chapter fundamental problems of collaborative computational intelligence are discussed. The problems are distilled from the seminal research of Alan Turing and John von Neumann. For Turing the creation of machines with human-like intelligence was only a question of programming time. In his research he identified the most relevant problems concerning evolutionary computation, learning, and structure of an artificial brain. Many problems are still unsolved, especially efficient higher learning methods which Turing

called initiative. Von Neumann was more cautious. He doubted that human-like intelligent behavior could be described unambiguously in finite time and finite space. Von Neumann focused on self-reproducing automata to create more complex systems out of simpler ones. An early proposal from John Holland is analyzed. It centers on adaptability and population of programs. The early research of Newell, Shaw, and Simon is discussed. They use the logical calculus to discover proofs in logic. Only a few recent research projects have the broad perspectives and the ambitious goals of Turing and von Neumann. As examples the projects Cyc, Cog, and JANUS are discussed.

61.5 Natural Image Statistics [116]

61.5.1 Original Abstract

None

61.6 A Novel Connectionist System for Unconstrained Handwriting Recognition [88]

61.6.1 Original Abstract

Recognizing lines of unconstrained handwritten text is a challenging task. The difficulty of segmenting cursive or overlapping characters, combined with the need to exploit surrounding context, has led to low recognition rates for even the best current recognizers. Most recent progress in the field has been made either through improved preprocessing or through advances in language modeling. Relatively little work has been done on the basic recognition algorithms. Indeed, most systems rely on the same hidden Markov models that have been used for decades in speech and handwriting recognition, despite their well-known shortcomings. This paper proposes an alternative approach based on a novel type of recurrent neural network, specifically designed for sequence labeling tasks where the data is hard to segment and contains long-range bidirectional interdependencies. In experiments on two large unconstrained handwriting databases, our approach achieves word recognition accuracies of 79.7 percent on online data and 74.1 percent on offline data, significantly outperforming a state-of-the-art HMM-based system. In addition, we demonstrate the network's robustness to lexicon size, measure the individual influence of its hidden layers, and analyze its use of context. Last,

we provide an in-depth discussion of the differences between the network and HMMs, suggesting reasons for the network's superior performance.

61.6.2 Main points

61.7 What is the best multi-stage architecture for object recognition? [118]

61.7.1 Original Abstract

In many recent object recognition systems, feature extraction stages are generally composed of a filter bank, a non-linear transformation, and some sort of feature pooling layer. Most systems use only one stage of feature extraction in which the filters are hard-wired, or two stages where the filters in one or both stages are learned in supervised or unsupervised mode. This paper addresses three questions: 1. How does the non-linearities that follow the filter banks influence the recognition accuracy? 2. does learning the filter banks in an unsupervised or supervised manner improve the performance over random filters or hardwired filters? 3. Is there any advantage to using an architecture with two stages of feature extraction, rather than one? We show that using non-linearities that include rectification and local contrast normalization is the single most important ingredient for good accuracy on object recognition benchmarks. We show that two stages of feature extraction yield better accuracy than one. Most surprisingly, we show that a two-stage system with random filters can yield almost 63

61.7.2 Main points

- 1. Differences in non-linearities
  - Rectifying non-linearity is the most important factor
    - The polarization does not seem important
    - Or the possible cancelations are counterproductive
- 2. unsupervised, supervised, random, and hardwired filters
  - Hardwired filters have the worst performance
  - Random filters achieve good performance

- 3. Deep vs shallow
  - Two stages are better than one
- Background
  - Common approach steps:
    - Feature extraction with some filter banks
      - oriented edges
      - gabor filters
    - non-linear operation on the original features
      - Quantization
      - winner-take-all
      - sparsification
      - normalization
      - point-wise saturation
    - pooling operation
      - max pooling
      - average pooling
      - histogramming
    - Classify with supervised method
  - Example:
    - SIFT
      - apply oriented edges to some region
      - determines dominant orientation
      - agregate different regions
  - Feature extraction
    - Gabor wavelets
    - SIFT
    - HoG
    - statistics of input data on natural images creates gabor-like filters
    - Random filters
    - learn the filters with gradient descent

- Method
  - Layers
    - Filter Bank Layer F_{CSG}
      - Convolution filter
      - Sigmoid/tanh non-linearity
      - Gain coefficients
    - Rectification Layer R_{abs}
    - Local Contrast Normalization Layer N
    - Average Pooling and Subscampling Layer P_A
    - Max Pooling and Subscampling Layer P_M
  - Architectures
    - $F_{CSG} - P_A$
    - $F_{CSG} - R_{abs} - P_A$
    - $F_{CSG} - R_{abs} - N - P_A$
    - $F_{CSG} - N$
  - Training protocols
    - Random Features and Supervised Classifier - R and RR
    - Unsupervised Features, Supervised Classifier - U and UU
    - Random Features, Global Supervised Refinement - R+ and R+R+
    - Unsupervised Feature, Global Supervised Refinement - U+ and U+U+
  - Generation of Unsupervised filters using Predictive Sparse Decomposition
- Results
  - Random filters achieve good performance
  - Supervised Refinement improves
  - Two stages are better than one
  - Unsupervised pretraining achieves better results, but in case of using rectification and normalization the improvement is about 1%

- Rectification is very important
- One stage + PMK SVM gives good results
- Using handmade Gabbor filters goot worst results than random filters

61.8 Unsupervised feature learning for audio classification using convolutional deep belief networks. [152]

61.8.1 Original Abstract

In recent years, deep learning approaches have gained significant interest as a way of building hierarchical representations from unlabeled data. However, to our knowledge, these deep learning approaches have not been extensively studied for auditory data. In this paper, we apply convolutional deep belief networks to audio data and empirically evaluate them on various audio classification tasks. In the case of speech data, we show that the learned features correspond to phones/phonemes. In addition, our feature representations learned from unlabeled audio data show very good performance for multiple audio classification tasks. We hope that this paper will inspire more research on deep learning approaches applied to a wide range of audio recognition tasks.

61.9 Actions in context [163]

61.9.1 Original Abstract

This paper exploits the context of natural dynamic scenes for human action recognition in video. Human actions are frequently constrained by the purpose and the physical properties of scenes and demonstrate high correlation with particular scene classes. For example, eating often happens in a kitchen while running is more common outdoors. The contribution of this paper is three-fold: (a) we automatically discover relevant scene classes and their correlation with human actions, (b) we show how to learn selected scene classes from video without manual supervision and (c) we develop a joint framework for action and scene recognition and demonstrate improved recognition of both in natural video. We use movie scripts as a means of automatic supervision for training. For selected action classes we identify correlated scene classes

in text and then retrieve video samples of actions and scenes for training using script-to-video alignment. Our visual models for scenes and actions are formulated within the bag-of-features framework and are combined in a joint scene-action SVM-based classifier. We report experimental results and validate the method on a new large dataset with twelve action classes and ten scene classes acquired from 69 movies.

61.10 Evaluation of local spatio-temporal features for action recognition [258]

61.10.1 Original Abstract

Local space-time features have recently become a popular video representation for action recognition. Several methods for feature localization and description have been proposed in the literature and promising recognition results were demonstrated for a number of action classes. The comparison of existing methods, however, is often limited given the different experimental settings used. The purpose of this paper is to evaluate and compare previously proposed space-time features in a common experimental setup. In particular, we consider four different feature detectors and six local feature descriptors and use a standard bag-of-features SVM approach for action recognition. We investigate the performance of these methods on a total of 25 action classes distributed over three datasets with varying difficulty. Among interesting conclusions, we demonstrate that regular sampling of space-time features consistently outperforms all tested space-time interest point detectors for human actions in realistic settings. We also demonstrate a consistent ranking for the majority of methods over different datasets and discuss their advantages and limitations.

61.10.2 Main points

- Detectors
  - Harris3D
  - Cuboid
  - Hessian
  - Dense sampling

- Descriptors
  - HOG/HOF
  - HOG3D
  - ESURF (extended SURF)
- Datasets
  - KTH actions
    - 6 human action classes
    - walking, jogging, running, boxing, waving and clapping
    - 25 subjects
    - 4 scenarios
    - 2391 video samples
    - `http://www.nada.kth.se/cvap/actions/`
  - UCF sport actions
    - 10 human action classes
    - winging, diving, kicking, weight-lifting, horse-riding, running, skateboarding, swinging, golf swinging and walking
    - 150 video samples
    - `http://crcv.ucf.edu/data/UCF_Sports_Action.php`
  - Hollywood2 actions
    - 12 action classes
    - answering the hone, driving car, eating, fighting, geting out of the car, hand shaking, hugging, kissing, running, sitting down, sitting up, and standing up.
    - 69 Hollywood movies
    - 1707 video samples
    - `http://www.di.ens.fr/~laptev/actions/hollywood2/`

61.11 Stacks of convolutional restricted Boltzmann machines for shift-invariant feature learning [189]

61.11.1 Original Abstract

In this paper we present a method for learning class-specific features for recognition. Recently a greedy layer-wise procedure was proposed to initialize

weights of deep belief networks, by viewing each layer as a separate restricted Boltzmann machine (RBM). We develop the convolutional RBM (C-RBM), a variant of the RBM model in which weights are shared to respect the spatial structure of images. This framework learns a set of features that can generate the images of a specific object class. Our feature extraction model is a four layer hierarchy of alternating filtering and maximum subsampling. We learn feature parameters of the first and third layers viewing them as separate C-RBMs. The outputs of our feature extraction hierarchy are then fed as input to a discriminative classifier. It is experimentally demonstrated that the extracted features are effective for object detection, using them to obtain performance comparable to the state of the art on handwritten digit recognition and pedestrian detection.

61.11.2 Main points

- New Convolutional Restricted Boltzmann Machine (C-RBM)
- Comparable state-of-the-art on handwritten digit recognition and pedestrian detection
- RBM
  - Probabilistic model
  - hidden variables independent given observerd data
  - Not capture explicitly spacial structure of images
- C-RBM
  - Include spatial locality and weight sharing
  - Favors filters with high response on training images
  - Unsupervised learning using Contrastive Divergence
  - Layerwise training for stacks of RBMs
  - Convolutional connections are employed in a generative Markov Random Field architecture
  - Hidden units divided into K feature maps
  - Convolution problems

        * Boundary units are withinb a smaller number of subwindows compared to the interior pixels
        * middle pixels may contribute to K_{xy} features
        * Separation of boundary variables (v^b) from middle variables (v^m)
        * Problems sampling from boundary pixels (not have nough features)
        * Over completeness because of K-features
        * Sampling creates images very similar to the original ones
        * Need of more Gibbs sampling steps
        * Their solution is to fix hidden bias terms c during training

- Multilayer C-RBMs
    - Subsampling takes maximum conditional feature probability over non-overlapping subwindows of feature maps
    - Architecture
        * discriminative layer (SVM)
        * max pooling
        * convolution
        * max pooling
        * convolution
        * input
    - On pedestrians also HOG is used in discriminative layer
- MNIST dataset
    - Discriminative layer with RBF kernel
    - 10 one-vs-rest binary SVMs
    - 1st layer 15 feature maps
    - 2nd layer 2x2 non-overlapping subwindos
    - 3rd layer 15 feature maps
    - 4th layer
- Comparison with Large CNN

  - C-RBM is better when training is small

- Pedestrian dataset
  - 1st layer 7x7 15 feature maps
  - 2nd layer 4x4 subsampling
  - 3rd layer 15x5x5 30 feature maps
  - 4th layer 2x2 subsampling
  - + HOG
  - Discriminative layer with linear kernel

61.12 Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations [151]

61.12.1 Original Abstract

There has been much interest in unsupervised learning of hierarchical generative models such as deep belief networks. Scaling such models to full-sized, high-dimensional images remains a difficult problem. To address this problem, we present the convolutional deep belief network, a hierarchical generative model which scales to realistic image sizes. This model is translation-invariant and supports efficient bottom-up and top-down probabilistic inference. Key to our approach is probabilistic max-pooling, a novel technique which shrinks the representations of higher layers in a probabilistically sound way. Our experiments show that the algorithm learns useful high-level visual features, such as object parts, from unlabeled images of objects and natural scenes. We demonstrate excellent performance on several visual recognition tasks and show that our model can perform hierarchical (bottom-up and top-down) inference over full-sized images.

61.12.2 Main points

- Probabilistic max-pooling
- Scale DBN to real-sized images
  - Computationally intractable

  - Need invariance in representation

- RBM
  - Binary valued: Independent Bernoulli random variables
  - Real valued: Gaussian with diagonal covariance
  - Training:
    * Stochastic gradient ascent on log-likelihood of training data
    * Contrastive divergence approximation

- Convolutional RBM
  - detection layers: convolving feature maps
  - pooling layers: shrink the representation
    * Block: CxC from bottom layer
    * Max-pooling : minimizes energy subject to only one unit can be active.
  - Sparsity regularization: hidden units have a mean activation close to a small constant

- Convolutional Deep belief network
  - Stacking CRBM on top of one another
  - Training:
    * Gibbs sampling
    * Mean-field (5 iterations in this paper)

61.13 Learning Deep Architectures for AI [21]

61.13.1 Original Abstract

Theoretical results suggest that in order to learn the kind of complicated functions that can represent high-level abstractions (e.g., in vision, language, and other AI-level tasks), one may need deep architectures. Deep architectures are composed of multiple levels of non-linear operations, such as in neural nets with many hidden layers or in complicated propositional formulae re-using many sub-formulae. Searching the parameter space of deep architectures is

a difficult task, but learning algorithms such as those for Deep Belief Networks have recently been proposed to tackle this problem with notable success, beating the state-of-the-art in certain areas. This monograph discusses the motivations and principles regarding learning algorithms for deep architectures, in particular those exploiting as building blocks unsupervised learning of single-layer models such as Restricted Boltzmann Machines, used to construct deeper models such as Deep Belief Networks.

61.14 Journal of Statistical Software [153]

61.14.1 Original Abstract

None

62 2010

62.1 Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules [64]

62.1.1 Original Abstract

In a mathematical approach to hypothesis tests, we start with a clearly defined set of hypotheses and choose the test with the best properties for those hypotheses. In practice, we often start with less precise hypotheses. For example, often a researcher wants to know which of two groups generally has the larger responses, and either a t-test or a Wilcoxon-Mann-Whitney (WMW) test could be acceptable. Although both t-tests and WMW tests are usually associated with quite different hypotheses, the decision rule and p-value from either test could be associated with many different sets of assumptions, which we call perspectives. It is useful to have many of the different perspectives to which a decision rule may be applied collected in one place, since each perspective allows a different interpretation of the associated p-value. Here we collect many such perspectives for the two-sample t-test, the WMW test and other related tests. We discuss validity and consistency under each perspective and discuss recommendations between the tests in light of these many different perspectives. Finally, we briefly discuss a decision rule for testing

genetic neutrality where knowledge of the many perspectives is vital to the proper interpretation of the decision rule.

62.1.2 Main points

62.2 High dynamic range imaging: acquisition, display, and image-based lighting [203]

62.2.1 Original Abstract

This landmark book is the first to describe HDRI technology in its entirety and covers a wide-range of topics, from capture devices to tone reproduction and image-based lighting. The techniques described enable you to produce images that have a dynamic range much closer to that found in the real world, leading to an unparalleled visual experience. As both an introduction to the field and an authoritative technical reference, it is essential to anyone working with images, whether in computer graphics, film, video, photography, or lighting design.New material includes chapters on High Dynamic Range Video Encoding, High Dynamic Range Image Encoding, and High Dynammic Range Display DevicesWritten by the inventors and initial implementors of High Dynamic Range ImagingCovers the basic concepts (including just enough about human vision to explain why HDR images are necessary), image capture, image encoding, file formats, display techniques, tone mapping for lower dynamic range display, and the use of HDR images and calculations in 3D renderingRange and depth of coverage is good for the knowledgeable researcher as well as those who are just starting to learn about High Dynamic Range imaging

62.3 Computer vision: algorithms and applications [241]

62.3.1 Original Abstract

Humans perceive the three-dimensional structure of the world with apparent ease. However, despite all of the recent advances in computer vision research, the dream of having a computer interpret an image at the same level as a ...

62.4 Computer Vision–ECCV 2010 [52]

62.4.1 Original Abstract

None

62.5 Tiled convolutional neural networks. [141]

62.5.1 Original Abstract

Convolutional neural networks (CNNs) have been successfully applied to many tasks such as digit and object recognition. Using convolutional (tied) weights significantly reduces the number of parameters that have to be learned, and also allows translational invariance to be hard-coded into the architecture. In this paper, we consider the problem of learning invariances, rather than relying on hardcoding. We propose tiled convolution neural networks (Tiled CNNs), which use a regular "tiled " pattern of tied weights that does not require that adjacent hidden units share identical weights, but instead requires only that hidden units k steps away from each other to have tied weights. By pooling over neighboring units, this architecture is able to learn complex invariances (such as scale and rotational invariance) beyond translational invariance. Further, it also enjoys much of CNNs' advantage of having a relatively small number of learned parameters (such as ease of learning and greater scalability). We provide an efficient learning algorithm for Tiled CNNs based on Topographic ICA, and show that learning complex invariant features allows us to achieve highly competitive results for both the NORB and CIFAR-10 datasets.

62.6 Convolutional Deep Belief Networks on CIFAR-10 [133]

62.6.1 Original Abstract

We describe how to train a two-layer convolutional Deep Belief Network (DBN) on the 1.6 million tiny imagesdataset.When training a convolutional DBN, one must decide what to do with the edge pixels of teh images. Asthe pixels near the edge of an image contribute to the fewest convolutional filter outputs, the model maysee it fit to tailor its few convolutional filters to better model the edge pixels. This is undesirable becaue itusually comes

at the expense of a good model for the interior parts of the image. We investigate several waysof dealing with the edge pixels when training a convolutional DBN. Using a combination of locally-connectedconvolutional units and globally-connected units, as well as a few tricks to reduce the effects of overfitting,we achieve state-of-the-art performance in the classification task of the CIFAR-10 subset of the tiny imagesdataset.

62.6.2 Main points

- Detectors
  - Harris3D
  - Cuboid
  - Hessian
  - Dense sampling
- Descriptors
  - HOG/HOF
  - HOG3D
  - ESURF (extended SURF)
- Datasets
  - KTH actions
    - 6 human action classes
    - walking, jogging, running, boxing, waving and clapping
    - 25 subjects
    - 4 scenarios
    - 2391 video samples
    - *http* : *//www.nada.kth.se/cvap/actions/*
  - UCF sport actions
    - 10 human action classes
    - winging, diving, kicking, weight-lifting, horse-riding, running, skateboarding, swinging, golf swinging and walking
    - 150 video samples

    - $http://crcv.ucf.edu/data/UCF_Sports_Action.php$
  - Hollywood2 actions
    - 12 action classes
    - answering the hone, driving car, eating, fighting, geting out of the car, hand shaking, hugging, kissing, running, sitting down, sitting up, and standing up.
    - 69 Hollywood movies
    - 1707 video samples
    - $http://www.di.ens.fr/\ laptev/actions/hollywood2/$

62.7 Convolutional learning of spatio-temporal features [243]

62.7.1 Original Abstract

We address the problem of learning good features for understanding video data. We introduce a model that learns latent representations of image sequences from pairs of successive images. The convolutional architecture of our model allows it to scale to realistic image sizes whilst using a compact parametrization. In experiments on the NORB dataset, we show our model extracts latent "flow fields" which correspond to the transformation between the pair of input frames. We also use our model to extract low-level motion features in a multi-stage architecture for action recognition, demonstrating competitive performance on both the KTH and Hollywood2 datasets.

