Iterative Neural Autoregressive Distribution Estimator (NADE-k)

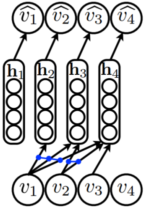
Tapani Raiko, Li Yao, KyungHyun Cho, and Yoshua Bengio

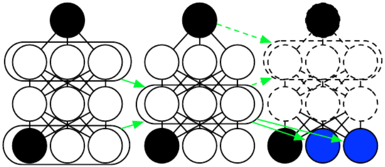
Aalto University, Université de Montréal

Submitted to NIPS 2014

We put two ideas together

- Neural Autoregressive Distribution Estimator (NADE)
- Multi-Predictive Deep Boltzmann Machine (MPDBM)





Neural Autoregressive Distribution Estimator (NADE)

- Learns an analytical $p(\mathbf{x})$, state-of-the-art
- Predicts components of $\mathbf{x}$ sequentially, given the ones before
- Trained by back-prop
- Larochelle & Murray (AISTATS 2011)
- Order-agnostic and deep version by Uria et al. (ICML 2014)

Multi-Predictive Deep Boltzmann Machine (MPDBM)

- Train a DBM with back-prop through inference procedure
- Does not require layerwise pretraining
- Outperforms standard DBM
- Goodfellow et al. (NIPS 2013)

Proposed Method

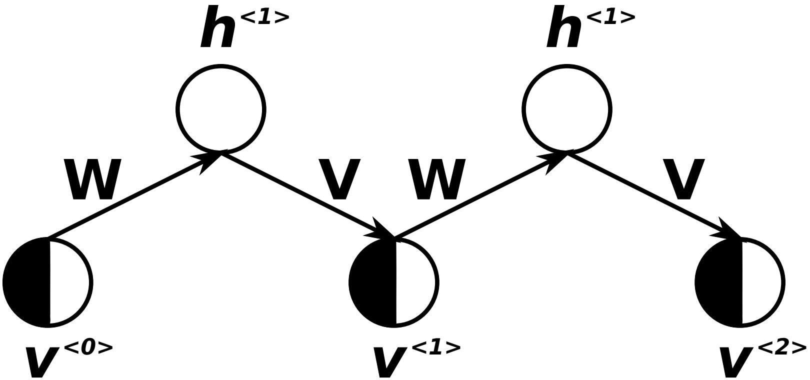
- Learn to impute missing values

$$p_\theta(\mathbf{x}_{\text{mis}} \mid \mathbf{x}_{\text{obs}}) = \prod_{i \in \text{mis}} p_\theta(x_i \mid \mathbf{x}_{\text{obs}})$$

to maximize (averaged) log-likelihood

$$\mathcal{L}(\theta) = \mathbb{E}_{o \in D!} \mathbb{E}_{\mathbf{x} \in \text{data}} - \log \prod_{d=1}^{D} p_\theta(x_{o_d} \mid \mathbf{x}_{o_{<d}})$$

