Convolutional Network Features for Scene Recognition

Markus Koskela and Jorma Laaksonen
Department of Information and Computer Science
Aalto University School of Science
PO Box 15400, FI-00076 AALTO, Finland
firstname.lastname@aalto.fi

ABSTRACT

Convolutional neural networks have recently been used to obtain record-breaking results in many vision benchmarks. In addition, the intermediate layer activations of a trained network when exposed to new data sources have been shown to perform very well as generic image features, even when there are substantial differences between the original training data of the network and the new domain. In this paper, we focus on scene recognition and show that convolutional networks trained on mostly object recognition data can successfully be used for feature extraction in this task as well. We train a total of four networks with different training data and architectures, and show that the proposed method combining multiple scales and multiple features obtains state-of-the-art performance on four standard scene datasets.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis

Keywords

scene recognition; convolutional networks; spatial pyramid; linear classifiers; explicit kernel maps

1. INTRODUCTION

Scene recognition is an important problem in many application areas of image and video processing. A standard approach during the recent years has been to extract several sets of local patch descriptors, encode them into high-dimensional vectors, pool them into an image-level signatures, employ some standard classification algorithm, and possibly use late fusion to combine the results of multiple features. Widely-used encodings have included SIFT bag-of-words with spatial pyramids [11], sparse coding [6, 14], and Fisher vectors [17, 9].

MM'14, November 3–7, 2014, Orlando, Florida, USA.

ACM 978-1-4503-3063-3/14/11 ...$15.00.
http://dx.doi.org/10.1145/2647868.2655024.

Part-based approaches have recently shown high performance in scene recognition [20, 9, 2]. These methods require only weak supervision, that is, no annotations at the parts-level, but only at the class-level. The part-based methods can be considered as extensions of low-level SIFT-type descriptors towards mid-level features that capture more informative visual elements. Another way to see the part-based methods is to view them as extensions of the deformable parts model [5], which has been very succesful in object detection, but where the parts are assumed to have fixed locations, to a more general image classification task.

Another recent development in image classification has been the use of deep convolutional neural networks (CNNs), with excellent results [10, 23, 18, 7]. Still, one drawback with CNNs is that they require huge amounts of training data and delicate tuning of the training parameters. It has, however, been observed that CNNs trained with one visual dataset can function as highly discriminative features even for considerably different data domains and tasks [3, 16, 7].

In this work, CNNs trained with external data and with two different architectures are used as feature extractors in a standard linear-SVM-based multi-feature visual recognition framework for scene recognition. We consider scene recognition to be a particularly interesting task to study using CNNs, as the higher-layer convolutional filters are visually very similar to image parts in part-based methods [23]. We apply our recognition framework to four standard datasets for scene recognition and obtain competitive recognition accuracies, exceeding the current state-of-the-art results, to the best of our knowledge, for all four datasets.

2. RELATED WORK

There has been lots of work done on scene recognition, but we limit the discussion to only the most relevant recent contributions in this section. The Fisher vector encoding can arguably be considered as the current state-of-the-art in local feature based image classification [17, 9]. By measuring the deviation of a sample from a GMM-based generative model in the SIFT descriptor space, one ends up, however, with very high-dimensional image signatures. Comparable results on standard benchmarks have also been obtained using sparse coding [6, 14] and superpixel segmentation [1].

Several part-based methods have recently been proposed as extensions to the methods based on local feature encoding. The part-based methods can be based on training data containing bounding boxes [24] or image-level class annotations [20, 9, 2]. To further increase the classification accu-

racy, the part-based methods can be combined with local feature encodings, e.g. with Fisher vectors [9, 2].

In particular, the very recent ImageNet classification challenges have been dominated by CNN-based methods [10, 23]. The convolutional networks based on the structure of Krizhevsky et al [10] typically contain five convolutional layers, followed by two fully-connected layers of size 4096 neurons, and the output layer. Existing CNNs, often trained with ImageNet data, have also been used for generic image feature extraction. Several tools exist for this purpose, including OverFeat [18], DeCAF [3], and Caffe [8].

The most similar recent publications to our work are [3, 16, 7]. In [3], a CNN trained with ImageNet 2012 data is applied to object recognition, subcategory recognition, and scene recognition. Our work has been inspired by the scene recognition experiments in [3], as we feel that their baseline classification setup does not fully illustrate the potential of CNN features in this task. The publicly available OverFeat CNN [18], trained with ImageNet 2013 data, is used for several image classification and recognition tasks in [16]. In [7], average, max, and VLAD pooling and multi-scale spatial pyramids are proposed for CNN features, and state-of-the-art results are obtained with several datasets. In this work, we focus on scene recognition. We train several CNNs with different input data and apply a multi-feature recognition framework with hard negative mining [13] and linear SVMs with approximative kernel maps [19], which has been previously succesfully used especially with local features. We compare our features to previous methods based on CNNs, to part-based approaches, and to local feature encodings.