62.7.2 Main points

62.8 Learning Convolutional Feature Hierarchies for Visual Recognition [125]

62.8.1 Original Abstract

We propose an unsupervised method for learning multi-stage hierarchies of sparseconvolutional features. While sparse coding has become an increasingly popularmethod for learning visual features, it is most often trained at the patch level.Applying the resulting filters convolutionally results in highly redundant codesbecause overlapping patches are encoded in isolation. By train-

ing convolutionallyover large image windows, our method reduces the redudancy between featurevectors at neighboring locations and improves the efficiency of the overall repre-sentation. In addition to a linear decoder that reconstructs the image from sparsefeatures, our method trains an efficient feed-forward encoder that predicts quasi-sparse features from the input. While patch-based training rarely produces any-thing but oriented edge detectors, we show that convolutional training produceshighly diverse filters, including center-surround filters, corner detectors, cross de-tectors, and oriented grating detectors. We show that using these filters in multi-stage convolutional network architecture improves performance on a number ofvisual recognition and detection tasks

62.9 Tiled convolutional neural networks [184]

62.9.1 Original Abstract

Convolutional neural networks (CNNs) have been successfully applied to manytasks such as digit and object recognition. Using convolutional (tied) weightssignificantly reduces the number of parameters that have to be learned, and alsoallows translational invariance to be hard-coded into the architecture. In this pa-per, we consider the problem of learning invariances, rather than relying on hard-coding. We propose tiled convolution neural networks (Tiled CNNs), which usea regular "tiled" pattern of tied weights that does not require that adjacent hiddenunits share identical weights, but instead requires only that hidden units k stepsaway from each other to have tied weights. By pooling over neighboring units,this architecture is able to learn complex invariances (such as scale and rotationalinvariance) beyond translational invariance. Further, it also enjoys much of CNNs'advantage of having a relatively small number of learned parameters (such as easeof learning and greater scalability). We provide an efficient learning algorithm forTiled CNNs based on Topographic ICA, and show that learning complex invariantfeatures allows us to achieve highly competitive results for both the NORB andCIFAR-10 datasets.

62.10 Why does unsupervised pre-training help deep learning? [58]

62.10.1 Original Abstract

Much recent research has been devoted to learning algorithms for deep architectures such as Deep Belief Networks and stacks of auto-encoder variants, with impressive results obtained in several areas, mostly on vision and language data sets. The best results obtained on supervised learning tasks involve an unsupervised learning component, usually in an unsupervised pre-training phase. Even though these new algorithms have enabled training deep models, many questions remain as to the nature of this difficult learning problem. The main question investigated here is the following: how does unsupervised pre-training work? Answering this questions is important if learning in deep architectures is to be further improved. We propose several explanatory hypotheses and test them through extensive simulations. We empirically show the influence of pre-training with respect to architecture depth, model capacity, and number of training examples. The experiments confirm and clarify the advantage of unsupervised pre-training. The results suggest that unsupervised pre-training guides the learning towards basins of attraction of minima that support better generalization from the training data set; the evidence from these results supports a regularization explanation for the effect of pre-training.

62.11 Rectified linear units improve restricted boltzmann machines [179]

62.11.1 Original Abstract

Restricted Boltzmann machines were developed using binary stochastic hidden units. These can be generalized by replacing each binary unit by an infinite number of copies that all have the same weights but have progressively more negative biases. The learning and inference rules for these "Stepped Sigmoid Units " are unchanged. They can be approximated efficiently by noisy, rectified linear units. Compared with binary units, these units learn features that are better for object recognition on the NORB dataset and face verification on the Labeled Faces in the Wild dataset. Unlike binary units, rectified linear units preserve information about relative intensities as information travels through multiple layers of feature detectors. 1.

62.11.2 Main points

63 2011

63.1 Kernel Adaptive Filtering: A Comprehensive Introduction [158]

63.1.1 Original Abstract

Online learning from a signal processing perspectiveThere is increased interest in kernel learning algorithms in neural networks and a growing need for nonlinear adaptive algorithms in advanced signal processing, communications, and controls. "Kernel Adaptive Filtering" is the first book to present a comprehensive, unifying introduction to online learning algorithms in reproducing kernel Hilbert spaces. Based on research being conducted in the Computational Neuro-Engineering Laboratory at the University of Florida and in the Cognitive Systems Laboratory at McMaster University, Ontario, Canada, this unique resource elevates the adaptive filtering theory to a new level, presenting a new design methodology of nonlinear adaptive filters.Covers the kernel least mean squares algorithm, kernel affine projection algorithms, the kernel recursive least squares algorithm, the theory of Gaussian process regression, and the extended kernel recursive least squares algorithmPresents a powerful model-selection method called maximum marginal likelihoodAddresses the principal bottleneck of kernel adaptive filters–their growing structureFeatures twelve computer-oriented experiments to reinforce the concepts, with MATLAB codes downloadable from the authors' Web siteConcludes each chapter with a summary of the state of the art and potential future directions for original research"Kernel Adaptive Filtering" is ideal for engineers, computer scientists, and graduate students interested in nonlinear adaptive systems for online applications (applications where the data stream arrives one sample at a time and incremental optimal solutions are desirable). It is also a useful guide for those who look for nonlinear adaptive filtering methodologies to solve practical problems.

63.2 Structured learning and prediction in computer vision [191]

63.2.1 Original Abstract

Powerful statistical models that can be learned efficiently from large amounts of data are currently revolutionizing computer vision. These models possess a rich internal structure reflecting task-specific relations and constraints. This monograph introduces the reader to the most popular classes of structured models in computer vision. Our focus is discrete undirected graphical models which we cover in detail together with a description of algorithms for both probabilistic inference and maximum a posteriori inference. We discuss separately recently successful techniques for prediction in general structured models. In the second part of this monograph we describe methods for parameter learning where we distinguish the classic maximum likelihood based methods from the more recent prediction-based parameter learning methods. We highlight developments to enhance current models and discuss kernelized models and latent variable models. To make the monograph more practical and to provide links to further study we provide examples of successful application of many methods in the computer vision literature.

63.3 Action recognition by dense trajectories [257]

63.3.1 Original Abstract

Feature trajectories have shown to be efficient for representing videos. Typically, they are extracted using the KLT tracker or matching SIFT descriptors between frames. However, the quality as well as quantity of these trajectories is often not sufficient. Inspired by the recent success of dense sampling in image classification, we propose an approach to describe videos by dense trajectories. We sample dense points from each frame and track them based on displacement information from a dense optical flow field. Given a state-of-the-art optical flow algorithm, our trajectories are robust to fast irregular motions as well as shot boundaries. Additionally, dense trajectories cover the motion information in videos well. We, also, investigate how to design descriptors to encode the trajectory information. We introduce a novel descriptor based on motion boundary histograms, which is robust to camera motion. This descriptor consistently outperforms other state-of-the-art descriptors, in particular in uncontrolled realistic videos. We evaluate our video description

in the context of action classification with a bag-of-features approach. Experimental results show a significant improvement over the state of the art on four datasets of varying difficulty, i.e. KTH, YouTube, Hollywood2 and UCF sports.

63.4 Face Recognition in Unconstrained Videos with Matched Background Similarity [268]

63.4.1 Original Abstract

Recognizing faces in unconstrained videos is a task of mounting importance. While obviously related to face recognition in still images, it has its own unique characteristics and algorithmic requirements. Over the years several methods have been suggested for this problem, and a few benchmark data sets have been assembled to facilitate its study. However, there is a sizable gap between the actual application needs and the current state of the art. In this paper we make the following contributions. (a) We present a comprehensive database of labeled videos of faces in challenging, uncontrolled conditions (i.e., 'in the wild'), the 'YouTube Faces' database, along with benchmark, pair-matching tests1. (b) We employ our benchmark to survey and compare the performance of a large variety of existing video face recognition techniques. Finally, (c) we describe a novel set-to-set similarity measure, the Matched Background Similarity (MBGS). This similarity is shown to considerably improve performance on the benchmark tests.

63.5 Are sparse representations really relevant for image classification? [204]

63.5.1 Original Abstract

Recent years have seen an increasing interest in sparse representations for image classification and object recognition, probably motivated by evidence from the analysis of the primate visual cortex. It is still unclear, however, whether or not sparsity helps classification. In this paper we evaluate its impact on the recognition rate using a shallow modular architecture, adopting both standard filter banks and filter banks learned in an unsupervised way. In our experiments on the CIFAR-10 and on the Caltech-101 datasets, enforcing sparsity constraints actually does not improve recognition performance. This

has an important practical impact in image descriptor design, as enforcing these constraints can have a heavy computational cost.

63.5.2 Main points

<m:note/>

63.6 Adaptive deconvolutional networks for mid and high level feature learning [271]

63.6.1 Original Abstract

We present a hierarchical model that learns image decompositions via alternating layers of convolutional sparse coding and max pooling. When trained on natural images, the layers of our model capture image information in a variety of forms: low-level edges, mid-level edge junctions, high-level object parts and complete objects. To build our model we rely on a novel inference scheme that ensures each layer reconstructs the input, rather than just the output of the layer directly beneath, as is common with existing hierarchical approaches. This makes it possible to learn multiple layers of representation and we show models with 4 layers, trained on images from the Caltech-101 and 256 datasets. When combined with a standard classifier, features extracted from these models outperform SIFT, as well as representations from other feature learning methods

63.7 Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis [139]

63.7.1 Original Abstract

Previous work on action recognition has focused on adapting hand-designed local features, such as SIFT or HOG, from static images to the video domain. In this paper, we propose using unsupervised feature learning as a way to learn features directly from video data. More specifically, we present an extension of the Independent Subspace Analysis algorithm to learn invariant spatio-temporal features from unlabeled video data. We discovered that, despite its simplicity, this method performs surprisingly well when combined with deep

learning techniques such as stacking and convolution to learn hierarchical representations. By replacing hand-designed features with our learned features, we achieve classification results superior to all previous published results on the Hollywood2, UCF, KTH and YouTube action recognition datasets. On the challenging Hollywood2 and YouTube action datasets we obtain 53.3

63.7.2 Main points

<m:note/>

63.8 Audio-based music classification with a pretrained convolutional network [55]

63.8.1 Original Abstract

Recently the 'Million Song Dataset', containing audio features and metadata for one million songs, was made available. In this paper, we build a convolutional network that is then trained to perform artist recognition, genre recognition and key detection. The network is tailored to summarize the audio features over musically significant timescales. It is infeasible to train the network on all available data in a supervised fashion, so we use unsupervised pretraining to be able to harness the entire dataset: we train a convolutional deep belief network on all data, and then use the learnt parameters to initialize a convolutional multilayer perceptron with the same architecture. The MLP is then trained on a labeled subset of the data for each task. We also train the same MLP with randomly initialized weights. We find that our convolutional approach improves accuracy for the genre recognition and artist recognition tasks. Unsupervised pretraining improves convergence speed in all cases. For artist recognition it improves accuracy as well.

63.9 Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis [140]

63.9.1 Original Abstract

Previous work on action recognition has focused on adapting hand-designed local features, such as SIFT or HOG, from static images to the video domain.

In this paper, we propose using unsupervised feature learning as a way to learn features directly from video data. More specifically, we present an extension of the Independent Subspace Analysis algorithm to learn invariant spatio-temporal features from unlabeled video data. We discovered that, despite its simplicity, this method performs surprisingly well when combined with deep learning techniques such as stacking and convolution to learn hierarchical representations. By replacing hand-designed features with our learned features, we achieve classification results superior to all previous published results on the Hollywood2, UCF, KTH and YouTube action recognition datasets. On the challenging Hollywood2 and YouTube action datasets we obtain 53.3

63.10 Stacked convolutional auto-encoders for hierarchical feature extraction [164]

63.10.1 Original Abstract

We present a novel convolutional auto-encoder (CAE) for unsupervised feature learning. A stack of CAEs forms a convolutional neural network (CNN). Each CAE is trained using conventional on-line gradient descent without additional regularization terms. A max-pooling layer is essential to learn biologically plausible features consistent with those found by previous approaches. Initializing a CNN with filters of a trained CAE stack yields superior performance on a digit (MNIST) and an object recognition (CIFAR10) benchmark.

63.11 Building high-level features using large scale unsupervised learning [142]

63.11.1 Original Abstract

We consider the problem of building high-level, class-specific feature detectors from only unlabeled data. For example, is it possible to learn a face detector using only unlabeled images? To answer this, we train a deep sparse autoencoder on a large dataset of images (the model has 1 billion connections, the dataset has 10 million 200×200 pixel images downloaded from the Internet). We train this network using model parallelism and asynchronous SGD on a cluster with 1,000 machines (16,000 cores) for three days. Contrary to what appears to be a widely-held intuition, our experimental results reveal that it is possible to train a face detector without having to label images as containing

a face or not. Control experiments show that this feature detector is robust not only to translation but also to scaling and out-of-plane rotation. We also find that the same network is sensitive to other high-level concepts such as cat faces and human bodies. Starting from these learned features, we trained our network to recognize 22,000 object categories from ImageNet and achieve a leap of 70

63.11.2 Main points

63.12 Generating text with recurrent neural networks [236]

63.12.1 Original Abstract

None

64 2012

64.1 Alan Turing's Electronic Brain: The Struggle to Build the ACE, the World's Fastest Computer [44]

64.1.1 Original Abstract

The mathematical genius Alan Turing, now well known for his crucial wartime role in breaking the ENIGMA code, was the first to conceive of the fundamental principle of the modern computer-the idea of controlling a computing machine's operations by means of a program of coded instructions, stored in the machine's 'memory'. In 1945 Turing drew up his revolutionary design for an electronic computing machine-his Automatic Computing Engine ('ACE'). A pilot model of the ACE ran its first program in 1950 and the production version, the 'DEUCE', went on to become a cornerstone of the fledgling British computer industry. The first 'personal' computer was based on Turing's ACE. Alan Turing's Automatic Computing Engine describes Turing's struggle to build the modern computer. The first detailed history of Turing's contributions to computer science, this text is essential reading for anyone interested in the history of the computer and the history of mathematics. It contains first hand accounts by Turing and by the pioneers of computing who worked with him. As well as relating the story of the invention of

the computer, the book clearly describes the hardware and software of the ACE-including the very first computer programs. The book is intended to be accessible to everyone with an interest in computing, and contains numerous diagrams and illustrations as well as original photographs. The book contains chapters describing Turing's path-breaking research in the fields of Artificial Intelligence (AI) and Artificial Life (A-Life). The book has an extensive system of hyperlinks to The Turing Archive for the History of Computing, an on-line library of digital facsimiles of typewritten documents by Turing and the other scientists who pioneered the electronic computer.

64.1.2 Main points

64.2 Connectionism [78]

64.2.1 Original Abstract

Connectionism is a movement in cognitive science which hopes toexplain human intellectual abilities using artificial neural networks(also known as 'neural networks' or 'neuralnets'). Neural networks are simplified models of the braincomposed of large numbers of units (the analogs of neurons) togetherwith weights that measure the strength of connections between theunits. These weights model the effects of the synapses that link oneneuron to another. Experiments on models of this kind have demonstratedan ability to learn such skills as face recognition, reading, and thedetection of simple grammatical structure., Philosophers have become interested in connectionism because itpromises to provide an alternative to the classical theory of the mind:the widely held view that the mind is something akin to a digitalcomputer processing a symbolic language. Exactly how and to what extentthe connectionist paradigm constitutes a challenge to classicism hasbeen a matter of hot debate in recent years.

64.2.2 Main points

64.3 Computer Vision - ECCV 2012 [66]

64.3.1 Original Abstract

None

64.4 Combining gradient histograms using orientation tensors for human action recognition [197]

64.4.1 Original Abstract

We present a method for human action recognition based on the combination of Histograms of Gradients into orientation tensors. It uses only information from HOG3D: no features or points of interest are extracted. The resulting raw histograms obtained per frame are combined into an orientation tensor, making it a simple, fast to compute and effective global descriptor. The addition of new videos and/or new action cathegories does not require any recomputation or changes to the previously computed descriptors. Our method reaches 92.01

64.5 Unsupervised and Transfer Learning Challenge: a Deep Learning Approach. [168]

64.5.1 Original Abstract

Learning good representations from a large set of unlabeled data is a particularlychallenging task. Recent work (see Bengio (2009) for a review) shows that training deep architectures is a good way to extract such representations, by extractingand disentangling gradually higher-level factors of variation characterizing the inputdistribution. In this paper, we describe different kinds of layers we trained for learning representations in the setting of the Unsupervised and Transfer Learning Challenge. The strategy of our team won the final phase of the challenge. It combined andstacked different one-layer unsupervised learning algorithms, adapted to each of thefive datasets of the competition. This paper describes that strategy and the particularone-layer learning algorithms feeding a simple linear classifier with a tiny number oflabeled training samples (1 to 64 per class).

64.6 Attribute learning for understanding unstructured social activity [72]

64.6.1 Original Abstract

The rapid development of social video sharing platforms has created a huge demand for automatic video classification and annotation techniques, in par-

ticular for videos containing social activities of a group of people (e.g. YouTube video of a wedding reception). Recently, attribute learning has emerged as a promising paradigm for transferring learning to sparsely labelled classes in object or single-object short action classification. In contrast to existing work, this paper for the first time, tackles the problem of attribute learning for understanding group social activities with sparse labels. This problem is more challenging because of the complex multi-object nature of social activities, and the unstructured nature of the activity context. To solve this problem, we (1) contribute an unstructured social activity attribute (USAA) dataset with both visual and audio attributes, (2) introduce the concept of semi-latent attribute space and (3) propose a novel model for learning the latent attributes which alleviate the dependence of existing models on exact and exhaustive manual specification of the attribute-space. We show that our framework is able to exploit latent attributes to outperform contemporary approaches for addressing a variety of realistic multi-media sparse data learning tasks including: multi-task learning, N-shot transfer learning, learning with label noise and importantly zero-shot learning.

64.7 TRECVid 2012 Semantic Video Concept Detection by NTT-MD-DUT [235]

64.7.1 Original Abstract

In this paper, we describe the TRECVid 2012 videoconcept detection system first developed at the NTTMedia Intelligence Laboratories in collaborationwith Dalian University of Technology. For thisyear's task, we adopted a subspace partition basedscheme for classifier learning, with emphasis on thereduction of classifier complexity, aiming atimproving the training efficiency and boosting theclassifier performance. As the video corpus used forTRECVid evaluation is ever increasing, two practicalissues are becoming more and more challenging forbuilding concept detection systems. The first one isthe time-consuming training and testing procedures,which have taken up most of the evaluation activities,preventing the design and testing of novel algorithms.The second and the more important issue is thatwhen using whole data for classifier training, thederived separating hyperplanes would be rathercomplex and thus degrade the classificationperformance. To address these issues, we propose toadopt the "divide-and-conquer" strategy for conceptdetector construction as follows. We first partitionthe whole training feature

space into multiplesub-space with a scalable clustering method, andthen build sub-classifiers on these sub-spacesseparately for each concept. The decision of a testingsample is the fusion of the results a few firedsub-classifiers. Experimental results demonstrate theefficiency and effectiveness of our proposedapproach.

64.8 AXES at TRECVid 2012: KIS, INS, and MED [11]

64.8.1 Original Abstract

The AXES project participated in the interactive instance search task (INS), the known-item search task (KIS), and the multimedia event detection task (MED) for TRECVid 2012. As in our TRECVid 2011 system, we used nearly identical search systems and user interfaces for both INS and KIS. Our interactive INS and KIS systems focused this year on using classifiers trained at query time with positive examples collected from external search engines. Participants in our KIS experiments were media professionals from the BBC; our INS experiments were carried out by students and researchers at Dublin City University. We performed comparatively well in both experiments. Our best KIS run found 13 of the 25 topics, and our best INS runs outperformed all other submitted runs in terms of P@100. For MED, the system presented was based on a minimal number of low-level descriptors, which we chose to be as large as computationally feasible. These descriptors are aggregated to produce high-dimensional video-level signatures, which are used to train a set of linear classifiers. Our MED system achieved the second-best score of all submitted runs in the main track, and best score in the ad-hoc track, suggesting that a simple system based on state-of-the-art low-level descriptors can give relatively high performance. This paper describes in detail our KIS, INS, and MED systems and the results and findings of our experiments.

64.9 Deep Neural Networks for Acoustic Modeling in Speech Recognition [106]

64.9.1 Original Abstract

Most current speech recognition systems use hidden Markov models (HMMs) to deal with the temporal variability of speech and Gaussian mixture models (GMMs) to determine how well each state of each HMM fits a frame or a

short window of frames of coefficients that represents the acoustic input. An alternative way to evaluate the fit is to use a feed-forward neural network that takes several frames of coefficients as input and produces posterior probabilities over HMM states as output. Deep neural networks (DNNs) that have many hidden layers and are trained using new methods have been shown to outperform GMMs on a variety of speech recognition benchmarks, sometimes by a large margin. This article provides an overview of this progress and represents the shared views of four research groups that have had recent successes in using DNNs for acoustic modeling in speech recognition.