Recurrent NN for Imputation



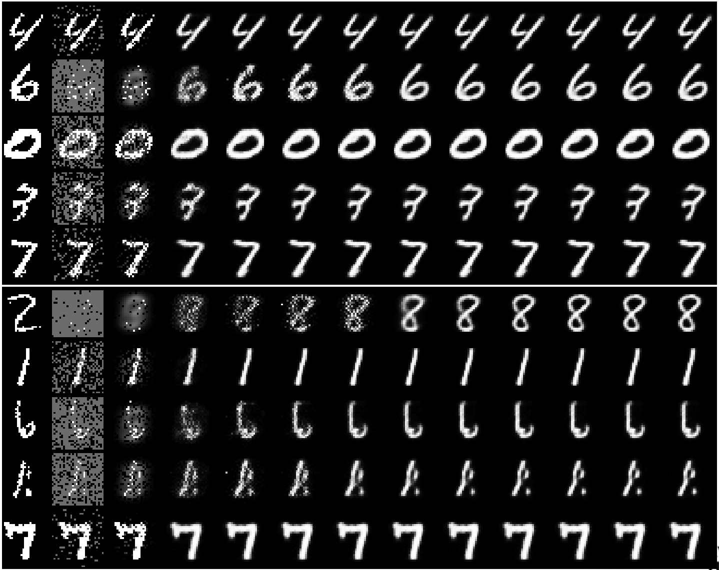
Recurrent NN for Imputation

- Input $\mathbf{v}^{\langle 0 \rangle} = (\mathbf{1} - \mathbf{m}) \odot \mathbf{x}$
  $\mathbf{m}$ is a binary mask indicating missing values
- Do imputation iteratively ($t = 1 \dots k$)

$$\mathbf{h}^{\langle t \rangle} = \phi(\mathbf{W}\mathbf{v}^{\langle t-1 \rangle} + \mathbf{c})$$
$$\mathbf{v}^{\langle t \rangle} = \mathbf{m} \odot \sigma(\mathbf{V}\mathbf{h}^{\langle t \rangle} + \mathbf{b}) + (\mathbf{1} - \mathbf{m}) \odot \mathbf{x}$$

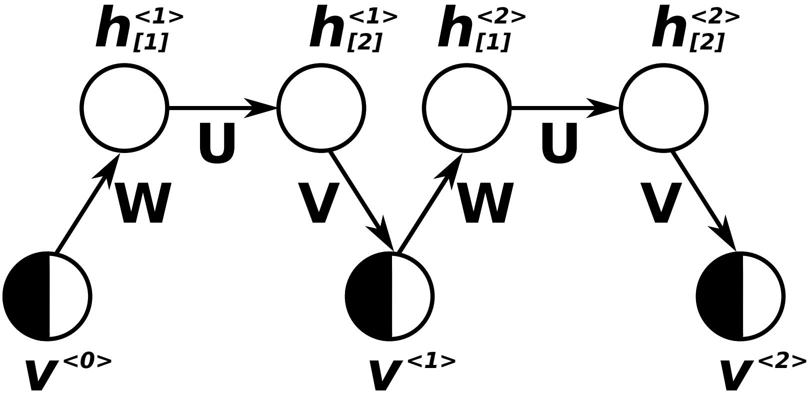
- Output $p_\theta(x_i = 1 \mid \mathbf{x}_{\text{obs}}) = v_i^{\langle k \rangle}$

Input $\mathbf{x}$, masked input, and $\mathbf{v}^{\langle 0 \rangle} \ldots \mathbf{v}^{\langle 10 \rangle}$