3. SCENE RECOGNITION FRAMEWORK

We extracted four different CNN features from the images. The used CNNs were trained on ImageNet 2010 and 2012 training datasets (about 1.3M images each) using Caffe [8], following as closely as possible the network structure parameters of Krizhevsky et al [10] ("201x-distort" in our notation) and Zeiler & Fergus [23] ("201x-aratio"). Due to different software tools used, our networks are not exact reimplementations. For example, we did not use the convolutional layer RMS renormalization of [23]. The input images for the CNNs were resized to 256×256 pixels, by distorting the aspect ratio for 201x-distort and by cropping the center square for 201x-aratio. A random crop of 227×227 pixels (with a 0.5 probability of horizonal flip) was extracted from input images, and the training data was processed for 90 epochs in mini-batches of 256 (201x-distort) or 128 images (201x-aratio). Due to space constraints, we omit the details of the networks here and point interested readers to [10, 23, 8]. On a 6 GB NVidia GTX Titan GPU card, the training lasted about 7 and 13 days per CNN for 201x-distort and 201x-aratio, respectively.

We use the activations of the first fully-connected layers of each network as our features, as they have been observed to provide the best results (our experiments, and [3, 16]), which results in 4096-dimensional feature vectors. The test images are resized similarly as the training data, and then the center crops of 227×227 pixels are forward-passed through the networks to extract the features. Furthermore, we use the "reverse" spatial pyramid pooling proposed in [7] with two scale levels. Our first level corresponds to the full image, and the second level consists of nine regions of 128×128 pixels with a stride of 64 pixels. The CNN activations of the regions are then pooled using average pooling, and the activations of the different scales are concatenated. The resulting spatial pyramid features are therefore 8192-dimensional.

Table 1: Classification accuracies (and standard deviations) for the scenes-15 dataset

Method	Accuracy	
Lazebnik et al (2006) [11]	0.814 (0.005)	
Sun & Ponce (2013) [20]	0.860 (0.008)	
Zheng et al (2012) [24]	0.863	
Paris et al (2012) [14]	0.870 (0.005)	
Bu et al (2013) [1]	0.894 (0.007)	
Gao et al (2010) [6]	0.898 (0.005)	
	full image	spatial pyr.
2010-aratio	0.887 (0.003)	0.915 (0.003)
2010-distort	0.881 (0.006)	0.908 (0.003)
2012-aratio	0.887 (0.005)	0.913 (0.005)
2012-distort	0.884 (0.006)	0.907 (0.005)
fusion	0.907 (0.003)	**0.921** (0.004)

Table 2: Classification accuracies (and standard deviations) for the uiuc-sports dataset

Method	Accuracy	
Li & Fei-Fei (2007) [12]	0.734	
Gao et al (2010) [6]	0.853 (0.005)	
Sun & Ponce (2013) [20]	0.864 (0.009)	
Zheng et al (2012) [24]	0.872	
Paris et al (2012) [14]	0.877 (0.011)	
	full image	spatial pyr.
2010-aratio	0.935 (0.009)	0.941 (0.007)
2010-distort	0.942 (0.010)	0.944 (0.008)
2012-aratio	0.938 (0.008)	0.946 (0.010)
2012-distort	0.938 (0.011)	0.941 (0.006)
fusion	0.947 (0.010)	**0.948** (0.009)

A one-vs-all linear classifier is trained for each scene category in all experiments, and the scene category with the maximum confidence score is used as the result of the multi-class classification. We apply the homogeneous kernel map approximations of the intersection kernel [21], and use the LIBLINEAR [4] library with the L_2-regularized logistic regression solver. In the fusion experiments, the four features are always late-fused using geometric mean.

4. EXPERIMENTS

We present scene recognition results on four widely used datasets. The Fifteen Scene Categories (*scenes-15*) dataset [11] contains 4485 greyscale images assigned to 15 categories, with 200 to 400 images belonging to each category. We use 100 images per class for training and the rest for testing. The UIUC sports (*uiuc-sports*) dataset [12] contains 8 sports event categories with 137 to 250 images per sports event and a total of 1579 images. We use 70 randomly selected images for training and 60 for testing. For both these datasets, we always use all available negative samples.