64.9.2 Main points

<m:note/>

64.10 Local-feature-map Integration Using Convolutional Neural Networks for Music Genre Classification [180]

64.10.1 Original Abstract

None

64.11 Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition [2]

64.11.1 Original Abstract

Convolutional Neural Networks (CNN) have showed success in achieving translation invariance for many image processing tasks. The success is largely attributed to the use of local filtering and max-pooling in the CNN architecture. In this paper, we propose to apply CNN to speech recognition within the framework of hybrid NN-HMM model. We propose to use local filtering and max-pooling in frequency domain to normalize speaker variance to achieve higher multi-speaker speech recognition performance. In our method, a pair of local filtering layer and max-pooling layer is added at the lowest end of neural network (NN) to normalize spectral variations of speech signals. In our experiments, the proposed CNN architecture is evaluated in a

speaker independent speech recognition task using the standard TIMIT data sets. Experimental results show that the proposed CNN method can achieve over 10

64.11.2 Main points

<m:note/>

64.12 Recognizing 50 human action categories of web videos [202]

64.12.1 Original Abstract

Action recognition on large categories of unconstrained videos taken from the web is a very challenging problem compared to datasets like KTH (6 actions), IXMAS (13 actions), and Weizmann (10 actions). Challenges like camera motion, different viewpoints, large interclass variations, cluttered background, occlusions, bad illumination conditions, and poor quality of web videos cause the majority of the state-of-the-art action recognition approaches to fail. Also, an increased number of categories and the inclusion of actions with high confusion add to the challenges. In this paper, we propose using the scene context information obtained from moving and stationary pixels in the key frames, in conjunction with motion features, to solve the action recognition problem on a large (50 actions) dataset with videos from the web. We perform a combination of early and late fusion on multiple features to handle the very large number of categories. We demonstrate that scene context is a very important feature to perform action recognition on very large datasets. The proposed method does not require any kind of video stabilization, person detection, or tracking and pruning of features. Our approach gives good performance on a large number of action categories; it has been tested on the UCF50 dataset with 50 action categories, which is an extension of the UCF YouTube Action (UCF11) dataset containing 11 action categories. We also tested our approach on the KTH and HMDB51 datasets for comparison.

64.12.2 Main points

Test on UCF50

64.13 Improving neural networks by preventing co-adaptation of feature detectors [104]

64.13.1 Original Abstract

When a large feedforward neural network is trained on a small training set, it typically performs poorly on held-out test data. This "overfitting" is greatly reduced by randomly omitting half of the feature detectors on each training case. This prevents complex co-adaptations in which a feature detector is only helpful in the context of several other specific feature detectors. Instead, each neuron learns to detect a feature that is generally helpful for producing the correct answer given the combinatorially large variety of internal contexts in which it must operate. Random "dropout" gives big improvements on many benchmark tasks and sets new records for speech and object recognition.

64.13.2 Main points

- Paper about Dropout
- Standard way to reduce test error
  - averaging different models
  - Computationally expensive in training and test
- Dropout
  - Small training set
  - Prevents “overfitting”
  - They use 50%
  - Instead of L2 norm, they set an upper bound for each individual neuron.
  - Mean network : At test time divide all the outgoing weights by 2 to compensate dropout
  - Specific case
    - Single hidden layer network
    - N hidden units
    - “Softmax” output

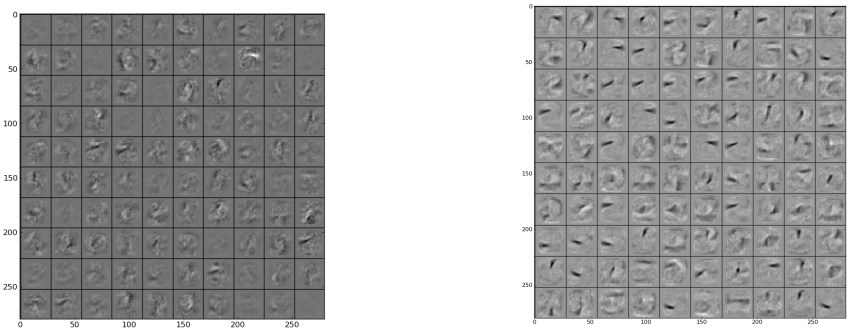

Figure 1: **Visualization of features learned by first layer hidden units** left without dropout and right using dropout

        * 50% dropout
        * during test using mean network
        * Exactly equivalent to taking the geometric mean of the probability distributions over labels predicted by all 2^N possible networks

- Results
  - MNIST
    * No dropout : 160 errors
    * Dropbout : 130 errors
    * Dropout + rm random 20% pixels : 110 errors
    * Deep Boltzmann machine : 88 errors
    * + Dropout : 79 errors
  - TIMIT
    * 4 Fully-connected hidden layers 4000 units per layer
    * + 185 "softmax" output units
    * Without dropout : 22.7%
    * Dropout on hidden units : 19.7%
  - CIFAR-10
    * Best published : 18.5%

    - 3 Conv+Max-pool 1 Fully : 16.6%
    - + Dropout in last hidden layer : 15.6%
  - ImageNet
    - Average of 6 separate models : 47.2%
    - state-of-the-art : 45.7%
    - 5 Conv+Max-pool
    - + 2 Fully
    - + 1000 "softmax"
    - Without dropout : 48.6%
    - Dropout in the 6th : 42.4%
  - Reuters
    - 2 fully of 2000 hidden units
    - Without dropout : 31.05%
    - Dropout : 29.62%

64.14 Gated boltzmann machine in texture modeling [92]

64.14.1 Original Abstract

In this paper, we consider the problem of modeling complex texture information using undirected probabilistic graphical models. Texture is a special type of data that one can better understand by considering its local structure. For that purpose, we propose a convolutional variant of the Gaussian gated Boltzmann machine (GGBM) [12], inspired by the co-occurrence matrix in traditional texture analysis. We also link the proposed model to a much simpler Gaussian restricted Boltzmann machine where convolutional features are computed as a preprocessing step. The usefulness of the model is illustrated in texture classification and reconstruction experiments.

64.15 ImageNet Classification with Deep Convolutional Neural Networks [134]

64.15.1 Original Abstract

We trained a large, deep convolutional neural network to classify the 1.2 millionhigh-resolution images in the ImageNet LSVRC-2010 contest into the

1000 dif-ferent classes. On the test data, we achieved top-1 and top-5 error rates of 37.5

64.15.2 Main points

- CNN architecture:
  - 650.000 neurons (60 million parameters)
  - 5 convolutional layers
  - Some of them followed by a max-pooling layer
  - 3 fully-connected layers
  - 1 1000-way softmax
- Dropout regularization method to reduce overfitting in 3 fully-connected layers
- Training time: 5-6 days on two GTX 580 3GB GPUs
- Dataset:
  - ILSVRC-2010
  - Down-sampled images to a fixed resolution of 256x256
  - Substract the mean activity ofver training set from each pixel
- ReLU:
  - $f(x) = \max(0, x)$
  - Faster than tanh
  - ReLU: 6 epochs
  - tanh: 36 more epochs to achieve same performance
- Local Response Normalization
  - 1.2 and 1.4% error reduction
  - Helps generalization
  - $b^i_{x,y} = a^i_{x,y} / \left(k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a^j_{x,y})^2 \right)^\beta$

  - $k = 2, n = 5, \alpha = 10^-4$, and $\beta = 0.75$

- Overlapping Pooling
  - 0.3 and 0.4% error reduction
  - grid $3 \prod 3$
  - stride = 2
  - Overlap each pooling one column pixel

- Overall Architecture
  - 224x224x3 (RGB image)
  - Conv 96 kernels of size 11x11x3 with stride of 4 pixels
  - Response-Normalized and max-pooling
  - Conv 256 kernels of size 5x5x48 with stride of ? pixels
  - Response-Normalized and max-pooling
  - Conv 384 kernels of size 3x3x256
  - Conv 384 kernels of size 3x3x192
  - Conv 256 kernels of size 3x3x192
  - ¿Response-Normalized? and Max-pooling
  - Fully connected 4096
  - Fully connected 4096
  - Fully connected 1000
  - Softmax

- Data augmentation
  - 0.1 error reduction
  - Original images escaled scaled and croped to 256x256
  - Extract 5 images of 224x224 from corners plus center
  - Mirror horizontally and get 5 more images
  - Augment data altering RGB channels:
    * Perform PCA on RGB throughout the training set

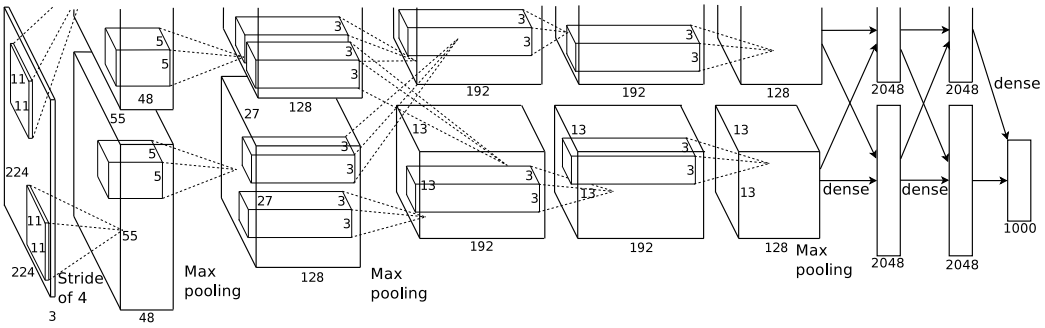

Figure 2: **Architecture of the CNN**

∗ Each training image add multiples of PCs with gaussian noise

- Dropout
  - Put to zero the output of neurons with probability 0.5
  - At test time multiply the outputs by 0.5
  - Two first fully-connected layers
  - Solves overfitting
  - Dobules the number of iterations required to ocnverge
- Details of learning
  - batch size = 128
  - momentum 0.9
  - weight decay 0.0005
  - Initial weights from zero-mean Gaussian std=0.01
  - biases = 1 on second, fourth, fifth Conv and fully-connected
  - biases = 0 on the rest
- Evaluation
  - Consider the feature activations induced by an image at the last, 4096-dimensional hidden layer

64.16 The Stanford / Technicolor / Fraunhofer HHI Video [12]

64.16.1 Original Abstract

Video search has become a very important tool, with the ever-growing size of multimedia collections. This work introduces our Video Semantic Indexing system. Our experiments show that Residual Vectors provide an efficient way of aggregat- ing local descriptors, with complementary gain with respect to BoVW. Also, we show that systems using a limited number of descriptors and machine learning techniques can still be quite effective. Our first participation at the TRECVID evaluation has been very fruitful: our team was ranked 6th in the light version of the Semantic Indexing task.

64.17 Learning hierarchical features for scene labeling [61]

64.17.1 Original Abstract

Scene labeling consists of labeling each pixel in an image with the category of the object it belongs to. We propose a method that uses a multiscale convolutional network trained from raw pixels to extract dense feature vectors that encode regions of multiple sizes centered on each pixel. The method alleviates the need for engineered features, and produces a powerful representation that captures texture, shape, and contextual information. We report results using multiple postprocessing methods to produce the final labeling. Among those, we propose a technique to automatically retrieve, from a pool of segmentation components, an optimal set of components that best explain the scene; these components are arbitrary, for example, they can be taken from a segmentation tree or from any family of oversegmentations. The system yields record accuracies on the SIFT Flow dataset (33 classes) and the Barcelona dataset (170 classes) and near-record accuracy on Stanford background dataset (eight classes), while being an order of magnitude faster than competing approaches, producing a 320×240 image labeling in less than a second, including feature extraction.

64.18 Machine learning: a probabilistic perspective [178]

64.18.1 Original Abstract

None

64.19 Differential feedback modulation of center and surround mechanisms in parvocellular cells in the visual thalamus [122]

64.19.1 Original Abstract

Many cells in both the central visual system and other sensory systems exhibit a center surround organization in their receptive field, where the response to a centrally placed stimulus is modified when a surrounding area is also stimulated. This can follow from laterally directed connections in the local circuit at the level of the cell in question but could also involve more complex interactions. In the lateral geniculate nucleus (LGN), the cells relaying the retinal input display a concentric, center surround organization that in part follows from the similar organization characterizing the retinal cells providing their input. However, local thalamic inhibitory interneurons also play a role, and as we examine here, feedback from the visual cortex too. Here, we show in the primate (macaque) that spatially organized cortical feedback provides a clear and differential influence serving to enhance both responses to stimulation within the center of the receptive field and the ability of the nonclassical surround mechanism to attenuate this. In short, both center and surround mechanisms are influenced by the feedback. This dynamically sharpens the spatial focus of the receptive field and introduces nonlinearities from the cortical mechanism into the LGN.

64.19.2 Main points

65 2013

65.1 Discrete geometry and optimization [25]

65.1.1 Original Abstract

?Optimization has long been a source of both inspiration and applications for geometers, and conversely, discrete and convex geometry have provided the foundations for many optimization techniques, leading to a rich interplay between these subjects. The purpose of the Workshop on Discrete Geometry, the Conference on Discrete Geometry and Optimization, and the Workshop on Optimization, held in September 2011 at the Fields Institute, Toronto, was to further stimulate the interaction between geometers and optimizers. This volume reflects the interplay between these areas. The inspiring Fejes Tóth Lecture Series, delivered by Thomas Hales of the University of Pittsburgh, exemplified this approach. While these fields have recently witnessed a lot of activity and successes, many questions remain open. For example, Fields medalist Stephen Smale stated that the question of the existence of a strongly polynomial time algorithm for linear optimization is one of the most important unsolved problems at the beginning of the 21st century. The broad range of topics covered in this volume demonstrates the many recent and fruitful connections between different approaches, and features novel results and state-of-the-art surveys as well as open problems.

65.1.2 Main points

65.2 Maxout Networks [83]

65.2.1 Original Abstract

We consider the problem of designing models to leverage a recently introduced approximate model averaging technique called dropout. We define a simple new model called maxout (so named because its output is the max of a set of inputs, and because it is a natural companion to dropout) designed to both facilitate optimization by dropout and improve the accuracy of dropout's fast approximate model averaging technique. We empirically verify that the model successfully accomplishes both of these tasks. We use maxout and dropout to

demonstrate state of the art classification performance on four benchmark datasets: MNIST, CIFAR-10, CIFAR-100, and SVHN.

65.2.2 Main points

65.3 Deep Generative Stochastic Networks Trainable by Backprop [23]

65.3.1 Original Abstract

We introduce a novel training principle for probabilistic models that is an alternative to maximum likelihood. The proposed Generative Stochastic Networks (GSN) framework is based on learning the transition operator of a Markov chain whose stationary distribution estimates the data distribution. The transition distribution of the Markov chain is conditional on the previous state, generally involving a small move, so this conditional distribution has fewer dominant modes, being unimodal in the limit of small moves. Thus, it is easier to learn because it is easier to approximate its partition function, more like learning to perform supervised function approximation, with gradients that can be obtained by backprop. We provide theorems that generalize recent work on the probabilistic interpretation of denoising autoencoders and obtain along the way an interesting justification for dependency networks and generalized pseudolikelihood, along with a definition of an appropriate joint distribution and sampling mechanism even when the conditionals are not consistent. GSNs can be used with missing inputs and can be used to sample subsets of variables given the rest. We validate these theoretical results with experiments on two image datasets using an architecture that mimics the Deep Boltzmann Machine Gibbs sampler but allows training to proceed with simple backprop, without the need for layerwise pretraining.

65.4 Improving Deep Neural Networks with Probabilistic Maxout Units [231]

65.4.1 Original Abstract

We present a probabilistic variant of the recently introduced maxout unit. The success of deep neural networks utilizing maxout can partly be attributed to favorable performance under dropout, when compared to rectified linear units. It however also depends on the fact that each maxout unit performs

a pooling operation over a group of linear transformations and is thus partially invariant to changes in its input. Starting from this observation we ask the question: Can the desirable properties of maxout units be preserved while improving their invariance properties ? We argue that our probabilistic maxout (probout) units successfully achieve this balance. We quantitatively verify this claim and report classification performance matching or exceeding the current state of the art on three challenging image classification benchmarks (CIFAR-10, CIFAR-100 and SVHN).

65.4.2 Main points

65.5 Network In Network [154]

65.5.1 Original Abstract

We propose a novel deep network structure called "Network In Network" (NIN) to enhance model discriminability for local patches within the receptive field. The conventional convolutional layer uses linear filters followed by a nonlinear activation function to scan the input. Instead, we build micro neural networks with more complex structures to abstract the data within the receptive field. We instantiate the micro neural network with a multilayer perceptron, which is a potent function approximator. The feature maps are obtained by sliding the micro networks over the input in a similar manner as CNN; they are then fed into the next layer. Deep NIN can be implemented by stacking mutiple of the above described structure. With enhanced local modeling via the micro network, we are able to utilize global average pooling over feature maps in the classification layer, which is easier to interpret and less prone to overfitting than traditional fully connected layers. We demonstrated the state-of-the-art classification performances with NIN on CIFAR-10 and CIFAR-100, and reasonable performances on SVHN and MNIST datasets.

65.6 An Empirical Investigation of Catastrophic Forgeting in Gradient-Based Neural Networks [86]

65.6.1 Original Abstract

Catastrophic forgetting is a problem faced by many machine learning models and algorithms. When trained on one task, then trained on a second task, many machine learning models "forget" how to perform the first task. This is

widely believed to be a serious problem for neural networks. Here, we investigate the extent to which the catastrophic forgetting problem occurs for modern neural networks, comparing both established and recent gradient-based training algorithms and activation functions. We also examine the effect of the relationship between the first task and the second task on catastrophic forgetting. We find that it is always best to train using the dropout algorithm–the dropout algorithm is consistently best at adapting to the new task, remembering the old task, and has the best tradeoff curve between these two extremes. We find that different tasks and relationships between tasks result in very different rankings of activation function performance. This suggests the choice of activation function should always be cross-validated.

65.7 Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks [84]

65.7.1 Original Abstract

Recognizing arbitrary multi-character text in unconstrained natural photographs is a hard problem. In this paper, we address an equally hard subproblem in this domain viz. recognizing arbitrary multi-digit numbers from Street View imagery. Traditional approaches to solve this problem typically separate out the localization, segmentation, and recognition steps. In this paper we propose a unified approach that integrates these three steps via the use of a deep convolutional neural network that operates directly on the image pixels. We employ the DistBelief implementation of deep neural networks in order to train large, distributed neural networks on high quality images. We find that the performance of this approach increases with the depth of the convolutional network, with the best performance occurring in the deepest architecture we trained, with eleven hidden layers. We evaluate this approach on the publicly available SVHN dataset and achieve over 96

65.8 Coloring Action Recognition in Still Images [127]

65.8.1 Original Abstract

In this article we investigate the problem of human action recognition in static images. By action recognition we intend a class of problems which in-

cludes both action classification and action detection (i.e. simultaneous localization and classification). Bag-of-words image representations yield promising results for action classification, and deformable part models perform very well object detection. The representations for action recognition typically use only shape cues and ignore color information. Inspired by the recent success of color in image classification and object detection, we investigate the potential of color for action classification and detection in static images. We perform a comprehensive evaluation of color descriptors and fusion approaches for action recognition. Experiments were conducted on the three datasets most used for benchmarking action recognition in still images: Willow, PASCAL VOC 2010 and Stanford-40. Our experiments demonstrate that incorporating color information considerably improves recognition performance, and that a descriptor based on color names outperforms pure color descriptors. Our experiments demonstrate that late fusion of color and shape information outperforms other approaches on action recognition. Finally, we show that the different color–shape fusion approaches result in complementary information and combining them yields state-of-the-art performance for action classification.

65.8.2 Main points

<m:note/>

65.9 Do Deep Nets Really Need to be Deep? [15]

65.9.1 Original Abstract

Currently, deep neural networks are the state of the art on problems such as speech recognition and computer vision. In this extended abstract, we show that shallow feed-forward networks can learn the complex functions previously learned by deep nets and achieve accuracies previously only achievable with deep models. Moreover, in some cases the shallow neural nets can learn these deep functions using a total number of parameters similar to the original deep model. We evaluate our method on the TIMIT phoneme recognition task and are able to train shallow fully-connected nets that perform similarly to complex, well-engineered, deep convolutional architectures. Our success in training shallow neural nets to mimic deeper models suggests that there probably exist better algorithms for training shallow feed-forward nets than

those currently available.