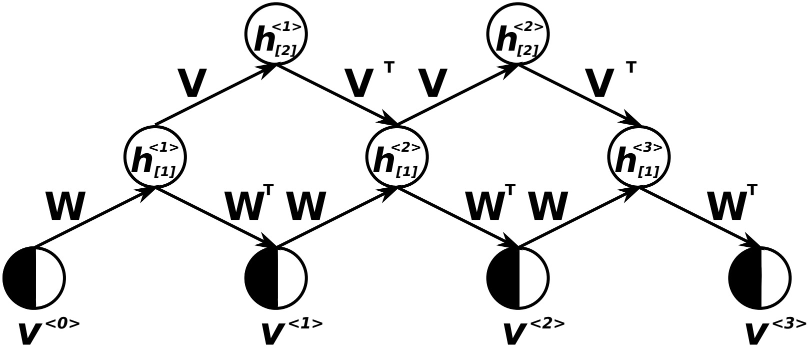


Aalto University

Depth as in NADE



Depth as in MPDBM



Properties

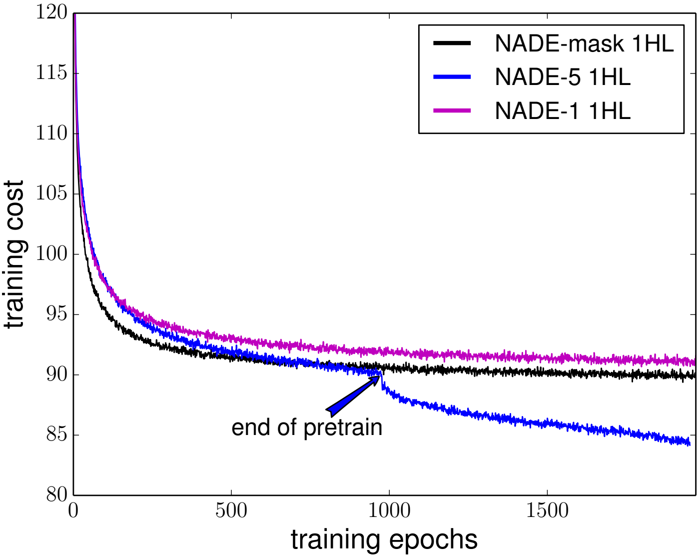
- Parallel training with back-prop through inference
- Tracktable likelihood (sequential)
- Ancestral sampling (iid, no MCMC)
- Generalizes variational mean-field in RBM/DBM
- Flexible deep structures

Details

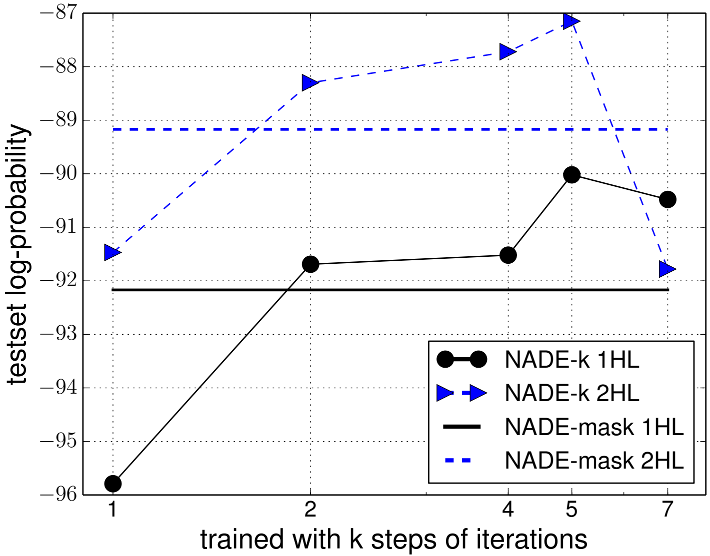
- Uria et al. (2014) gave the mask $\mathbf{m}$ as extra input
- Instead, we simply initialize missing values to the mean

- As pretraining, we aim at good reconstructions $\mathbf{v}^{\langle t \rangle}$ at each step $t = 1 \dots k$