The MIT indoor database (*indoor-67*) contains 67 indoor scene categories and a total of 15620 images [15]. On average, 80 images of each class are used for training and 20 for testing. The *sun397* scene benchmark contains 397 scene

Table 3: Averages of per-class classification accuracies for the indoor-67 dataset

Method		Accuracy	
Zheng et al (2012) [24]		0.472	
Bu et al (2013) [1]		0.483	
Sun & Ponce (2013) [20]		0.514	
Razavian et al (2014) [16]		0.584	
Juneja et al (2013) [9]	(BoP)	0.461	
	(FV)	0.608	
	(BoP+FV)	0.631	
Doersch el al (2013) [2]	(MLE)	0.640	
	(MLE+FV)	0.669	
Gong et al (2014) [7]	(full image)	0.537	
	(avg pool)	0.656	
	(VLAD)	0.689	
		full	s. pyr.
2010-aratio		0.624	0.674
2010-distort		0.601	0.657
2012-aratio		0.627	0.669
2012-distort		0.617	0.657
fusion		0.689	**0.701**

Table 4: Classification accuracies (and standard deviations) for the sun397 dataset

Method		Accuracy	
Xiao et al (2010) [22]		0.380	
Donahue et al (2014) [3]		0.409 (0.003)	
Sánchez et al (SIFT)		0.433 (0.002)	
(2013) [17] (SIFT+LCS)		0.472 (0.002)	
Gong et al	(full image)	0.396	
(2014) [7]	(avg pool)	0.475	
	(VLAD)	0.520	
		full image	spatial pyr.
2010-aratio		0.456 (0.002)	0.503 (0.001)
2010-distort		0.456 (0.004)	0.504 (0.002)
2012-aratio		0.460 (0.002)	0.506 (0.002)
2012-distort		0.454 (0.003)	0.506 (0.003)
fusion		0.519 (0.003)	**0.547** (0.002)

categories and a total 108 756 images [22]. 50 images per category are used for training and for testing. For these datasets, we include two rounds of hard negative mining [13] and sample 1000 negative examples on each round.

For the first two datasets, we use 10 random splits into training and test sets, with *indoor-67* we use the partitions of [15], and with *sun397* we use the same 10 partitions as in [22]. For *indoor-67*, we use average of per-class accuracies (mean of diagonal values of the confusion matrix) as the performance metric. For other datasets, the mean and standard deviation of average multiclass accuracies are calculated.

The results of the experiments are shown in Tables 1–4. For *scenes-15* and *uiuc-sports*, the confusion matrices of the runs achieving the highest accuracies are shown in Figs. 1–2. For *indoor-67* and *sun397*, the detailed results of our experiments can be found at *http://research.ics.aalto.fi/cbir/*.

For all datasets, our results are reported with the four used CNNs and with the full image and spatial pyramid features. In the literature, full image CNN feature results have been reported for *indoor-67* and *sun397* [16, 3, 7]. Our best single features outperform the previously reported results: 0.627 vs. 0.584 [16] for *indoor-67* and 0.460 vs. 0.409 [3] for *sun397*. Using the two-scale pyramid, which doubles the feature dimensionality and requires one to forward-pass several (in our case, ten) image regions through the CNN, further brings a notable performance increase. Overall, the four used CNNs achieve rather similar results, without any clear differences based on either the used architecture or data. Feature fusion systematically improves the results, although the improvement is not as large with the spatial pyramid features. Remarkably, our results with spatial pyramids and feature fusion considerably improve on the best reported results, according to our knowledge, on all datasets.

The SIFT-like local feature encodings can in some sense be considered to form a baseline for any higher-level approaches. On *scenes-15* and *uiuc-sports*, the highest published accuracies, known to us, have been obtained with sparse coding [6, 14]. On *indoor-67* and *sun397*, the best reported accuracies with local features, 0.608 for *indoor-67* [9], and 0.433 for *sun397* (single feature) [17], have been obtained with Fisher vectors. However, these methods result in signature dimensionalities in the order of $O(10^5)$.

The part-based approaches, such as bag-of-parts [9] or [20, 24], have been shown to produce promising results, but often not quite reaching the top-performing local features, except for the mid-level elements [2] in *indoor-67*. Part-based methods can, however, be easily applied as complementary features to local feature encodings to improve the overall recognition accuracy [9, 2] (see Table 3).

The CNN activation features have recently been used to obtain the best known results on the *indoor-67* and *sun397* datasets. The accuracy of the activation features as such (our full image results and [16, 3, 7]) is rather close to the Fisher vectors (whose dimensionality can be almost two orders of magnitude higher). By using multiple scales and concatenating the activations, one can, however, further improve the CNN features. One possibility is to use average pooling, as in our spatial pyramid results and in [7]. It should be noted that we use two scales whereas [7] has three scale levels and a stride of 32, which requires considerably more forward-passes of image regions on a single CNN than our method even with the four CNNs used here. Using a more sophisticated pooling strategy, as the VLAD pooling, including pre- and post-PCA steps to reduce the dimensionality, proposed also in [7], can also improve the results.