65.10 Intriguing properties of neural networks [240]

65.10.1 Original Abstract

Deep neural networks are highly expressive models that have recently achieved state of the art performance on speech and visual recognition tasks. While their expressiveness is the reason they succeed, it also causes them to learn uninterpretable solutions that could have counter-intuitive properties. In this paper we report two such properties. First, we find that there is no distinction between individual high level units and random linear combinations of high level units, according to various methods of unit analysis. It suggests that it is the space, rather than the individual units, that contains of the semantic information in the high layers of neural networks. Second, we find that deep neural networks learn input-output mappings that are fairly discontinuous to a significant extend. We can cause the network to misclassify an image by applying a certain imperceptible perturbation, which is found by maximizing the network's prediction error. In addition, the specific nature of these perturbations is not a random artifact of learning: the same perturbation can cause a different network, that was trained on a different subset of the dataset, to misclassify the same input.

65.11 3D convolutional neural networks for human action recognition. [119]

65.11.1 Original Abstract

We consider the automated recognition of human actions in surveillance videos. Most current methods build classifiers based on complex handcrafted features computed from the raw inputs. Convolutional neural networks (CNNs) are a type of deep model that can act directly on the raw inputs. However, such models are currently limited to handling 2D inputs. In this paper, we develop a novel 3D CNN model for action recognition. This model extracts features from both the spatial and the temporal dimensions by performing 3D convolutions, thereby capturing the motion information encoded in multiple adjacent frames. The developed model generates multiple channels of information from the input frames, and the final feature representation combines

information from all channels. To further boost the performance, we propose regularizing the outputs with high-level features and combining the predictions of a variety of different models. We apply the developed models to recognize human actions in the real-world environment of airport surveillance videos, and they achieve superior performance in comparison to baseline methods.

65.11.2 Main points

- Participated on TRECVID 2009
- Videos with static camera
- 3D - CNN
  - input 7@60x40
  - Hardwired 33@60x40
    - Gray
    - gradient-x
    - gradient-y
    - opticalflow-x
    - opticalflow-y
  - Convolution 7x7x3
  - C2 = 23*2@54x34
  - Subsampling 2x2
  - S3 = 23*2@27x17
  - Convolution 7x6x3
  - C4 = 13*6@21x12
  - Subsampling 3x3
  - S5 = 13*6@7x4
  - Convolution 7x4
  - C6 = 128@7x4
  - Fully connected layer
- Datasets

- Surveillance Event Detection
- Action classes
  - CellToEar
  - ObjectPut
  - Pointing
  - method
    - Humman detector to locate human head
    - Create a bounding box with 7 time frames and 60x40 spatial pixels
    -
  - Best results on three tasks
- KTH
  - Comptetitive performance

65.12 Learned versus Hand-Designed Feature Representations for 3d Agglomeration [28]

65.12.1 Original Abstract

For image recognition and labeling tasks, recent results suggest that machine learning methods that rely on manually specified feature representations may be outperformed by methods that automatically derive feature representations based on the data. Yet for problems that involve analysis of 3d objects, such as mesh segmentation, shape retrieval, or neuron fragment agglomeration, there remains a strong reliance on hand-designed feature descriptors. In this paper, we evaluate a large set of hand-designed 3d feature descriptors alongside features learned from the raw data using both end-to-end and unsupervised learning techniques, in the context of agglomeration of 3d neuron fragments. By combining unsupervised learning techniques with a novel dynamic pooling scheme, we show how pure learning-based methods are for the first time competitive with hand-designed 3d shape descriptors. We investigate data augmentation strategies for dramatically increasing the size of the training set, and show how combining both learned and hand-designed features leads to the highest accuracy.

65.13 Comparison of Artificial Neural Networks ; and training an Extreme Learning Machine [246]

65.13.1 Original Abstract

None

65.14 Mitosis detection in breast cancer histology images with deep neural networks [42]

65.14.1 Original Abstract

We use deep max-pooling convolutional neural networks to detect mitosis in breast histology images. The networks are trained to classify each pixel in the images, using as context a patch centered on the pixel. Simple postprocessing is then applied to the network output. Our approach won the ICPR 2012 mitosis detection competition, outperforming other contestants by a significant margin.

65.14.2 Main points

65.15 Action and event recognition with Fisher vectors on a compact feature set [192]

65.15.1 Original Abstract

Action recognition in uncontrolled video is an important and challenging computer vision problem. Recent progress in this area is due to new local features and models that capture spatio-temporal structure between local features, or human-object interactions. Instead of working towards more complex models, we focus on the low-level features and their encoding. We evaluate the use of Fisher vectors as an alternative to bag-of-word histograms to aggregate a small set of state-of-the-art low-level descriptors, in combination with linear classifiers. We present a large and varied set of evaluations, considering (i) classification of short actions in five datasets, (ii) localization of such actions in feature-length movies, and (iii) large-scale recognition of complex events. We find that for basic action recognition and localization MBH features alone are enough for state-of-the-art performance. For complex events we find that SIFT and MFCC features provide complementary cues. On all three problems

we obtain state-of-the-art results, while using fewer features and less complex models.

65.16 Semi-supervised Learning of Feature Hierarchies for Object Detection in a Video [269]

65.16.1 Original Abstract

We propose a novel approach to boost the performance of generic object detectors on videos by learning video-specific features using a deep neural network. The insight behind our proposed approach is that an object appearing in different frames of a video clip should share similar features, which can be learned to build better detectors. Unlike many supervised detector adaptation or detection-by-tracking methods, our method does not require any extra annotations or utilize temporal correspondence. We start with the high-confidence detections from a generic detector, then iteratively learn new video-specific features and refine the detection scores. In order to learn discriminative and compact features, we propose a new feature learning method using a deep neural network based on auto en-coders. It differs from the existing unsupervised feature learning methods in two ways: first it optimizes both discriminative and generative properties of the features simultaneously, which gives our features better discriminative ability, second, our learned features are more compact, while the unsupervised feature learning methods usually learn a redundant set of over-complete features. Extensive experimental results on person and horse detection show that significant performance improvement can be achieved with our proposed method.

65.17 MediaMill at TRECVID 2013: Searching Concepts, Objects, Instances and Events in Video [229]

65.17.1 Original Abstract

In this paper we summarize our TRECVID 2013 video retrieval experiments. The MediaMill team participated in four tasks: concept detection, object localization, in- stance search, and event recognition. For all tasks the starting point is our top-performing bag-of-words system of TRECVID 2008-2012,

which uses color SIFT descrip- tors, average and difference coded into codebooks with spa- tial pyramids and kernel-based machine learning. New this year are concept detection with deep learning, concept detec- tion without annotations, object localization using selective search, instance search by reranking, and event recognition based on concept vocabularies. Our experiments focus on es- tablishing the video retrieval value of the innovations. The 2013 edition of the TRECVID benchmark has again been a fruitful participation for the MediaMill team, resulting in the best result for concept detection, concept detection with- out annotation, object localization, concept pair detection, and visual event recognition with few examples.

65.18 TRECVID 2013 - An Introduction to the Goals , Tasks , Data , Evaluation Mechanisms , and Metrics [193]

65.18.1 Original Abstract

None

65.19 Quaero at TRECVid 2013 : Semantic Indexing [212]

65.19.1 Original Abstract

The Quaero group is a consortium of French and German organizations working on Multimedia Indexing and Retrieval. LIG, INRIA and KIT participated to the semantic indexing task and LIG participated to the organization of this task. This paper describes these participations. For the semantic indexing task, our approach uses a six-stages processing pipelines for computing scores for the likelihood of a video shot to contain a target concept. These scores are then used for producing a ranked list of images or shots that are the most likely to contain the target concept. The pipeline is composed of the following steps: descriptor extraction, descriptor optimization, classi cation, fusion of descriptor variants, higher-level fusion, and re-ranking. We used a number of di erent descriptors and a hierarchical fusion strategy. We also used conceptual feedback by adding a vector of classi cation score to the pool of descriptors. The best Quaero run has a Mean Inferred Average Precision

of 0.2692, which ranked us 3rd out of 16 participants. We also organized the TRECVid SIN 2012 collaborative annotation.

65.20 Understanding Deep Architectures using a Recursive Convolutional Network [57]

65.20.1 Original Abstract

A key challenge in designing convolutional network models is sizing them appropriately. Many factors are involved in these decisions, including number of layers, feature maps, kernel sizes, etc. Complicating this further is the fact that each of these influence not only the numbers and dimensions of the activation units, but also the total number of parameters. In this paper we focus on assessing the independent contributions of three of these linked variables: The numbers of layers, feature maps, and parameters. To accomplish this, we employ a recursive convolutional network whose weights are tied between layers; this allows us to vary each of the three factors in a controlled setting. We find that while increasing the numbers of layers and parameters each have clear benefit, the number of feature maps (and hence dimensionality of the representation) appears ancillary, and finds most of its benefit through the introduction of more weights. Our results (i) empirically confirm the notion that adding layers alone increases computational power, within the context of convolutional layers, and (ii) suggest that precise sizing of convolutional feature map dimensions is itself of little concern; more attention should be paid to the number of parameters in these layers instead.

65.20.2 Main points

- Deeper models are preferred over shallow ones
- Performance is independent of the number of units, when depth and parameters remains constant
- Recurrent Neural Network:
  - Convolutional architecture
  - all layers same number of feature maps
  - weights are tied across layers

– ReLU in all layers

– Max-pooling with non-overlaping windows

65.21 Visualizing and Understanding Convolutional Networks [272]

65.21.1 Original Abstract

Large Convolutional Network models have recently demonstrated impressive classification performance on the ImageNet benchmark. However there is no clear understanding of why they perform so well, or how they might be improved. In this paper we address both issues. We introduce a novel visualization technique that gives insight into the function of intermediate feature layers and the operation of the classifier. We also perform an ablation study to discover the performance contribution from different model layers. This enables us to find model architectures that outperform Krizhevsky et.al. on the ImageNet classification benchmark. We show our ImageNet model generalizes well to other datasets: when the softmax classifier is retrained, it convincingly beats the current state-of-the-art results on Caltech-101 and Caltech-256 datasets.

65.22 Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps [225]

65.22.1 Original Abstract

This paper addresses the visualisation of image classification models, learnt using deep Convolutional Networks (ConvNets). We consider two visualisation techniques, based on computing the gradient of the class score with respect to the input image. The first one generates an image, which maximises the class score [Erhan et al., 2009], thus visualising the notion of the class, captured by a ConvNet. The second technique computes a class saliency map, specific to a given image and class. We show that such maps can be employed for weakly supervised object segmentation using classification ConvNets. Finally, we establish the connection between the gradient-based ConvNet visualisation methods and deconvolutional networks [Zeiler et al., 2013].

65.23 Challenges in Representation Learning: A report on three machine learning contests [85]

65.23.1 Original Abstract

The ICML 2013 Workshop on Challenges in Representation Learning focused on three challenges: the black box learning challenge, the facial expression recognition challenge, and the multimodal learning challenge. We describe the datasets created for these challenges and summarize the results of the competitions. We provide suggestions for organizers of future challenges and some comments on what kind of knowledge can be gained from machine learning competitions.

65.24 Caffe: An open source convolutional architecture for fast feature embeding. [120]

65.24.1 Original Abstract

None

65.25 Deep Fisher networks for large-scale image classification [226]

65.25.1 Original Abstract

As massively parallel computations have become broadly available with modern GPUs, deep architectures trained on very large datasets have risen in popularity. Discriminatively trained convolutional neural networks, in particular, were recently shown to yield state-of-the-art performance in challenging image classification benchmarks such as ImageNet. However, elements of these architectures are similar to standard hand-crafted representations used in computer vision. In this paper, we explore the extent of this analogy, proposing a version of the state-of-the-art Fisher vector image encoding that can be stacked in multiple layers. This architecture significantly improves on standard Fisher vectors, and obtains competitive results with deep convolutional networks at a significantly smaller computational cost. Our hybrid architecture allows us to measure the performance improvement brought by a deeper image classification pipeline, while staying in the realms of conventional SIFT features and FV encodings.

65.26 Human vs. Computer in Scene and Object Recognition [29]

65.26.1 Original Abstract

Several decades of research in computer and primate vision have resulted in many models (some specialized for one problem, others more general) and invaluable experimental data. Here, to help focus research efforts onto the hardest unsolved problems, and bridge computer and human vision, we define a battery of 5 tests that measure the gap between human and machine performances in several dimensions (generalization across scene categories, generalization from images to edge maps and line drawings, invariance to rotation and scaling, local/global information with jumbled images, and object recognition performance). We measure model accuracy and the correlation between model and human error patterns. Experimenting over 7 datasets, where human data is available, and gauging 14 well-established models, we find that none fully resembles humans in all aspects, and we learn from each test which models and features are more promising in approaching humans in the tested dimension. Across all tests, we find that models based on local edge histograms consistently resemble humans more, while several scene statistics or "gist" models do perform well with both scenes and objects. While computer vision has long been inspired by human vision, we believe systematic efforts, such as this, will help better identify shortcomings of models and find new paths forward.

65.26.2 Main points

66 2014

66.1 Simultaneous Detection and Segmentation [93]

66.1.1 Original Abstract

We aim to detect all instances of a category in an image and, for each instance, mark the pixels that belong to it. We call this task Simultaneous Detection and Segmentation (SDS). Unlike classical bounding box detection, SDS requires a segmentation and not just a box. Unlike classical semantic segmentation, we require individual object instances. We build on recent work that uses convolutional neural networks to classify category-independent re-

gion proposals (R-CNN [16]), introducing a novel architecture tailored for SDS. We then use category-specific, top-down figure-ground predictions to refine our bottom-up proposals. We show a 7 point boost (16

66.1.2 Main points

- Segment one instance in a given image
- Work on top of region proposal R-CNN
- Dataset MSRC
- Mark each pixel belonging to the detected instance
- Algorithm: Simultaneous Detection and Segmentation
  - Proposal generation: 2.000 region candidates using MCG
  - Feature extraction: Extract features with pretrained CNN (Alexnet) with two paths, with and without background.
    - A : Extract CNN features from box and another with background masked
    - B : Second CNN is finetuned croping the box and removing background
    - C : Finetune both networks, one with the background and the other without
    - C + ref : refining the regions obtained from C
  - Region classification: linear SVM using fc6
  - Region refinement: non-maximum suppression on candidates and CNN for refinement
- Results
  - SegDPM detection PASCAL VOC2010: C+ref increases mean AP from 31.3 to 50.3
  - Pixel IU on VOC11: advance state-of-the-art about 5 points 10% relative

66.2 Part-based R-CNNs for fine-grained category detection [274]

66.2.1 Original Abstract

Semantic part localization can facilitate fine-grained categorization by explicitly isolating subtle appearance differences associated with specific object parts. Methods for pose-normalized representations have been proposed, but generally presume bounding box annotations at test time due to the difficulty of object detection. We propose a model for fine-grained categorization that overcomes these limitations by leveraging deep convolutional features computed on bottom-up region proposals. Our method learns whole-object and part detectors, enforces learned geometric constraints between them, and predicts a fine-grained category from a pose-normalized representation. Experiments on the Caltech-UCSD bird dataset confirm that our method outperforms state-of-the-art fine-grained categorization methods in an end-to-end evaluation without requiring a bounding box at test time.

66.2.2 Main points

- fine-grained category detection: detection and classification intra-class (Example: face recognition, dog breeds, and others)
- Got state-of-the-art without bounding box at test time
- Other approaches use Deformable Parts Model (DPM) plus engineered features (Example: HOG)
- Use of R-CNN to localize objects and generalizes to localize parts
- Use of Alexnet CNN pretrained with ImageNet and finetuned for detection
  - Substitute last fc8 1000 to 200
  - learning rate global = original:10
  - learning rate of fc8 globalx10
  - Decrease global by 10 during learning
- They add learned non-parametric geometric constraints

  - Mixture of Gaussians with 4 components and $\alpha = 0.1$
  - K nearest neighbors with $K = 20$
  - All hyperparameters found by 5 folds cross-validation
- Use of the fc6 to train the R-CNN object and part detector
- Use the pool5 for the geometric constraints
- Results
  - Caltech-UCSD bird with K-nearest Finetuning increases from 68.1% to 76%
  - Without bounding box at test time from 66% to 73.89%
- conclusion
  - For fine-grained discrimination is very useful pose and locality information
  - Future exploration on automatically discover and model parts as latent variables

66.3 Analyzing the performance of multilayer neural networks for object recognition [4]

66.3.1 Original Abstract

In the last two years, convolutional neural networks (CNNs) have achieved an impressive suite of results on standard recognition datasets and tasks. CNN-based features seem poised to quickly replace engineered representations, such as SIFT and HOG. However, compared to SIFT and HOG, we understand much less about the nature of the features learned by large CNNs. In this paper, we experimentally probe several aspects of CNN feature learning in an attempt to help practitioners gain useful, evidence-backed intuitions about how to apply CNNs to computer vision problems.

66.3.2 Main points

- Analysis of CNN (Alexnet)

- Findings
  - Effects of fine-tuning and pre-training:
    - Supervised pre-training is beneficial
    - Fine-tuning seems more significant for fc6 and fc7
  - ImageNet Pre-training does not Overfit:
    - pre-training time increases performance, and seems to not increase generalization error
    - For generalization quite quick 15k - 50k iterations (80%-90% of final performance)
  - Grandmother cells and distributed codes:
    - there are some "grandmother cells" for bicycle, person, cars and cats (from 15 to 30 filters)
    - but most of the features are distributed (from 30 to 40 filters)
  - Importance of feature location and magnitude:
    - CNN encoding:
      - Filters with non-zero response
      - Magnitude of the response
      - Spatial layout
    - spatial location critical for detection, but not for classification
    - Binarization gives similar results on fc6 and fc7 but not in early conv layers
    - Loosing spatial information drops performance on detection
- Datasets
  - PASCAL VOC 2007
  - SUN dataset
  - ImageNet (pretraining)

66.4 Rich feature hierarchies for accurate object detection and semantic segmentation [79]

66.4.1 Original Abstract

Object detection performance, as measured on the canonical PASCAL VOC dataset, has plateaued in the last few years. The best-performing methods are complex ensemble systems that typically combine multiple low-level image features with high-level context. In this paper, we propose a simple and scalable detection algorithm that improves mean average precision (mAP) by more than 30

66.4.2 Main points

- Region proposals using CNN (R-CNN)
- Object detection in three steps:

  Region proposal using selective search

  Feature extraction using Alexnet for 4096-dimentional feature vectors in each region (Caffe implementation)

  Classification using one-vs-all linear SVMs

- Pretraining Alexnet architecture
  - Changed last softmax layer from 1000 to 21 size (ILSVRC 2012 vs PASCAL)
  - Fine-tuning with PASCAL
  - mini-batch 128 (32 positive vs 96 background)
- Results on PASCAL VOC 2010-12
  - Compared to UVA system from Uijlings et. al. "Selective search for object recognition" Uijlings2013
  - Improved from 35.1% to 53.7% mAP
- Analysis of pretrained Alexnet
  - FC6 generalizes better than FC7

  - Good results if remove FC6 and FC7 (that only keeps 6% of the parameters)

- "[...] classical tools from computer vision and deep learning [...] the two are natural and inevitable partners."

66.5 Caffe: Convolutional architecture for fast feature embedding [121]

66.5.1 Original Abstract

Caffe provides multimedia scientists and practitioners with a clean and modifiable framework for state-of-the-art deep learning algorithms and a collection of reference models. The framework is a BSD-licensed C++ library with Python and MATLAB bindings for training and deploying general-purpose convolutional neural networks and other deep models efficiently on commodity architectures. Caffe fits industry and internet-scale media needs by CUDA GPU computation, processing over 40 million images a day on a single K40 or Titan GPU (≈ 2.5 ms per image). By separating model representation from actual implementation, Caffe allows experimentation and seamless switching among platforms for ease of development and deployment from prototyping machines to cloud environments. Caffe is maintained and developed by the Berkeley Vision and Learning Center (BVLC) with the help of an active community of contributors on GitHub. It powers ongoing research projects, large-scale industrial applications, and startup prototypes in vision, speech, and multimedia.