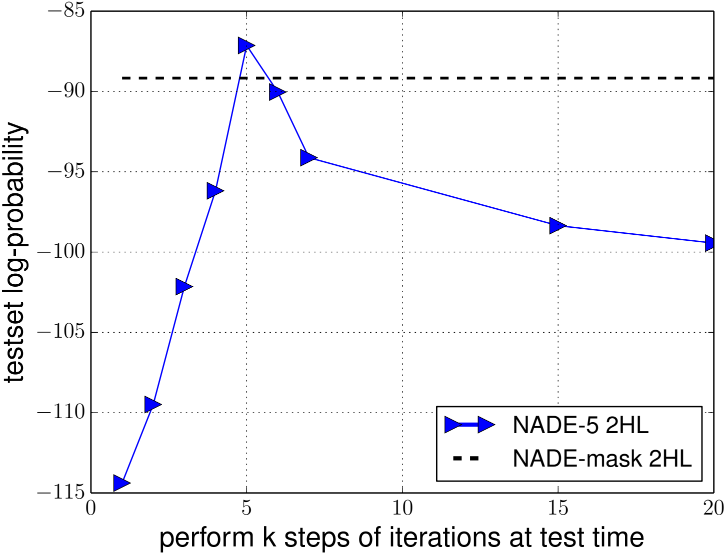
Learning Curves



Varying k and depth in training

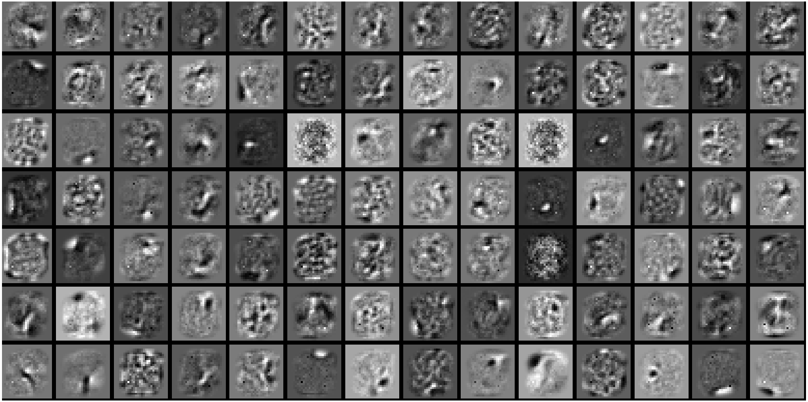


Vary k in testing (trained with $k = 5$)

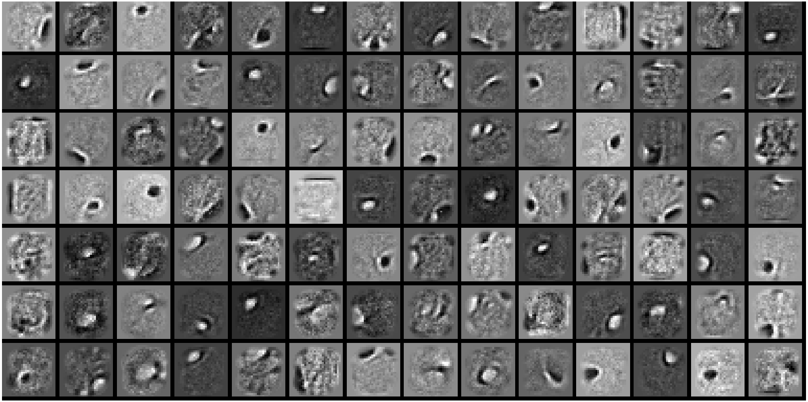


Generated samples (no MCMC!)

Encoding filters

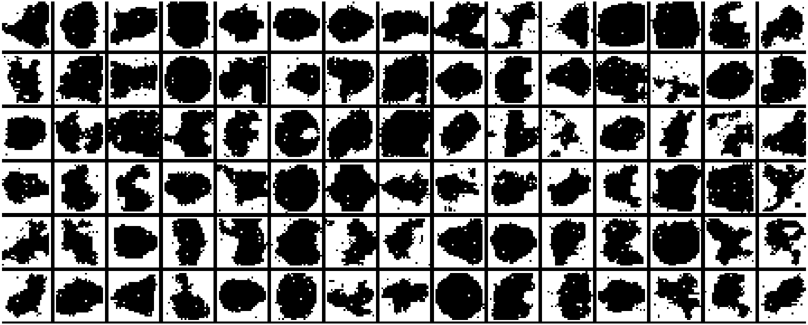


Decoding filters



	Test Log-Prob.
NADE (fixed order)	-88.86
RBM (500h, CD-25)	$\approx$ -86.34
DBN (500h+2000h)	$\approx$ **-84.55**
NADE-mask 1HL	-92.17
NADE-mask 2HL	-89.17
EoNADE-mask 1HL	-87.71
EoNADE-mask 2HL	-85.10
NADE-5 1HL	-90.02
NADE-5 2HL	-87.14
EoNADE-5 1HL	-86.23
EoNADE-5 2HL	**-84.68**

Caltech-101 Silhouettes



- Test LL -107.28 (state-of-the-art)

Conclusions and Discussion

- NADE-k retains the tractability of NADE
- Performs on par with intractable methods (DBN/DBM)
- Li Yao and Antti Rasmus are working on extensions
  - Real-valued data
  - Adjust confidence based on number of observed values