5. CONCLUSIONS

The CNN activation features have a great promise as universal representation for various image classification and retrieval tasks. Compared to many existing image signatures, CNN activations are fast to extract, even when applying some kind of a spatial pyramid structure, as the pipeline consist only of image transformations, forward-passes through the CNN, and possible pooling. At the same time, the feature dimensionalities remain modest, especially when compared to e.g. the Fisher vector encoding which in practice requires a separate quantization scheme.

In recent works, part and object based approaches have emerged as a way to construct higher-order models than with SIFT-like local feature encodings. CNN activation features extracted on multiple scales, with the current state-

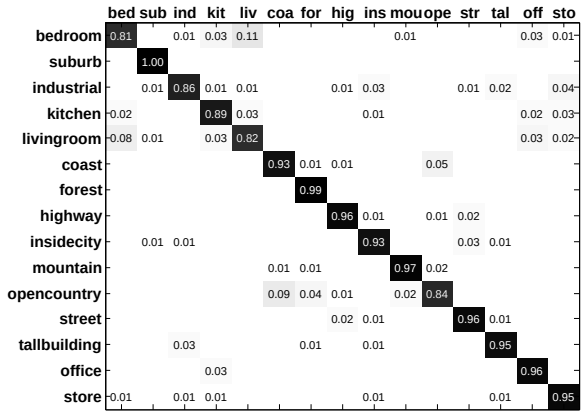

Figure 1: Confusion matrix for the spatial pyramid feature fusion experiment on the scenes-15 dataset

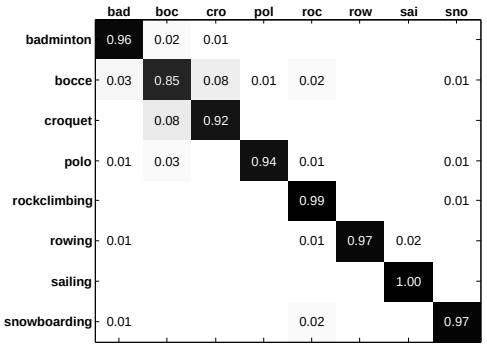

Figure 2: Confusion matrix for the spatial pyramid feature fusion experiment on the uiuc-sports dataset

of-the-art performance reported on several datasets, can be considered as an alternative approach for this purpose.

On the other hand, combining several CNNs can also result in a performance gain as was demonstrated in the experiments of this paper. This has also been observed e.g. in the ImageNet challenges, even when using several networks of identical architecture and trained with the same input data [10, 23]. Even better results can be expected if the variety of the used CNNs would be larger.

6. ACKNOWLEDGMENTS

This work has been funded by the grants 255745 and 251170 of the Academy of Finland, SSP-14183 of EIT ICT Labs, and the D2I SHOK project. The calculations were performed using computer resources within the Aalto University School of Science "Science-IT" project.

7. REFERENCES

[1] S. Bu, Z. Liu, J. Han, and J. Wu. Superpixel segmentation based structural scene recognition. In *ACM Multimedia*, MM '13, 2013.

[2] C. Doersch, A. Gupta, and A. A. Efros. Mid-level visual element discovery as discriminative mode seeking. In *NIPS*, 2013.

[3] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014.

[4] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 9:1871–1874, 2008.

[5] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9), 2010.

[6] S. Gao, I. Tsang, L.-T. Chia, and P. Zhao. Local features are not lonely – Laplacian sparse coding for image classification. In *CVPR*, 2010.

[7] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. arXiv.org:1403.1840, March 2014.

[8] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. `http://caffe.berkeleyvision.org/`, 2013.

[9] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *CVPR*, 2013.

[10] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.

[11] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[12] L.-J. Li and L. Fei-Fei. What, where and who? Classifying events by scene and object recognition. In *ICCV*, 2007.

[13] X. Li, C. Snoek, M. Worring, D. Koelma, and A. Smeulders. Bootstrapping visual categorization with relevant negatives. In *IEEE-TMM*, 15(4), 2013.

[14] S. Paris, X. Halkias, and H. Glotin. Sparse coding for histograms of local binary patterns applied for image categorization: Toward a bag-of-scenes analysis. In *ICPR*, 2012.

[15] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009.

[16] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. arXiv:1403.6382, March 2014.

[17] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the Fisher Vector: Theory and practice. *IJCV*, 105(3):222–245, 2013.

[18] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. OverFeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR*, April 2014.

[19] M. Sjöberg, M. Koskela, S. Ishikawa, and J. Laaksonen. Large-scale visual concept detection with explicit kernel maps and power mean SVM. In *ICMR*, April 2013.

[20] J. Sun and J. Ponce. Learning discriminative part detectors for image classification and cosegmentation. In *ICCV*, 2013.

[21] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *CVPR*, 2010.

[22] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[23] M. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. arXiv:1311.2901, November 2013.

[24] Y. Zheng, Y.-G. Jiang, and X. Xue. Learning hybrid part filters for scene recognition. In *ECCV*, 2012.