66.5.2 Main points

Comment: Tech report for the Caffe software at http://github.com/BVLC/Caffe/

66.6 Efficient Object Localization Using Convolutional Networks [245]

66.6.1 Original Abstract

Recent state-of-the-art performance on human-body pose estimation has been achieved with Deep Convolutional Networks (ConvNets). Traditional ConvNet architectures include pooling layers which reduce computational require-

ments, introduce invariance and prevent over-training. These benefits of pooling come at the cost of reduced localization accuracy. We introduce a novel architecture which includes an efficient 'position refinement' model that is trained to estimate the joint offset location within a small region of the image. This refinement model is jointly trained in cascade with a state-of-the-art ConvNet model to achieve improved accuracy in human joint location estimation. We show that the variance of our detector approaches the variance of human annotations on the FLIC dataset and outperforms all existing approaches on the MPII-human-pose dataset.

66.6.2 Main points

Comment: 8 pages with 1 page of citations

66.7 Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition [97]

66.7.1 Original Abstract

Existing deep convolutional neural networks (CNNs) require a fixed-size (e.g., 224x224) input image. This requirement is "artificial" and may reduce the recognition accuracy for the images or sub-images of an arbitrary size/scale. In this work, we equip the networks with a more principled pooling strategy, "spatial pyramid pooling", to eliminate the above requirement. The new network structure, called SPP-net, can generate a fixed-length representation regardless of image size/scale. Pyramid pooling is also robust to object deformations. With these advantages, SPP-net should in general improve all CNN-based image classification methods. On the ImageNet 2012 dataset, we demonstrate that SPP-net boosts the accuracy of a variety of published CNN architectures despite their different designs. On the Pascal VOC 2007 and Caltech101 datasets, SPP-net achieves state-of-the-art classification results using a single full-image representation and no fine-tuning. The power of SPP-net is also significant in object detection. Using SPP-net, we compute the feature maps from the entire image only once, and then pool features in arbitrary regions (sub-images) to generate fixed-length representations for training the detectors. This method avoids repeatedly computing the convolutional features. In processing test images, our method computes convolutional features 30-170x faster than the recent and most accurate method

R-CNN (and 24-64x faster overall), while achieving better or comparable accuracy on Pascal VOC 2007. In ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014, our methods rank 2 in object detection and 3 in image classification among all 38 teams. This manuscript also introduces the improvement made for this competition.

66.7.2 Main points

Comment: This manuscript (v2) is an extended technical report of our ECCV 2014 paper. This manuscript introduces the details of our methods for ILSVRC 2014 (rank 2 in DET and 3 in CLS)

66.8 Dropout: A Simple Way to Prevent Neural Networks from Overfitting [233]

66.8.1 Original Abstract

Deep neural nets with a large number of parameters are very powerful machine learning systems. However, overfitting is a serious problem in such networks. Large networks are also slow to use, making it difficult to deal with overfitting by combining the predictions of many different large neural nets at test time. Dropout is a technique for addressing this problem. The key idea is to randomly drop units (along with their connections) from the neural network during training. This prevents units from co-adapting too much. During training, dropout samples from an exponential number of different "thinned" networks. At test time, it is easy to approximate the effect of averaging the predictions of all these thinned networks by simply using a single unthinned network that has smaller weights. This significantly reduces overfitting and gives major improvements over other regularization methods. We show that dropout improves the performance of neural networks on supervised learning tasks in vision, speech recognition, document classification and computational biology, obtaining state-of-the-art results on many benchmark data sets.

66.9 DeepFace: Closing the Gap to Human-Level Performance in Face Verification [242]

66.9.1 Original Abstract

None

66.10 Return of the Devil in the Details: Delving Deep into Convolutional Nets [40]

66.10.1 Original Abstract

The latest generation of Convolutional Neural Networks (CNN) have achieved impressive results in challenging benchmarks on image recognition and object detection, significantly raising the interest of the community in these methods. Nevertheless, it is still unclear how different CNN methods compare with each other and with previous state-of-the-art shallow representations such as the Bag-of-Visual-Words and the Improved Fisher Vector. This paper conducts a rigorous evaluation of these new techniques, exploring different deep architectures and comparing them on a common ground, identifying and disclosing important implementation details. We identify several useful properties of CNN-based representations, including the fact that the dimensionality of the CNN output layer can be reduced significantly without having an adverse effect on performance. We also identify aspects of deep and shallow methods that can be successfully shared. A particularly significant one is data augmentation, which achieves a boost in performance in shallow methods analogous to that observed with CNN-based methods. Finally, we are planning to provide the configurations and code that achieve the state-of-the-art performance on the PASCAL VOC Classification challenge, along with alternative configurations trading-off performance, computation speed and compactness.

66.11 Deep Learning in Neural Networks: An Overview [215]

66.11.1 Original Abstract

In recent years, deep artificial neural networks (including recurrent ones) have won numerous contests in pattern recognition and machine learning. This historical survey compactly summarises relevant work, much of it from

the previous millennium. Shallow and deep learners are distinguished by the depth of their credit assignment paths, which are chains of possibly learnable, causal links between actions and effects. I review deep supervised learning (also recapitulating the history of backpropagation), unsupervised learning, reinforcement learning
evolutionary computation, and indirect search for short programs encoding deep and large networks.

66.12 Deep Learning: Methods and Applications [?]

66.12.1 Original Abstract

This book is aimed to provide an overview of general deep learning methodology and its applications to a variety of signal and information processing tasks. The application areas are chosen with the following three criteria: 1) expertise or knowledge of the authors; 2) the application areas that have already been transformed by the successful use of deep learning technology, such as speech recognition and computer vision; and 3) the application areas that have the potential to be impacted significantly by deep learning and that have gained concentrated research efforts, including natural language and text processing, information retrieval, and multimodal information processing empowered by multi-task deep learning. In Chapter 1, we provide the background of deep learning, as intrinsically connected to the use of multiple layers of nonlinear transformations to derive features from the sensory signals such as speech and visual images. In the most recent literature, deep learning is embodied also as representation learning, which involves a hierarchy of features or concepts where higher-level representations of them are defined from lower-level ones and where the same lower-level representations help to define higher-level ones. In Chapter 2, a brief historical account of deep learning is presented. In particular, selected chronological development of speech recognition is used to illustrate the recent impact of deep learning that has become a dominant technology in speech recognition industry within only a few years since the start of a collaboration between academic and industrial researchers in applying deep learning to speech recognition. In Chapter 3, a three-way classification scheme for a large body of work in deep learning is developed. We classify a growing number of deep learning techniques into unsupervised, supervised, and hybrid categories, and present qualitative descriptions and a literature survey for each category. From Chapter 4 to Chapter 6, we discuss

in detail three popular deep networks and related learning methods, one in each category. Chapter 4 is devoted to deep autoencoders as a prominent example of the unsupervised deep learning techniques. Chapter 5 gives a major example in the hybrid deep network category, which is the discriminative feed-forward neural network for supervised learning with many layers initialized using layer-by-layer generative, unsupervised pre-training. In Chapter 6, deep stacking networks and several of the variants are discussed in detail, which exemplify the discriminative or supervised deep learning techniques in the three-way categorization scheme. In Chapters 7-11, we select a set of typical and successful applications of deep learning in diverse areas of signal and information processing and of applied artificial intelligence. In Chapter 7, we review the applications of deep learning to speech and audio processing, with emphasis on speech recognition organized according to several prominent themes. In Chapters 8, we present recent results of applying deep learning to language modeling and natural language processing. Chapter 9 is devoted to selected applications of deep learning to information retrieval including Web search. In Chapter 10, we cover selected applications of deep learning to image object recognition in computer vision. Selected applications of deep learning to multi-modal processing and multi-task learning are reviewed in Chapter 11. Finally, an epilogue is given in Chapter 12 to summarize what we presented in earlier chapters and to discuss future challenges and directions.

66.13 Learning Multi-modal Latent Attributes [71]

66.13.1 Original Abstract

The rapid development of social media sharing has created a huge demand for automatic media classification and annotation techniques. Attribute learning has emerged as a promising paradigm for bridging the semantic gap and addressing data sparsity via transferring attribute knowledge in object recognition and relatively simple action classification. In this paper, we address the task of attribute learning for understanding multimedia data with sparse and incomplete labels. In particular, we focus on videos of social group activities, which are particularly challenging and topical examples of this task because of their multimodal content and complex and unstructured nature relative to the density of annotations. To solve this problem, we 1) introduce a concept of semilatent attribute space, expressing user-defined and latent at-

tributes in a unified framework, and 2) propose a novel scalable probabilistic topic model for learning multimodal semilatent attributes, which dramatically reduces requirements for an exhaustive accurate attribute ontology and expensive annotation effort. We show that our framework is able to exploit latent attributes to outperform contemporary approaches for addressing a variety of realistic multimedia sparse data learning tasks including: multitask learning, learning with label noise, N-shot transfer learning, and importantly zero-shot learning.

66.14 On the saddle point problem for non-convex optimization [195]

66.14.1 Original Abstract

A central challenge to many fields of science and engineering involves minimizing non-convex error functions over continuous, high dimensional spaces. Gradient descent or quasi-Newton methods are almost ubiquitously used to perform such minimizations, and it is often thought that a main source of difficulty for the ability of these local methods to find the global minimum is the proliferation of local minima with much higher error than the global minimum. Here we argue, based on results from statistical physics, random matrix theory, and neural network theory, that a deeper and more profound difficulty originates from the proliferation of saddle points, not local minima, especially in high dimensional problems of practical interest. Such saddle points are surrounded by high error plateaus that can dramatically slow down learning, and give the illusory impression of the existence of a local minimum. Motivated by these arguments, we propose a new algorithm, the saddle-free Newton method, that can rapidly escape high dimensional saddle points, unlike gradient descent and quasi-Newton methods. We apply this algorithm to deep neural network training, and provide preliminary numerical evidence for its superior performance.

66.15 Feature selection and hierarchical classifier design with applications to human motion recognition [68]

66.15.1 Original Abstract

The performance of a classifier is affected by a number of factors including classifiertype, the input features and the desired output. This thesis examines the impact of featureselection and classification problem division on classification accuracy and complexity.Proper feature selection can reduce classifier size and improve classifier performanceby minimizing the impact of noisy, redundant and correlated features. Noisy features cancause false association between the features and the classifier output. Redundant andcorrelated features increase classifier complexity without adding additional information.Output selection or classification problem division describes the division of a large clas-sification problem into a set of smaller problems. Problem division can improve accuracyby allocating more resources to more difficult class divisions and enabling the use of morespecific feature sets for each subproblem.The first part of this thesis presents two methods for creating featureselected hierarchi-cal classifiers. The feature-selected hierarchical classification method jointly optimizes thefeatures and classification tree-design using genetic algorithms. The multi-modal binarytree (MBT) method performs the class division and feature selection sequentially and tol-erates misclassifications in the higher nodes of the tree. This yields a piecewise separationfor classes that cannot be fully separated with a single classifier. Experiments show thatthe accuracy of MBT is comparable to other multi-class extensions, but with lower testtime. Furthermore, the accuracy of MBT is significantly higher on multi-modal data sets.The second part of this thesis focuses on input feature selection measures. A numberof filter-based feature subset evaluation measures are evaluated with the goal of assessingtheir performance with respect to specific classifiers. Although there are many featureselection measures proposed in literature, it is unclear which feature selection measuresare appropriate for use with different classifiers. Sixteen common filter-based measures aretested on 20 real and 20 artificial data sets, which are designed to probe for specific featureselection challenges. The strengths and weaknesses of each measure are discussed withrespect to the specific feature selection challenges in the artificial data sets, correlationwith classifier accuracy and their ability to identify known informative features.The results indicate that the best filter

measure is classifier-specific. K-nearest neigh-bours classifiers work well with subset-based RELIEF, correlation feature selection or con-ditional mutual information maximization, whereas Fisher's interclass separability criterionand conditional mutual information maximization work better for support vector machines.Based on the results of the feature selection experiments, two new filter-based measuresare proposed based on conditional mutual information maximization, which performs well but cannot identify dependent features in a set and does not include a check for corre-lated features. Both new measures explicitly check for dependent features and the secondmeasure also includes a term to discount correlated features. Both measures correctly iden-tify known informative features in the artificial data sets and correlate well with classifieraccuracy.The final part of this thesis examines the use of feature selection for time-series databy using feature selection to determine important individual time windows or key framesin the series. Time-series feature selection is used with the MBT algorithm to createclassification trees for time-series data. The feature selected MBT algorithm is tested ontwo human motion recognition tasks: full-body human motion recognition from joint angledata and hand gesture recognition from electromyography data. Results indicate that thefeature selected MBT is able to achieve high classification accuracy on the time-series datawhile maintaining a short test time.

66.16 Deep Learning: Methods and Applications [54]

66.16.1 Original Abstract

This book is aimed to provide an overview of general deep learning methodology and its applications to a variety of signal and information processing tasks. The application areas are chosen with the following three criteria: 1) expertise or knowledge of the authors; 2) the application areas that have already been transformed by the successful use of deep learning technology, such as speech recognition and computer vision; and 3) the application areas that have the potential to be impacted significantly by deep learning and that have gained concentrated research efforts, including natural language and text processing, information retrieval, and multimodal information processing empowered by multi-task deep learning. In Chapter 1, we provide the background of deep learning, as intrinsically connected to the use of multiple layers of nonlinear transformations to derive features from the sensory signals such as speech and visual images. In the most recent literature, deep learning is

embodied also as representation learning, which involves a hierarchy of features or concepts where higher-level representations of them are defined from lower-level ones and where the same lower-level representations help to define higher-level ones. In Chapter 2, a brief historical account of deep learning is presented. In particular, selected chronological development of speech recognition is used to illustrate the recent impact of deep learning that has become a dominant technology in speech recognition industry within only a few years since the start of a collaboration between academic and industrial researchers in applying deep learning to speech recognition. In Chapter 3, a three-way classification scheme for a large body of work in deep learning is developed. We classify a growing number of deep learning techniques into unsupervised, supervised, and hybrid categories, and present qualitative descriptions and a literature survey for each category. From Chapter 4 to Chapter 6, we discuss in detail three popular deep networks and related learning methods, one in each category. Chapter 4 is devoted to deep autoencoders as a prominent example of the unsupervised deep learning techniques. Chapter 5 gives a major example in the hybrid deep network category, which is the discriminative feed-forward neural network for supervised learning with many layers initialized using layer-by-layer generative, unsupervised pre-training. In Chapter 6, deep stacking networks and several of the variants are discussed in detail, which exemplify the discriminative or supervised deep learning techniques in the three-way categorization scheme. In Chapters 7-11, we select a set of typical and successful applications of deep learning in diverse areas of signal and information processing and of applied artificial intelligence. In Chapter 7, we review the applications of deep learning to speech and audio processing, with emphasis on speech recognition organized according to several prominent themes. In Chapters 8, we present recent results of applying deep learning to language modeling and natural language processing. Chapter 9 is devoted to selected applications of deep learning to information retrieval including Web search. In Chapter 10, we cover selected applications of deep learning to image object recognition in computer vision. Selected applications of deep learning to multi-modal processing and multi-task learning are reviewed in Chapter 11. Finally, an epilogue is given in Chapter 12 to summarize what we presented in earlier chapters and to discuss future challenges and directions.

66.17 Large-scale Video Classification with Convolutional Neural Networks [124]

66.17.1 Original Abstract

Convolutional Neural Networks (CNNs) have been es-tablished as a powerful class of models for image recog-nition problems. Encouraged by these results, we pro-vide an extensive empirical evaluation of CNNs on large- scale video classification using a new dataset of 1 millionYouTube videos belonging to 487 classes. We study mul-tiple approaches for extending the connectivity of a CNNin time domain to take advantage of local spatio-temporalinformation and suggest a multiresolution, foveated archi-tecture as a promising way of speeding up the training.Our best spatio-temporal networks display significant per-formance improvements compared to strong feature-basedbaselines (55.3

66.17.2 Main points

- Compare different CNN architectures for video classification
- Create a new dataset with 1 million of YouTube sport videos and 487 classes
- They required one month of training
- Multiresolution CNNs: New CNN with low resolution context and high resolution center
  - Context stream: seems to learn color filters
  - Fovea stream: learns grayscale features
- Compare with and without pretraining on other dataset UCF-101
- Architectures (increasing spatio-temporal relations)
  - Single frame: Classify with one single shot
  - Late Fusion: Classify with separate-in-time shots
  - Early Fusion: Classify with adjacent shots merging on first convolution layer
  - Slow Fusion: Classify with adjacent shots progressively mergin in upper layers

- Results (best models):
  - clip Hit, Video Hit, Video Hit top5
  - 42.4 60.0 78.5 Single-Frame + Multiresolution
  - 41.9 60.9 80.2 Slow Fusion
- Results on UCF-101 with pretraining:
  - 41.3 No pretraining
  - 64.1 Fine-tune top layer
  - 65.4 Fine-tune top 3 layers
  - 62.2 Fine-tune all layers
- Conclusions:
  - From video classification can be derived that camera movements deteriorate the predictions
  - Single frame gives very good results
- Further work:
  - Apply some filter for camera movements
  - Explore RNN from clip-level into video-level

66.18 Spectral Networks and Deep Locally Connected Networks on Graphs [33]

66.18.1 Original Abstract

Convolutional Neural Networks are extremely efficient architectures in image and audio recognition tasks, thanks to their ability to exploit the local translational invariance of signal classes over their domain. In this paper we consider possi- ble generalizations of CNNs to signals defined on more general domains without the action of a translation group. In particular, we propose two constructions, one based upon a hierarchical clustering of the domain, and another based on the spectrum of the graph Laplacian. We show through experiments that for low- dimensional graphs it is possible to learn convolutional layers with a number of parameters independent of the input size, resulting in efficient deep architectures.

66.19 Towards Real-Time Image Understanding with Convolutional Networks [62]

66.19.1 Original Abstract

One of the open questions of artificial computer vision is how to produce good internal representations of the visual world. What sort of internal representation would allow an artificial vision system to detect and classify objects into categories, independently of pose, scale, illumination, conforma- tion, and clutter? More interestingly, how could an artificial vision system learn appropriate internal representations automatically, the way animals and humans seem to learn by simply looking at the world? Another related question is that of computational tractability, and more precisely that of computational efficiency. Given a good visual represen- tation, how efficiently can it be trained, and used to encode new sensorial data. Efficiency has several dimensions: power requirements, processing speed, and memory usage. In this thesis I present three new contributions to the field of computer vision: (1) a multiscale deep convolutional network architecture to easily capture long-distance relationships between input variables in image data, (2) a tree-based algorithm to efficiently explore multiple segmentation can- didates, to produce maximally confident semantic segmentations of images, (3) a custom dataflow computer architecture optimized for the computation of convolutional networks, and similarly dense image processing models. All three contributions were produced with the common goal of getting us closer to real-time image understanding. Scene parsing consists in labeling each pixel in an image with the category of the object it belongs to. In the first part of this thesis, I propose a method that uses a multiscale convolutional network trained from raw pixels to extract dense feature vectors that encode regions of multiple sizes centered on each pixel. The method alleviates the need for engineered features. Inparallel to feature extraction, a tree of segments is computed from a graph of pixel dissimilarities. The feature vectors associated with the segments covered by each node in the tree are aggregated and fed to a classifier which produces an estimate of the distribution of object categories contained in the segment. A subset of tree nodes that cover the image are then selected so as to maximize the average "purity" of the class distributions, hence maximizing the overall likelihood that each segment contains a single object. The system yields record accuracies on several public benchmarks. The computation of convolutional networks, and related models

heavily relies on a set of basic operators that are particularly fit for dedicated hardware implementations. In the second part of this thesis I introduce a scalable dataflow hardware architecture optimized for the computation of general-purpose vision algorithms—neuFlow —and a dataflow compiler—luaFlow —that transforms high-level flow-graph representations of these algorithms into machine code for neuFlow. This system was designed with the goal of providing real-time detection, categorization and localization of objects in complex scenes, while consuming 10 Watts when implemented on a Xilinx Virtex 6 FPGA platform, or about ten times less than a lap- top computer, and producing speedups of up to 100 times in real-world applications (results from 2011).

66.20 Learning Deep Face Representation [59]

66.20.1 Original Abstract

Face representation is a crucial step of face recognition systems. An optimal face representation should be discriminative, robust, compact, and very easy-to-implement. While numerous hand-crafted and learning-based representations have been proposed, considerable room for improvement is still present. In this paper, we present a very easy-to-implement deep learning framework for face representation. Our method bases on a new structure of deep network (called Pyramid CNN). The proposed Pyramid CNN adopts a greedy-filter-and-down-sample operation, which enables the training procedure to be very fast and computation-efficient. In addition, the structure of Pyramid CNN can naturally incorporate feature sharing across multi-scale face representations, increasing the discriminative ability of resulting representation. Our basic network is capable of achieving high recognition accuracy (85.8

66.20.2 Main points

- New deep structure Pyramid CNN
- Labeled Faces in the Wild (LFW)
  - $>$ 13.000 faces
  - 1680 of the people have two or more distinct photos
  - Detected by Viola-Jones detector

  - http://vis-www.cs.umass.edu/lfw/
- State-of-the-art performance on LFW benchmark (97.3%)
- Good face representation
  - Identity-preserving: Same person pictures close in feature space
  - Abstract and Compact: from high to low dimensionality
  - Uniform and Automatic: NO hand-crafted and hard-wired parts
- Pyramid CNN
  - ID-preserving Representation Learning: Loss functions measures distance in output feature space
  - Convolutions and Down-sampling
  - Deeper give best results, but increases rapidly the training time
  - Each CNN own private output layer and gets the input from the previous shared layer
  - Only the output of the last level network is used for the represetnation
  - The rest of the outputs is just for training
- Results
  - 164 incorrect predictions
  - Some of them are incorrectly labeled
  - Others are very difficult for humans, because of the age or pose
  - On LFW benchmark achieves state-of-the-art and close to human on croped images
- With ROC curve as a mesure there is an improvement of 0.07-0.12 with Baseline
- Face recognition does not contemplate affine transformations or perspectives,
- Can be difficult to apply in task such as ImageNet, where the object can be in any place and position

66.21 OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks [220]

66.21.1 Original Abstract

We present an integrated framework for using Convolutional Networks for classification, localization and detection. We show how a multiscale and sliding window approach can be efficiently implemented within a ConvNet. We also introduce a novel deep learning approach to localization by learning to predict object boundaries. Bounding boxes are then accumulated rather than suppressed in order to increase detection confidence. We show that different tasks can be learned simultaneously using a single shared network. This integrated framework is the winner of the localization task of the ImageNet Large Scale Visual Recognition Challenge 2013 (ILSVRC2013) and obtained very competitive results for the detection and classifications tasks. In post-competition work, we establish a new state of the art for the detection task. Finally, we release a feature extractor from our best model called OverFeat.

66.21.2 Main points

- Framework for using CNN
  - classification
  - localization
  - detection
- Winner on localization task of ILSVRC2013
- ConvNets are trained enterily with the raw pixels
- Other approaches for detection and localization
- appling a sliding window over multiples scales
- ...
- ...

66.22 LSDA: Large Scale Detection through Adaptation [107]

66.22.1 Original Abstract

A major challenge in scaling object detection is the difficulty of obtaining labeled images for large numbers of categories. Recently, deep convolutional neural networks (CNNs) have emerged as clear winners on object classification benchmarks, in part due to training with 1.2M+ labeled classification images. Unfortunately, only a small fraction of those labels are available for the detection task. It is much cheaper and easier to collect large quantities of image-level labels from search engines than it is to collect detection data and label it with precise bounding boxes. In this paper, we propose Large Scale Detection through Adaptation (LSDA), an algorithm which learns the difference between the two tasks and transfers this knowledge to classifiers for categories without bounding box annotated data, turning them into detectors. Our method has the potential to enable detection for the tens of thousands of categories that lack bounding box annotations, yet have plenty of classification data. Evaluation on the ImageNet LSVRC-2013 detection challenge demonstrates the efficacy of our approach. This algorithm enables us to produce a >7.6K detector by using available classification data from leaf nodes in the ImageNet tree. We additionally demonstrate how to modify our architecture to produce a fast detector (running at 2fps for the 7.6K detector). Models and software are available at

66.22.2 Main points

- Converting a classifier into a detector
- ImageNet only contains 200 annotated classes for detection
- Other approaches Multiple Instance Learning
- Take Alexnet change last layer to desired number of classes and finetune
- Finetune for detection using also background class
- Compute category score as $\text{score}_{c}ategory - score_{b}ackgroundExperiment$
- • ILSVRC2013 detection dataset

- 1.000 images per class
- 200 categories
- val1 : 100 categories with bounding box for detection training
- val2 : 100 categories for evaluation

Results

- full LSDA 50% relative mAP boost over only classifier
- Classifier only focus on most discriminative parts (Ex: face of an animal)
- After detection finetuning detects all the body
- False positive errors
  - localization errors (Loc):
  - confusion with background (BG):
  - other (Oth): Most of errors because confusion of the class

They released the 7.6K model for detection in lsda.berkeleyvision.org

minimize the gap between classifiers and detectors

66.23 Deformable part models are convolutional neural networks [80]

66.23.1 Original Abstract

Deformable part models (DPMs) and convolutional neural networks (CNNs) are two widely used tools for visual recognition. They are typically viewed as distinct approaches: DPMs are graphical models (Markov random fields), while CNNs are "black-box" non-linear classifiers. In this paper, we show that a DPM can be formulated as a CNN, thus providing a novel synthesis of the two ideas. Our construction involves unrolling the DPM inference algorithm and mapping each step to an equivalent (and at times novel) CNN layer. From this perspective, it becomes natural to replace the standard image features used in DPM with a learned feature extractor. We call the resulting model DeepPyramid DPM and experimentally validate it on PASCAL VOC. DeepPyramid DPM significantly outperforms DPMs based on histograms of oriented gradients features (HOG) and slightly outperforms a comparable version of the recently introduced R-CNN detection system, while running an order of magnitude faster.

66.23.2 Main points

- A DPM can be expresed as a CNN
- when using the new distance transform pooling that generalizes max pooling
- and maxout units
- DeepPyramid DPM takes an image pyramid and produces a pyramid of object detectors
- Instead of using HOG uses a CNN
  - from pretrained Alexnet (CNN)
  - remove fc6, fc7, fc8 and pool5
  - only interested on conv5 (256 feature channels)
  - Each pixel on conv5 feature map corresponds to 16 pixels in the original image
- Create the image pyramid
  - Resize image largest dimension to 1.713 pixels
  - Conv5 sees 108 cells in longest side
  - 7 pyramid levels with scale factor 1/sqrt(2)
  - total of 25k output cells per image
  - For comparison: 1k5 in OverFeat and 250k commonly with HOG
- Results
  - Conv5 only fires to certain scales per class
  - On the other hand, HOG in all scales

66.24 Do Convnets Learn Correspondence? [160]

66.24.1 Original Abstract

Convolutional neural nets (convnets) trained from massive labeled datasets have substantially improved the state-of-the-art in image classification and

object detection. However, visual understanding requires establishing correspondence on a finer level than object category. Given their large pooling regions and training from whole-image labels, it is not clear that convnets derive their success from an accurate correspondence model which could be used for precise localization. In this paper, we study the effectiveness of convnet activation features for tasks requiring correspondence. We present evidence that convnet features localize at a much finer scale than their receptive field sizes, that they can be used to perform intraclass aligment as well as conventional hand-engineered features, and that they outperform conventional features in keypoint prediction on objects from PASCAL VOC 2011.

66.25 Two-stream convolutional networks for action recognition in videos [227]

66.25.1 Original Abstract

We investigate architectures of discriminatively trained deep Convolutional Networks (ConvNets) for action recognition in video. The challenge is to capture the complementary information on appearance from still frames and motion between frames. We also aim to incorporate into the network design aspects of the best performing hand-crafted features. Our contribution is three-fold. First, we propose a two-stream ConvNet architecture which incorporates spatial and temporal networks. Second, we demonstrate that a ConvNet trained on multi-frame dense optical flow is able to achieve very good performance in spite of limited training data. Finally, we show that multi-task learning, applied to two different action classification datasets, can be used to increase the amount of training data and improve the performance on both. Our architecture is trained and evaluated on the standard video actions benchmarks of UCF-101 and HMDB-51, where it matches the state of the art. It also exceeds by a large margin previous attempts to use deep nets for video classification.

66.25.2 Main points

- Action recognition using two paths on a CNN
- First one using only frames
- Second one using optical flow

- Datasets
  - ImageNet ILSVRC-2012 (pretraining)
  - UCF-101: 9.5K videos
  - HMDB-15: 3.7K videos
- Biological inspiration by two paths on our visual cortex
  - Ventral stream performs object recognition
  - Dorsal stream recognises motion
- Action recognition approaches commonly use
  - High dimensional encodings of spatio-temporal features
  - Classification with shallow methods
  - Some of the features extracted by:
    * Histogram of Oriented Gradients (HOG)
    * Histogram of Optical Flow (HOF)
  - Then features merged with Bag Of Features (BoF)
  - Final classification using SVM
  - state-of-the-art Motion Boundary Histogram (MBH)
  - Compensation of camera motion is very important
  - Fisher vector encodings (deep version on [226]
- Two methods for merging the two CNN softmax layers
  - Averaging their outputs
  - Training a multi-class linear SVM
- The two CNN
  - Spatial stream ConvNet:
    * Individual frames
    * It can be pretrained with image datasets (Ex: ImageNet)
  - Optical flow ConvNet options:

    - Optical flow stacking: From L frames extract L+1 optical flow input channels
    - Trajectory stacking: This follows the optical flows as following the different objects. -> ——> —> -> ->
    - Bi-directional optical flow: Like mirroring in images, it is possible to use Forward and Backward optical flows (data augmentation?)
    - Mean flow subtraction: To center the inputs on the non-linearity center. Accentuated sometimes by camera motion. They solve this problem subtracting the mean of each displacement, this is less computational costly, but also less precise

- Multi-task learning
  - On top of the CNN two softmax layers are added
  - One is only trained for HMDB-51 while the other on UCF-101
- Training
  - Random crop
  - Random mirroring
  - Random RGB jittering
  - learning rate: 0.01, 0.001, 0.0001
- Results
  - Pretraining with ILSVRC-2012 improves results
  - Optical flow in general works better than extracting this information from pairs of images
  - Temporal and spatial information is complementary
  - Augmenting the data is very beneficial
  - Pretraining with large amounts of images improves the generalization

66.26 Deep Networks with Internal Selective Attention through Feedback Connections [234]

66.26.1 Original Abstract

Traditional convolutional neural networks (CNN) are stationary and feedforward. They neither change their parameters during evaluation nor use feedback from higher to lower layers. Real brains, however, do. So does our Deep Attention Selective Network (dasNet) architecture. DasNet's feedback structure can dynamically alter its convolutional filter sensitivities during classification. It harnesses the power of sequential processing to improve classification performance, by allowing the network to iteratively focus its internal attention on some of its convolutional filters. Feedback is trained through direct policy search in a huge million-dimensional parameter space, through scalable natural evolution strategies (SNES). On the CIFAR-10 and CIFAR-100 datasets, dasNet outperforms the previous state-of-the-art model on unaugmented datasets.

66.26.2 Main points

66.27 How transferable are features in deep neural networks? [270]

66.27.1 Original Abstract

Many deep neural networks trained on natural images exhibit a curious phenomenon in common: on the first layer they learn features similar to Gabor filters and color blobs. Such first-layer features appear not to be specific to a particular dataset or task, but general in that they are applicable to many datasets and tasks. Features must eventually transition from general to specific by the last layer of the network, but this transition has not been studied extensively. In this paper we experimentally quantify the generality versus specificity of neurons in each layer of a deep convolutional neural network and report a few surprising results. Transferability is negatively affected by two distinct issues: (1) the specialization of higher layer neurons to their original task at the expense of performance on the target task, which was expected, and (2) optimization difficulties related to splitting networks between co-adapted neurons, which was not expected. In an example network trained on ImageNet, we demonstrate that either of these two issues may dominate,

depending on whether features are transferred from the bottom, middle, or top of the network. We also document that the transferability of features decreases as the distance between the base task and target task increases, but that transferring features even from distant tasks can be better than using random features. A final surprising result is that initializing a network with transferred features from almost any number of layers can produce a boost to generalization that lingers even after fine-tuning to the target dataset.

66.28 Histograms of pattern sets for image classification and object recognition [256]

66.28.1 Original Abstract

This paper introduces a novel image representation capturing feature dependencies through the mining of meaningful combinations of visual features. This representation leads to a compact and discriminative encoding of images that can be used for image classification, object detection or object recognition. The method relies on (i) multiple random projections of the input space followed by local binarization of projected histograms encoded as sets of items, and (ii) the representation of images as Histograms of Pattern Sets (HoPS). The approach is validated on four publicly available datasets (Daimler Pedestrian, Oxford Flowers, KTH Texture and PASCAL VOC2007), allowing comparisons with many recent approaches. The proposed image representation reaches state-of-the-art performance on each one of these datasets.

66.28.2 Main points

- Pattern mining
- dimentionality reduction
- feature selection
- feature augmentation
- Common image representations with real-valued histograms
  - Local Binary Patterns (LBP)
  - Histograms of Oriented Gradients (HOG)

  - Bag-of-Words (BoW)

- Authors propose Histograms of Pattern Sets (HoPS).
  - Extract some features from the images, for example BoW.
  - Randomly select N features (in this case visual words)
  - Binarize the selected feature histograms
    * The top-K selected features with higher occurrencies are set to one.
    * The rest is set to zero.
    * Group the features with value 1 as a transaction of size K
  - Repeat the random selection P times and create one transaction at each step
  - Apply data mining techniques to select the most discriminative transactions, for example:
    * Frequent Patterns (FPs) [5]
    * Jumping Emerging Patterns (JEPs) [56]
      · positive JEPs: random projections that are found only in the positive images
      · negative JEPs: random projections that are found onluy in the negative images
  - The final representaiton is a histogram of 2xP bins where P are the total number of projections and one positive JEP and negative JEP per projection
  - Train a classifier with this representation
  - Results
    * A linear SVM trained with HoPS improved the performance on the original features from 68.8 to 74.1
  * A RBF-Chi$^2 trained with HoPs improved the performance on the original features from 71$
  * Image classification: state-of-the-art in Oxford-Flowers 17 dataset
  - Texture recognition: good results on KTH-TIPS2a
  - Object detection: state-of-the-art in PASCAL VOC 2007 dataset
  - Pedestrian recognition: state-of-the-art in pedestrian recognition

66.29 Recurrent Models of Visual Attention [176]

66.29.1 Original Abstract

Applying convolutional neural networks to large images is computationally expensive because the amount of computation scales linearly with the number of image pixels. We present a novel recurrent neural network model that is capable of extracting information from an image or video by adaptively selecting a sequence of regions or locations and only processing the selected regions at high resolution. Like convolutional neural networks, the proposed model has a degree of translation invariance built-in, but the amount of computation it performs can be controlled independently of the input image size. While the model is non-differentiable, it can be trained using reinforcement learning methods to learn task-specific policies. We evaluate our model on several image classification tasks, where it significantly outperforms a convolutional neural network baseline on cluttered images, and on a dynamic visual control problem, where it learns to track a simple object without an explicit training signal for doing so.

66.30 From Captions to Visual Concepts and Back [60]

66.30.1 Original Abstract

This paper presents a novel approach for automatically generating image descriptions: visual detectors and language models learn directly from a dataset of image captions. We use Multiple Instance Learning to train visual detectors for words that commonly occur in captions, including many different parts of speech such as nouns, verbs, and adjectives. The word detector outputs serve as conditional inputs to a maximum-entropy language model. The language model learns from a set of over 400,000 image descriptions to capture the statistics of word usage. We capture global semantics by re-ranking caption candidates using sentence-level features and a deep multimodal similarity model. When human judges compare the system captions to ones written by other people, the system captions have equal or better quality over 23

66.30.2 Main points

Comment: Added appendix

66.31 Going deeper with convolutions [239]

66.31.1 Original Abstract

We propose a deep convolutional neural network architecture codenamed "Inception", which was responsible for setting the new state of the art for classification and detection in the ImageNet Large-Scale Visual Recognition Challenge 2014 (ILSVRC 2014). The main hallmark of this architecture is the improved utilization of the computing resources inside the network. This was achieved by a carefully crafted design that allows for increasing the depth and width of the network while keeping the computational budget constant. To optimize quality, the architectural decisions were based on the Hebbian principle and the intuition of multi-scale processing. One particular incarnation used in our submission for ILSVRC 2014 is called GoogLeNet, a 22 layers deep network, the quality of which is assessed in the context of classification and detection.

66.31.2 Main points

66.32 Deep Learning: Methods and Applications [54]

66.32.1 Original Abstract

This book is aimed to provide an overview of general deep learning methodology and its applications to a variety of signal and information processing tasks. The application areas are chosen with the following three criteria: 1) expertise or knowledge of the authors; 2) the application areas that have already been transformed by the successful use of deep learning technology, such as speech recognition and computer vision; and 3) the application areas that have the potential to be impacted significantly by deep learning and that have gained concentrated research efforts, including natural language and text processing, information retrieval, and multimodal information processing empowered by multi-task deep learning. In Chapter 1, we provide the background of deep learning, as intrinsically connected to the use of multiple layers of nonlinear transformations to derive features from the sensory signals such as speech and visual images. In the most recent literature, deep learning is embodied also as representation learning, which involves a hierarchy of features or concepts where higher-level representations of them are defined from lower-level ones and where the same lower-level representations help to define

higher-level ones. In Chapter 2, a brief historical account of deep learning is presented. In particular, selected chronological development of speech recognition is used to illustrate the recent impact of deep learning that has become a dominant technology in speech recognition industry within only a few years since the start of a collaboration between academic and industrial researchers in applying deep learning to speech recognition. In Chapter 3, a three-way classification scheme for a large body of work in deep learning is developed. We classify a growing number of deep learning techniques into unsupervised, supervised, and hybrid categories, and present qualitative descriptions and a literature survey for each category. From Chapter 4 to Chapter 6, we discuss in detail three popular deep networks and related learning methods, one in each category. Chapter 4 is devoted to deep autoencoders as a prominent example of the unsupervised deep learning techniques. Chapter 5 gives a major example in the hybrid deep network category, which is the discriminative feed-forward neural network for supervised learning with many layers initialized using layer-by-layer generative, unsupervised pre-training. In Chapter 6, deep stacking networks and several of the variants are discussed in detail, which exemplify the discriminative or supervised deep learning techniques in the three-way categorization scheme. In Chapters 7-11, we select a set of typical and successful applications of deep learning in diverse areas of signal and information processing and of applied artificial intelligence. In Chapter 7, we review the applications of deep learning to speech and audio processing, with emphasis on speech recognition organized according to several prominent themes. In Chapters 8, we present recent results of applying deep learning to language modeling and natural language processing. Chapter 9 is devoted to selected applications of deep learning to information retrieval including Web search. In Chapter 10, we cover selected applications of deep learning to image object recognition in computer vision. Selected applications of deep learning to multi-modal processing and multi-task learning are reviewed in Chapter 11. Finally, an epilogue is given in Chapter 12 to summarize what we presented in earlier chapters and to discuss future challenges and directions.

66.32.2 Main points

66.33 Very deep convolutional networks for large-scale image recognition [228]

66.33.1 Original Abstract

In this work we investigate the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting. Our main contribution is a thorough evaluation of networks of increasing depth, which shows that a significant improvement on the prior-art configurations can be achieved by pushing the depth to 16–19 weight layers. These findings were the basis of our ImageNet Challenge 2014 submission, where our team secured the first and the second places in the localisation and classification tracks respectively. We also show that our representations generalise well to other datasets, where they achieve the state-of-the-art results. Importantly, we have made our two best-performing ConvNet models publicly available to facilitate further research on the use of deep visual representations in computer vision.

66.33.2 Main points

66.34 Imagenet large scale visual recognition challenge [211]

66.34.1 Original Abstract

The ImageNet Large Scale Visual Recognition Challenge is a benchmark in object category classification and detection on hundreds of object categories and millions of images. The challenge has been run annually from 2010 to present, attracting participation from more than fifty institutions. This paper describes the creation of this benchmark dataset and the advances in object recognition that have been possible as a result. We discuss the challenges of collecting large-scale ground truth annotation, highlight key breakthroughs in categorical object recognition, provide detailed a analysis of the current state of the field of large-scale image classification and object detection, and compare the state-of-the-art computer vision accuracy with human accuracy. We conclude with lessons learned in the five years of the challenge, and propose future directions and improvements.

66.34.2 Main points

Comment: 37 pages, 14 figures

References

[1] K Aas and L Eikvil. Text categorisation: A survey. *Raport NR*, 1999.

[2] O Abdel-Hamid and A Mohamed. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. *Acoustics, Speech and ...*, 2012.

[3] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A Learning Algorithm for Boltzmann Machines*. *Cognitive Science*, 9(1):147–169, January 1985.

[4] Pulkit Agrawal, Ross Girshick, and Jitendra Malik. Analyzing the performance of multilayer neural networks for object recognition. *Computer Vision-ECCV 2014*, pages 329–344, January 2014.

[5] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, pages 207–216, 1993.

[6] S.-I. Amari. Characteristics of Random Nets of Analog Neuron-Like Elements. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-2(5):643–657, November 1972.

[7] Shunichi Amari. A Theory of Adaptive Pattern Classifiers. *IEEE Transactions on Electronic Computers*, EC-16(3):299–307, June 1967.

[8] JA Anderson. A simple neural network generating an interactive memory. *Mathematical Biosciences*, 14(3-4):197–220, August 1972.

[9] James A. Anderson and Edward Rosenfeld. *Neurocomputing: foundations of research*. MIT Press, Cambridge, MA, USA, 1988.

[10] Robert Andrews, Joachim Diederich, and Alan B. Tickle. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, 8(6):373–389, December 1995.

[11] R Arandjelovic, Andrew Zisserman, and Basura Fernando. AXES at TRECVid 2012: KIS, INS, and MED. 2012.

[12] A F De Araujo, F Silveira, H Lakshman, J Zepeda, A Sheth, and B Girod. The Stanford / Technicolor / Fraunhofer HHI Video. 2012.

[13] WR Ashby. Design for a Brain: The Origin of Adaptive Behavior. 1960.

[14] Stéphane Ayache, Georges Quénot, and Jérôme Gensel. Classifier fusion for SVM-based multimedia semantic indexing. *Springer Berlin Heidelberg*, 2007.

[15] LJ Ba and R Caurana. Do Deep Nets Really Need to be Deep? *arXiv preprint arXiv:1312.6184*, pages 1–6, 2013.

[16] Alexander Bain. *Mind and body. The theories of their relation.* New York : D. Appleton and company, 1873.

[17] A.G. Barto, Richard S. Sutton, and Charles W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *Systems, Man and ...*, SMC-13(5):834–846, September 1983.

[18] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and L Van Gool. Speeded-up robust features (SURF). *Computer vision and image ...*, 110(3):346–359, June 2008.

[19] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. *Computer Vision-ECCV 2006*, 2006.

[20] AJ Bell and TJ Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, November 1995.

[21] Yoshua Bengio. *Learning Deep Architectures for AI*, volume 2. 2009.

[22] Yoshua Bengio and Y LeCun. Scaling learning algorithms towards AI. *Large-Scale Kernel Machines*, (1):1–41, 2007.

[23] Yoshua Bengio, Éric Thibodeau-Laufer, Guillaume Alain, and Jason Yosinski. Deep Generative Stochastic Networks Trainable by Backprop. June 2013.

[24] Jeremy Bernstein. A. I. *The New Yorker*, page 50, December 1981.

[25] Karoly Bezdek, Antoine Deza, and Yinyu Ye. *Discrete geometry and optimization*. Springer Science & Business Media, July 2013.

[26] Christopher M. Bishop. *Pattern recognition and machine learning.*, volume 1. New York: springer, 2006., 2006.

[27] MA Boden. *Mind as machine: A history of cognitive science*, volume 1. 2006.

[28] JA Bogovic, GB Huang, and Viren Jain. Learned versus Hand-Designed Feature Representations for 3d Agglomeration. *arXiv preprint arXiv:1312.6159*, pages 1–14, 2013.

[29] Ali Borji and Laurent Itti. Human vs. Computer in Scene and Object Recognition. pages 113–120, 2013.

[30] BE Boser, IM Guyon, and VN Vapnik. A training algorithm for optimal margin classifiers. *... of the fifth annual workshop on ...*, pages 144–152, 1992.

[31] DS Broomhead and David Lowe. Radial basis functions, multi-variable functional interpolation and adaptive networks. March 1988.

[32] M. Brown and D.G. Lowe. Unsupervised 3D object recognition and reconstruction in unordered datasets. *3-D Digital Imaging and Modeling, 2005. ...*, pages 56–63, June 2005.

[33] Matthew Brown and DG Lowe. Automatic panoramic image stitching using invariant features. *International journal of computer vision*, 74(1):59–73, August 2007.

[34] J. Bruce, T. Balch, and M. Veloso. Fast and inexpensive color image segmentation for interactive robots. volume 3, pages 2061–2066 vol.3, 2000.

[35] Joan Bruna, Arthur Szlam, Wojciech Zaremba, and Yann LeCun. Spectral Networks and Deep Locally Connected Networks on Graphs. pages 1–14, 2014.

[36] P.J. Burt and E.H. Adelson. The Laplacian pyramid as a compact image code. *Communications, IEEE Transactions on*, 31(4):532–540, April 1983.

[37] Santiago Ramon y Cajal. Histologie du systeme nerveux de l'homme & des vertebres. page 1014, 1909.

[38] P Le Callet. A convolutional neural network approach for objective video quality assessment. *Neural Networks, IEEE ...*, 5:1316–1327, 2006.

[39] MA Carreira-Perpinan and Geoffrey Hinton. On contrastive divergence learning. *... on artificial intelligence and ...*, 0, 2005.

[40] D. Chai and A. Bouzerdoum. A Bayesian approach to skin color classification in YCbCr color space. volume 2, pages 421–424 vol.2, 2000.

[41] D. Chai and K.N. Ngan. Locating facial region of a head-and-shoulders color image. pages 124–129, April 1998.

[42] D. Chai and K.N. Ngan. Face segmentation using skin-color map in videophone applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(4):551–564, June 1999.

[43] Ken Chatfield and Karen Simonyan. Return of the Devil in the Details: Delving Deep into Convolutional Nets. *arXiv preprint arXiv: ...*, pages 1–11, 2014.

[44] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. *... Vision and Pattern Recognition ...*, 1:539–546 vol. 1, June 2005.

[45] PS Churchland and TJ Sejnowski. Perspectives on cognitive neuroscience. *Science*, 242(4879):741–745, November 1988.

[46] DC Ciresan, Alessandro Giusti, and Jürgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. *Medical Image ...*, 2013.

[47] B. Jack Copeland and Diane Proudfoot. Alan Turing\'s forgotten ideas in Computer Science. *Scientific American*, pages 99–103, 1999.

[48] BJ Copeland. *Alan Turing's Electronic Brain: The Struggle to Build the ACE, the World's Fastest Computer*. Oxford University Press, May 2012.

[49] BJ Copeland and Diane Proudfoot. On Alan Turing's anticipation of connectionism. *Synthese*, 108(3):361–377, September 1996.

[50] B. G. Cragg and H. N. V. Temperley. Memory: The Analogy with Ferromagnetic Hysteresis. *Brain*, 78(2):304–316, June 1955.

[51] Daniel Crevier. AI: The tumultuous history of the search for artificial intelligence. 1993.

[52] Gabriella Csurka and C Dance. Visual categorization with bags of keypoints. *Workshop on statistical . . .*, 2004.

[53] Paul Cull. The mathematical biophysics of Nicolas Rashevsky. *Biosystems*, 88(3):178–184, April 2007.

[54] Y Le Cun. A theoretical framework for back-propagation. 1988.

[55] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *. . . and Pattern Recognition, 2005. CVPR 2005 . . .*, 1:886–893 vol. 1, June 2005.

[56] Kostas Daniilidis, Petros Maragos, and Nikos Paragios. *Computer Vision–ECCV 2010*. 2010.

[57] Michael R. W. Dawson. *Connectionism: A Hands-on Approach*. John Wiley & Sons, April 2008.

[58] Peter Dayan, GE Hinton, RM Neal, and RS Zemel. The helmholtz machine. *Neural computation*, 904:889–904, 1995.

[59] Li Deng and Dong Yu. *Deep Learning: Methods and Applications*. 2014.

[60] Sander Dieleman, P Brakel, and Benjamin Schrauwen. Audio-based music classification with a pretrained convolutional network. *. . . International Society for Music . . .*, (Ismir):669–674, 2011.

[61] Guozhu Dong and Jinyan Li. Efficient mining of emerging patterns: Discovering trends and differences. *Proceedings of the fifth ACM SIGKDD international ...*, pages 43–52, 1999.

[62] David Eigen, Jason Rolfe, Rob Fergus, and Y LeCun. Understanding Deep Architectures using a Recursive Convolutional Network. *arXiv preprint arXiv:1312.1847*, pages 1–9, 2013.

[63] D Erhan, Yoshua Bengio, and Aaron Courville. Why does unsupervised pre-training help deep learning? *... of Machine Learning ...*, 9(2007):201–208, 2010.

[64] Mark Everingham and SMA Eslami. The pascal visual object classes challenge: A retrospective. *International Journal of ...*, 111(1):98–136, June 2014.

[65] Mark Everingham and Luc Van Gool. The pascal visual object classes (voc) challenge. *International journal of ...*, 88(2):303–338, September 2010.

[66] Haoqiang Fan, Zhimin Cao, Yunin Jiang, Qi Yin, C Doudou, and Chinchilla Doudou. Learning Deep Face Representation. *arXiv preprint arXiv:1403.2802*, pages 1–10, 2014.

[67] Hao Fang, Saurabh Gupta, and Forrest Iandola. From Captions to Visual Concepts and Back. *arXiv preprint arXiv: ...*, November 2014.

[68] C Farabet, Camille Couprie, Laurent Najman, and Y LeCun. Learning hierarchical features for scene labeling. 8:1915–1929, 2012.

[69] Clément Farabet. *Towards Real-Time Image Understanding with Convolutional Networks*. PhD thesis, Université Paris-Est, 2014.

[70] B.G. Farley and W. Clark. Simulation of self-organizing systems by digital computer. *... of the IRE Professional Group on*, 4(4):76–84, September 1954.

[71] MP Fay and MA Proschan. Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics surveys*, 4:1–39, 2010.

[72] Martin A. Fischler. *Intelligence: The Eye, the Brain, and the Computer*. Addison-Wesley, January 1987.

[73] Matthew Fisher, Daniel Ritchie, Manolis Savva, Thomas Funkhouser, and Pat Hanrahan. Example-based Synthesis of 3D Object Arrangements. *ACM Trans. Graph.*, 31(6):135:1–135:11, November 2012.

[74] Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid. *Computer Vision - ECCV 2012*. 2012.

[75] DA Forsyth and J Ponce. *Computer vision: a modern approach*. 2002.

[76] Cecille Freeman. *Feature selection and hierarchical classifier design with applications to human motion recognition*. PhD thesis, 2014.

[77] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. 2001.

[78] LM Fu. Neural networks in computer intelligence. page 484, April 2003.

[79] Yanwei Fu, Timothy M Hospedales, Tao Xiang, and Shaogang Gong. Learning Multi-modal Latent Attributes. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):303–16, February 2014.

[80] Yanwei Fu, TM Hospedales, Tao Xiang, and Shaogang Gong. Attribute learning for understanding unstructured social activity. *Computer Vision-ECCV 2012*, 2012.

[81] K. Fukushima, S. Miyake, and T. Ito. Neocognitron: A neural network model for a mechanism of visual pattern recognition. *Systems, Man and Cybernetics, ...*, SMC-13(5):826–834, September 1983.

[82] Kunihiko Fukushima. Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. 202, 1980.

[83] Dov Gabbay, John Woods, and Paul Thagard. *Philosophy of Psychology and Cognitive Science: A Volume of the Handbook of the Philosophy of Science Series*. Elsevier, October 2006.

[84] D. Gabor. Communication theory and cybernetics. *Circuit Theory, Transactions of the IRE Professional ...*, CT-1(4):19–31, December 1954.

[85] G. David Garson. *Neural Networks: An Introductory Guide for Social Scientists*. SAGE, September 1998.

[86] James Garson. Connectionism. Winter 201 edition, 2012.

[87] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, November 2014.

[88] Ross Girshick, Forrest Iandola, Trevor Darrell, and Jitendra Malik. Deformable part models are convolutional neural networks. *arXiv preprint arXiv:1409.5403*, September 2014.

[89] AS Glassner. *Principles of digital image synthesis: Vol. 1*. Elsevier, 1995.

[90] JG Glimm, John Impagliazzo, and Isadore Singer. *The legacy of John von Neumann*. American Mathematical Soc., September 2006.

[91] Ian J. Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout Networks. pages 1319–1327, February 2013.

[92] IJ Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, and Vinay Shet. Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks. *arXiv preprint arXiv: ...*, pages 1–13, 2013.

[93] IJ Goodfellow, Dumitru Erhan, PL Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. Challenges in Representation

Learning: A report on three machine learning contests. *Neural Information ...*, pages 1–8, 2013.

[94] IJ Goodfellow, M Mirza, X Da, Aaron Courville, and Yoshua Bengio. An Empirical Investigation of Catastrophic Forgeting in Gradient-Based Neural Networks. *arXiv preprint arXiv: ...*, 2013.

[95] Luc Van Gool, Theo Moons, and Dorin Ungureanu. Affine / photometric invariants for planar intensity patterns. Lecture Notes in Computer Science, pages 642–651. Springer Berlin Heidelberg, January 1996.

[96] A Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber. A Novel Connectionist System for Unconstrained Handwriting Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868, May 2009.

[97] Madan M Gupta and George K Knopf. Neuro-vision systems: A tutorial. In *Neuro-Vision Systems: Principles and Applications*, pages 1–34. 1994.

[98] Fredric M. Ham and Ivica Kostanic. *Principles of Neurocomputing for Science and Engineering*. McGraw-Hill Higher Education, 1st edition, 2000.

[99] Fredric M. Ham and Ivica Kostanic. *Principles of Neurocomputing for Science and Engineering*. McGraw-Hill Higher Education, 1st edition, 2000.

[100] Tele Hao, Tapani Raiko, Alexander Ilin, and Juha Karhunen. Gated boltzmann machine in texture modeling. *... Neural Networks and Machine ...*, 2012.

[101] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous Detection and Segmentation. Lecture Notes in Computer Science, pages 297–312. Springer International Publishing, January 2014.

[102] C. Harris and M. Stephens. A Combined Corner and Edge Detector. *Procedings of the Alvey Vision Conference 1988*, pages 23.1–23.6, 1988.

[103] David Hartley. *Observations on man, his frame, his duty, and his expectations*. 00 L : Scholars' Facsimiles and Reprints, 1749.

[104] S Haykin. *Neural networks: a comprehensive foundation*. 1994.

[105] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *arXiv:1406.4729 [cs]*, June 2014.

[106] Donald Olding Hebb. *The Orgamization of Behavior a Neuropsychological Theory*. John Wiley & Sons Inc., New York, 1949.

[107] Robert Hecht-Nielsen. *Neurocomputing*. Addison-Wesley Publishing Company, January 1989.

[108] G E Hinton, P Dayan, B J Frey, and R M Neal. The "wake-sleep" algorithm for unsupervised neural networks. *Science (New York, N.Y.)*, 268(5214):1158–61, May 1995.

[109] GE Hinton, S Osindero, and YW Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 1554:1527–1554, 2006.

[110] GE Hinton, Simon Osindero, and YW Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 2006.

[111] GE Hinton and RR Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(July):504–507, 2006.

[112] GE Hinton, N Srivastava, Alex Krizhevsky, I Sutskever, and RR Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv: ...*, pages 1–18, 2012.

[113] Geoffrey Hinton. To recognize shapes, first learn to generate images. *Progress in brain research*, 2007.

[114] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, and Brian Kingsbury. Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine*, (November):82–97, 2012.

[115] Judy Hoffman, Sergio Guadarrama, and ES Tzeng. LSDA: Large Scale Detection through Adaptation. *Advances in Neural ...*, July 2014.

[116] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[117] Howard C. Warren. A history of the association psychology. page 355, 1921.

[118] P O Hoyer and A Hyvärinen. Independent component analysis applied to feature extraction from colour and stereo images. *Network (Bristol, England)*, 11(3):191–210, August 2000.

[119] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1-3):489–501, December 2006.

[120] DH Hubel and TN Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, pages 106–154, 1962.

[121] DH Hubel and TN Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, pages 215–243, 1968.

[122] A Hyvärinen and Patrik Hoyer. Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural computation*, 1720:1705–1720, 2000.

[123] a Hyvärinen and E Oja. Independent component analysis: algorithms and applications. *Neural networks : the official journal of the International Neural Network Society*, 13(4-5):411–30, 2000.

[124] Aapo Hyvärinen, Jarmo Hurri, and Patrik O. Hoyer. Natural Image Statistics. 39, 2009.

[125] Herbert Jaeger and Harald Haas. Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication. *Science*, 304(5667):78–80, April 2004.

[126] Kevin Jarrett, Koray Kavukcuoglu, Marc' Aurelio Ranzato, and Yann LeCun. What is the best multi-stage architecture for object recognition? *2009 IEEE 12th International Conference on Computer Vision*, pages 2146–2153, September 2009.

[127] Shuiwang Ji, Ming Yang, and Kai Yu. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–31, January 2013.

[128] Yangqing Jia. Caffe: An open source convolutional architecture for fast feature embeding., 2013.

[129] Yangqing Jia, Evan Shelhamer, and Jeff Donahue. Caffe: Convolutional architecture for fast feature embedding. *Proceedings of the ...*, June 2014.

[130] HE Jones and IM Andolina. Differential feedback modulation of center and surround mechanisms in parvocellular cells in the visual thalamus. *The Journal of ...*, 32(45):15946–15951, November 2012.

[131] Timor Kadir and Michael Brady. Saliency, Scale and Image Description. *International Journal of Computer Vision*, 45(2):83–105, November 2001.

[132] Evangelos Kalogerakis, Siddhartha Chaudhuri, Daphne Koller, and Vladlen Koltun. A Probabilistic Model for Component-based Shape Synthesis. *ACM Trans. Graph.*, 31(4):55:1–55:11, July 2012.

[133] Andrej Karpathy, G Toderici, S Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale Video Classification with Convolutional Neural Networks. *vision.stanford.edu*, 2014.

[134] Koray Kavukcuoglu, Pierre Sermanet, Y-lan Boureau, Yann LeCun, Karol Gregor, and Michaël Mathieu. Learning Convolutional Feature Hierarchies for Visual Recognition. *NIPS*, (1):1–9, 2010.

[135] Yan Ke and R. Sukthankar. PCA-SIFT: a more distinctive representation for local image descriptors. volume 2, pages II–506–II–513 Vol.2, June 2004.

[136] FS Khan and RM Anwer. Coloring Action Recognition in Still Images. *International journal of ...*, pages 1–18, 2013.

[137] HJ Kim, JS Lee, and HS Yang. Human action recognition using a modified convolutional neural network. *Advances in Neural Networks-ISNN 2007*, pages 715–723, January 2007.

[138] JC King. Why color management? *9th Congress of the International Color ...*, 2002.

[139] P. E. King-Smith and D. Carden. Luminance and opponent-color contributions to visual detection and adaptation and to temporal and spatial integration. *Journal of the Optical Society of America*, 66(7):709–717, July 1976.

[140] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, 1983.

[141] Teuvo Kohonen. Correlation matrix memories. *Computers, IEEE Transactions on*, 21(4):353–359, April 1972.

[142] Alex Krizhevsky. Convolutional Deep Belief Networks on CIFAR-10. pages 1–9, 2010.

[143] Alex Krizhevsky, I Sutskever, and GE Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *NIPS*, pages 1–9, 2012.

[144] Nicholas Lange, C. M. Bishop, and B. D. Ripley. Neural Networks for Pattern Recognition. *Journal of the American Statistical Association*, 92(440):1642, December 1997.

[145] Ivan Laptev and M Marszalek. Learning realistic human actions from movies. *Computer Vision and ...*, 2008.

[146] Hugo Larochelle, D Erhan, Aaron Courville, James Bergstra, and Yoshua Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. *Proceedings of the 24th ...*, (2006):8, 2007.

[147] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1265–1278, August 2005.

[148] Quoc V. Le, Will Y. Zou, Serena Y. Yeung, and Andrew Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. *Cvpr 2011*, pages 3361–3368, June 2011.

[149] Quoc V. Le, Will Y. Zou, Serena Y. Yeung, and Andrew Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. *Cvpr 2011*, pages 3361–3368, June 2011.

[150] QV Le, Jiquan Ngiam, Zhenghao Chen, DJ hao Chia, and PW Koh. Tiled convolutional neural networks. *NIPS*, pages 1–9, 2010.

[151] QV Le, MA Ranzato, R Monga, and Matthieu Devin. Building high-level features using large scale unsupervised learning. *arXiv preprint arXiv: ...*, 2011.

[152] Nicolas Le Roux and Yoshua Bengio. Representational power of restricted boltzmann machines and deep belief networks. *Neural computation*, 20(6):1631–49, June 2008.

[153] Y LeCun. Generalization and network design strategies. *Connections in Perspective. North-Holland, ...*, 1989.

[154] Y LeCun and Y Bengio. Convolutional networks for images, speech, and time series. *... handbook of brain theory and neural networks*, pages 1–14, 1995.

[155] Y LeCun, B Boser, JS Denker, D Henderson, RE Howard, W Hubbard, and LD Jackel. Backpropagation applied to handwritten zip code recognition. *Neural ...*, 1989.

[156] Y LeCun and L Bottou. Gradient-based learning applied to document recognition. *Proceedings of the ...*, 1998.

[157] Yann LeCun. Une procédure d'apprentissage pour réseau a seuil asymmetrique (a Learning Scheme for Asymmetric Threshold Networks). In *Proceedings of Cognitiva*, pages 599–604, Paris, France, 1985.

[158] Yann LeCun. Learning Process in an Asymmetric Threshold Network. NATO ASI Series, pages 233–240. Springer Berlin Heidelberg, January 1986.

[159] Yann LeCun, B. Boser, JS Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural ...*, pages 396–404, 1990.

[160] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pages 1–8, 2009.

[161] Honglak Lee, PT Pham, Y Largman, and AY Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. *NIPS*, pages 1–9, 2009.

[162] J Leeuw. Journal of Statistical Software. *Wiley Interdisciplinary Reviews: Computational*, 15(9), 2009.

[163] Min Lin, Qiang Chen, and Shuicheng Yan. Network In Network. page 10, December 2013.

[164] Tony Lindeberg. Feature detection with automatic scale selection. *International journal of computer vision*, 30(2):79–116, November 1998.

[165] Ralph Linsker. Self-organisation in a perceptual network. *Computer*, 21(3):105–117, March 1988.

[166] W. A. Little and Gordon L. Shaw. A statistical theory of short and long term memory. *Behavioral Biology*, 14(2):115–133, June 1975.

[167] Weifeng Liu, José C. Principe, and Simon Haykin. *Kernel Adaptive Filtering: A Comprehensive Introduction*. John Wiley & Sons, September 2011.

[168] John Locke. *An essay concerning human understanding*. Read Books, November 1700.

[169] Jonathan L Long, Ning Zhang, and Trevor Darrell. Do Convnets Learn Correspondence? pages 1601–1609. Curran Associates, Inc., 2014.

[170] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[171] D.G. Lowe. Object recognition from local scale-invariant features. volume 2, pages 1150–1157 vol.2, 1999.

[172] HB Mann and DR Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 18(1):50–60, March 1947.

[173] Vikash Mansinghka, Tejas D Kulkarni, Yura N Perov, and Josh Tenenbaum. Approximate Bayesian Image Interpretation using Generative Probabilistic Graphics Programs. pages 1520–1528. Curran Associates, Inc., 2013.

[174] M Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. *Computer Vision and ...*, (i):2929–2936, 2009.

[175] Jonathan Masci, Ueli Meier, D Ciresan, and J Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. *Artificial Neural Networks ...*, pages 52–59, 2011.

[176] J Matas, O Chum, M Urban, and T Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, September 2004.

[177] WS McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, December 1943.

[178] Kishan Mehrotra, CK Mohan, and Sanjay Ranka. *Elements of artificial neural networks*. 1997.

[179] G Mesnil, Y Dauphin, X Glorot, Salah Rifai, Yoshua Bengio, Ian Goodfellow, Erick Lavoie, Xavier Muller, Guillaume Desjardins, David Warde-Farley, Pascal Vincent, Aaron Courville, and James Bergstra. Unsupervised and Transfer Learning Challenge: a Deep Learning Approach. *... of Machine Learning ...*, 7:1–15, 2012.

[180] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, June 1953.

[181] K. Mikolajczyk, B. Leibe, and B. Schiele. Local features for object class recognition. *Computer Vision, 2005. ...*, 2:1792–1799 Vol. 2, October 2005.

[182] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *Pattern Analysis and Machine ...*, 27(10):1615–1630, October 2005.

[183] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International journal of ...*, 65(1-2):43–72, November 2005.

[184] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *International journal of computer vision*, 60(1):63–86, October 2004.

[185] Marvin Minsky and Seymour Papert. *Perceptrons*. November 1969.

[186] ML Minsky. *Theory of neural-analog reinforcement systems and its application to the brain model problem*. PhD thesis, 1954.

[187] Volodymyr Mnih, Nicolas Heess, and Alex Graves. Recurrent Models of Visual Attention. *Advances in Neural Information ...*, June 2014.

[188] Heinz Mühlenbein. Computational Intelligence: The Legacy of Alan Turing and John von Neumann. *Computational Intelligence*, pages 23–43, 2009.

[189] KP Murphy. *Machine learning: a probabilistic perspective*. 2012.

[190] Nadim Nachar. The Mann-Whitney U: a test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology*, 4(1):13–20, 2008.

[191] V Nair and GE Hinton. Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th International ...*, 2010.

[192] Toru Nakashika, Christophe Garcia, Tetsuya Takiguchi, and Insa De Lyon. Local-feature-map Integration Using Convolutional Neural Networks for Music Genre Classification. *INTERSPEECH*, pages 1–4, 2012.

[193] J. Nathans, TP Piantanida, and RL Eddy. Molecular genetics of inherited variation in human color vision. *Science*, 232(4747):203–210, April 1986.

[194] Radford M. Neal. Connectionist learning of belief networks. *Artificial Intelligence*, 56(1):71–113, July 1992.

[195] P. S. Neelakanta and Dolores DeGroff. *Neural Network Modeling: Statistical Mechanics and Cybernetic Perspectives*. CRC Press, July 1994.

[196] J Neumann and AW Burks. Theory of self-reproducing automata. 1966.

[197] Jiquan Ngiam, Zhenghao Chen, Daniel Chia, Pan Wei Koh, and Andrew Y. Ng. Tiled convolutional neural networks. *Advances in Neural ...*, pages 1–9, 2010.

[198] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. *International Conference on Machine Learning*, 28, 2011.

[199] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. *International Journal of Computer Vision*, 79(3):299–318, March 2008.

[200] Nils J. Nilsson. *Learning machines: foundations of trainable pattern-classifying systems*. McGraw-Hill, 1965.

[201] Feng Ning, Damien Delhomme, Yann LeCun, Fabio Piano, Léon Bottou, and Paolo Emilio Barbano. Toward automatic phenotyping of developing embryos from videos. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 14(9):1360–71, September 2005.

[202] Arthur L. Norberg. *Computers and Commerce: A Study of Technology and Management at Eckert-Mauchly Computer Company, Engineering Research Associates, and Remington Rand, 1946 – 1957*. MIT Press, 2005.

[203] Mohammad Norouzi, Mani Ranjbar, and Greg Mori. Stacks of convolutional restricted Boltzmann machines for shift-invariant feature learning. *Computer Vision and Pattern ...*, pages 2735–2742, 2009.

[204] A.B.J. Novikoff. On convergence proofs on perceptrons. *Proceedings of the Symposium on the Mathematical Theory of Automata, New York*, XII:615–622, 1962.

[205] Sebastian Nowozin and CH Lampert. Structured learning and prediction in computer vision. *... and Trends® in Computer Graphics and Vision*, 6(3–4):185–365, March 2011.

[206] D Oneata, Jakob Verbeek, and C Schmid. Action and event recognition with Fisher vectors on a compact feature set. *IEEE Intenational Conference on Computer Vision (ICCV)*, 2013.

[207] Paul Over, George Awad, Jon Fiscus, and Greg Sanders. TRECVID 2013 - An Introduction to the Goals , Tasks , Data , Evaluation Mechanisms , and Metrics. 2013.

[208] D. Parker. Learning-logic. Technical report, Invention Report S81-64, File 1, Cambridge, MA: Center for Computational Research in Economics and Management Science, MIT., 1982.

[209] Razvan Pascanu and YN Dauphin. On the saddle point problem for non-convex optimization. *arXiv preprint arXiv: ...*, pages 1–11, 2014.

[210] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1997.

[211] EA Perez, VF Mota, LM Maciel, Dhiego Sad, and Marcelo B. Vieira. Combining gradient histograms using orientation tensors for human action recognition. *Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE*, 2012.

[212] KB Petersen and MS Pedersen. The matrix cookbook. *Technical University of Denmark*, pages 1–56, 2008.

[213] S.L. Phung, A. Bouzerdoum, and Sr. D. Chai. Skin segmentation using color pixel classification: analysis and comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(1):148–154, January 2005.

[214] JB Pollack. Connectionism: Past, present, and future. *Artificial Intelligence Review*, pages 1–14, 1989.

[215] Marc'Aurelio Ranzato, Fu Jie Huang, Y-Lan Boureau, and Yann LeCun. Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007.

[216] Kishore K. Reddy and Mubarak Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5):971–981, November 2012.

[217] Erik Reinhard, Wolfgang Heidrich, and Paul Debevec. *High dynamic range imaging: acquisition, display, and image-based lighting*. Morgan Kaufmann, May 2010.

[218] Roberto Rigamonti, Matthew a. Brown, and Vincent Lepetit. Are sparse representations really relevant for image classification? *Cvpr 2011*, pages 1545–1552, June 2011.

[219] B. D. Ripley. Neural Networks and Related Methods for Classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(3):409–456, January 1994.

[220] Brian D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

[221] N. Rochester and J. Holland. Tests on a cell assembly theory of the action of the brain, using a large digital computer. *Information Theory, IRE ...*, 2(3):80–93, September 1956.

[222] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386–408, 1958.

[223] F Rosenblatt. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Laboratory, INC., Buffalo 21, N.Y., March 1961.

[224] Frank Rosenblatt. The Perceptron, a Perceiving and Recognizing Automaton. Technical report, Cornell Aeronautical Laboratory, Buffalo, NY, 1957.

[225] Olga Russakovsky, Jia Deng, and Hao Su. Imagenet large scale visual recognition challenge. *arXiv preprint arXiv: ...*, September 2014.

[226] Bahjat Safadi, Nadia Derbas, Abdelkader Hamadi, Thi-thu-thuy Vuong, Han Dong, Philippe Mulhem, and Georges Qu. Quaero at TRECVid 2013 : Semantic Indexing. 2013.

[227] L. K. Saul, T. Jaakkola, and M. I. Jordan. Mean Field Theory for Sigmoid Belief Networks. *arXiv:cs/9603102*, February 1996.

[228] Konrad Schindler and Luc van Gool. Action snippets: How many frames does human action recognition require? *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[229] Jürgen Schmidhuber. *Deep Learning in Neural Networks: An Overview.* Manno-Lugano, 2014.

[230] C Schuldt, I Laptev, and B Caputo. Recognizing human actions: a local SVM approach. *Pattern Recognition, 2004. ...*, pages 3–7, 2004.

[231] M. Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, November 1997.

[232] Matthias Seeger. *Gaussian processes for machine learning.*, volume 14. May 2004.

[233] TJ Sejnowski and CR Rosenberg. Parallel networks that learn to pronounce English text. *Complex systems*, 1:145–168, 1987.

[234] Pierre Sermanet, David Eigen, X Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. *arXiv preprint arXiv: ...*, pages 1–16, 2014.

[235] T. Serre, L. Wolf, and T. Poggio. Object Recognition with Features Inspired by Visual Cortex. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2:994–1000, 2005.

[236] Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio. Robust object recognition with cortex-like mechanisms. *IEEE transactions on pattern analysis and machine intelligence*, 29(3):411–26, March 2007.

[237] Claude Elwood Shannon and John McCarthy. *Automata Studies: Annals of Mathematics Studies. Number 34*. Princeton University Press, 1972.

[238] Jamie Shotton, Toby Sharp, and Alex Kipman. Real-time human pose recognition in parts from single depth images. *Communications of the ...*, 56(1):116–124, January 2013.

[239] P Simard, Dave Steinkraus, and JC Platt. Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis. *ICDAR*, 2003.

[240] Karen Simonyan, A Vedaldi, and A Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv preprint arXiv:1312.6034*, pages 1–8, 2013.

[241] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Fisher networks for large-scale image classification. *Advances in neural ...*, pages 163–171, 2013.

[242] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information ...*, June 2014.

[243] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, September 2014.

[244] CGM Snoek and KEA van de Sande. MediaMill at TRECVID 2013: Searching Concepts, Objects, Instances and Events in Video. *... of TRECVID*, 2013.

[245] Friedrich T Sommer, Thomas Wennekers, and El Camino Real. Models of distributed associative memory networks in the brain $*$. 122(1949):55–69, 2003.

[246] Jost Tobias Springenberg and Martin Riedmiller. Improving Deep Neural Networks with Probabilistic Maxout Units. December 2013.

[247] Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. *Learning Theory*, pages 545–560, 2005.

[248] N Srivastava and Geoffrey Hinton. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine ...*, 15:1929–1958, 2014.

[249] Marijn F Stollenga, Jonathan Masci, Faustino Gomez, and Jürgen Schmidhuber. Deep Networks with Internal Selective Attention through Feedback Connections. pages 3545–3553. Curran Associates, Inc., 2014.

[250] Yongqing Sun, Kyoko Sudo, Yukinobu Taniguchi, and H Li. TRECVid 2012 Semantic Video Concept Detection by NTT-MD-DUT. *Proc. TRECVID 2012 ...*, 2012.

[251] I Sutskever, James Martens, and Geoffrey Hinton. Generating text with recurrent neural networks. *Proceedings of the ...*, 2011.

[252] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[253] Daniel Svozil, V Kvasnicka, and Jie Pospichal. Introduction to multilayer feed-forward neural networks. *Chemometrics and intelligent ...*, 39:43–62, 1997.

[254] Christian Szegedy, Wei Liu, Yangqing Jia, and Pierre Sermanet. Going deeper with convolutions. *arXiv preprint arXiv: ...*, September 2014.

[255] Christian Szegedy, W Zaremba, and I Sutskever. Intriguing properties of neural networks. *arXiv preprint arXiv: ...*, pages 1–9, 2013.

[256] R Szeliski. *Computer vision: algorithms and applications*. 2010.

[257] Yaniv Taigman, Ming Yang, Marc Aurelio Ranzato, and Lior Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. 2014.

[258] GW Taylor, Rob Fergus, Y LeCun, and Christoph Bregler. Convolutional learning of spatio-temporal features. *Computer Vision-ECCV 2010*, 2010.

[259] Wilfrid K. Taylor. Electrical simulation of some nervous system functional activities. *Information theory 3*, pages 314–328, 1956.

[260] Jonathan Tompson, Ross Goroshin, and Arjun Jain. Efficient Object Localization Using Convolutional Networks. *arXiv preprint arXiv: ...*, November 2014.

[261] Advanced Topics and I N Computational. Comparison of Artificial Neural Networks ; and training an Extreme Learning Machine. (April):1–3, 2013.

[262] L Uhr. Highly parallel, hierarchical, recognition cone perceptual structures. *Parallel computer vision*, 1987.

[263] A. M Uttley. Conditional probability machines and conditional reflexes. *Automata studies*, pages 253–276, 1956.

[264] A. M. Uttley. Temporal and spatial patterns in a conditional probability machine. *Automata Studies*, pages 277– 285, 1956.

[265] VN Vapnik and AY Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, January 1971.

[266] V. Rao Vemuri. Artificial Neural Networks: Concepts and Control Applications. *ieeexplore.ieee.org*, page 509, September 1992.

[267] Chr von der Malsburg. Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, 14(2):85–100, June 1973.

[268] Christoph von der Malsburg and David J Willshaw. A mechanism for producing continuous neural mappings: ocularity dominance stripes and ordered retino-tectal projections. *Exp. Brain Res*, 1:463–469, 1976.

[269] John von Neumann. First Draft of a Report on the EDVAC. Technical Report 1, 1945.

[270] John von Neumann. Probabilistic logics and the synthesis of reliable organisms from unreliable components. *Automata studies*, pages 43–98, 1956.

[271] Winn Voravuthikunchai. Histograms of pattern sets for image classification and object recognition. *IEEE Conference on ...*, pages 224–231, 2014.

[272] Heng Wang, A Klaser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. *... and Pattern Recognition (...*, 2011.

[273] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, and Cordelia Schmid. Evaluation of local spatio-temporal features for action recognition. *Procedings of the British Machine Vision Conference 2009*, pages 124.1–124.11, 2009.

[274] P Werbos. Beyond regression: new tools for prediction and analysis in the behavioral sciences. 1974.

[275] Jason Weston, Frédéric Ratle, and Ronan Collobert. Deep learning via semi-supervised embedding. *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages 1168–1175, 2008.

[276] Bernard Widrow. An Adaptive "ADALINE" Neuron Using Chemical "Memistors". Technical report, Solid-State Electronics Laboratory, Stanford Electronics Laboratories, Stanford University, Stanford, California, 1960.

[277] Norbert Wiener. *Cybernetics or Control and Communication in the Animal and the Machine.* The Massachusetts Institute of Technology, 1948.

[278] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, December 1945.

[279] A. L. Wilkes and N. J. Wade. Bain on neural networks. *Brain and Cognition*, 33(3):295–305, April 1997.

[280] William James. *The principles of psychology.* New York : Henry Holt and company, 1890.

[281] RJ Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, June 1989.

[282] D. J. Willshaw, O. P. Buneman, and H. C. Longuet-Higgins. Non-Holographic Associative Memory. *Nature*, 222(5197):960–962, June 1969.

[283] D. J. Willshaw, O. P. Buneman, and H. C. Longuet-Higgins. Non-Holographic Associative Memory. *Nature*, 222(5197):960–962, June 1969.

[284] Lior Wolf, Tal Hassner, and Itay Maoz. Face Recognition in Unconstrained Videos with Matched Background Similarity. *Cvpr 2011*, pages 529–534, June 2011.

[285] Yang Yang, Guang Shu, and Mubarak Shah. Semi-supervised Learning of Feature Hierarchies for Object Detection in a Video. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1650–1657, June 2013.

[286] Yi-Ting Yeh, Lingfeng Yang, Matthew Watson, Noah D. Goodman, and Pat Hanrahan. Synthesizing Open Worlds with Constraints Using Locally Annealed Reversible Jump MCMC. *ACM Trans. Graph.*, 31(4):56:1–56:11, July 2012.

[287] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? pages 3320–3328. Curran Associates, Inc., 2014.

[288] Thomas Young. The Bakerian lecture: On the theory of light and colours. *Philosophical transactions of the Royal Society of . . .*, 92:12–48, January 1802.

[289] Matthew D. Zeiler, Graham W. Taylor, and Rob Fergus. Adaptive deconvolutional networks for mid and high level feature learning. *2011 International Conference on Computer Vision*, pages 2018–2025, November 2011.

[290] MD Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. *arXiv preprint arXiv:1311.2901*, 2013.

[291] J. Zhang and M. Marszalek. Local features and kernels for classification of texture and object categories: A comprehensive study. *International journal of ...*, 73(2):213–238, June 2007.

[292] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based R-CNNs for fine-grained category detection. *Computer Vision–ECCV 2014*, pages 834–849, January 2014.

[293] M.Z. Zia, M. Stark, B. Schiele, and K. Schindler. Detailed 3D Representations for Object Recognition and Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2608–2623, November 2013.