Making Restricted Boltzmann Machines Work

Tapani Raiko

Aalto University School of Science,
Department of Information and Computer Science

Oct 11, 2011

Background (1/2)

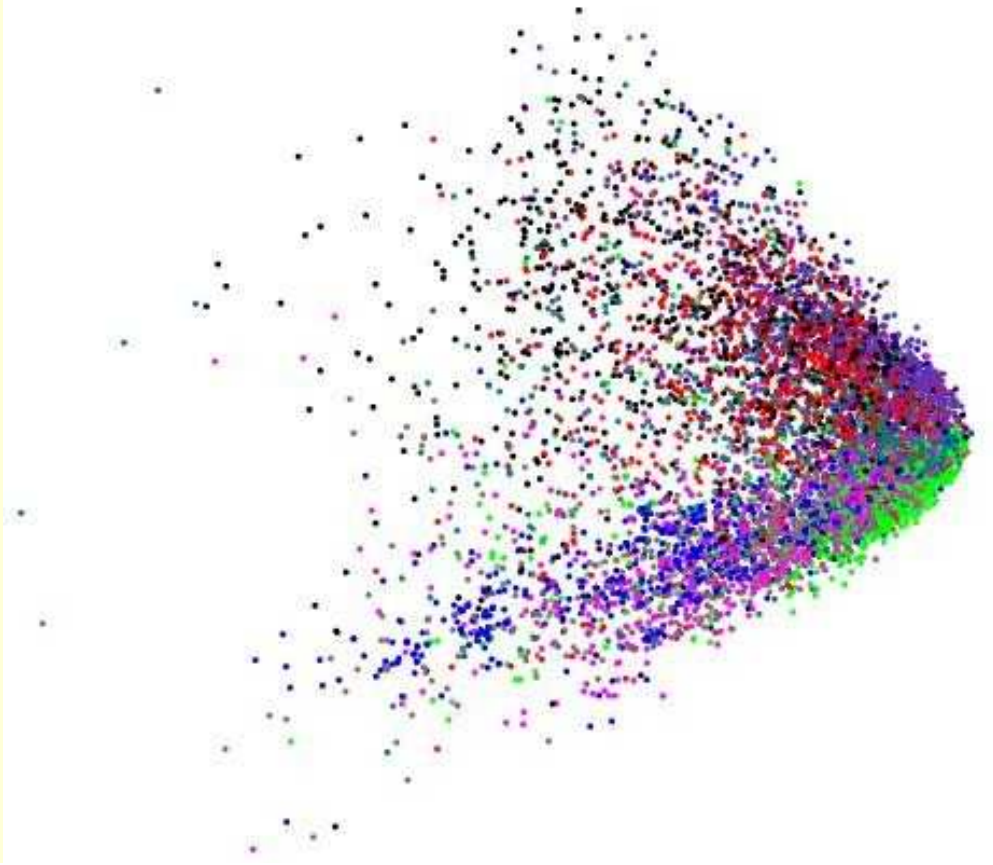
- Long tradition of studying unsupervised learning (SOM, ICA)
- Group: Bayesian algorithms for latent variable models
  Prof. Juha Karhunen, Prof. Erkki Oja,
  DSc Tapani Raiko, DSc Alexander Ilin,
  MSc KyungHyun Cho, MSc Jaakko Luttinen...
- Our focus is on probabilistic latent variable models (NFA, Valpola, 2000) and hierarchical representations (HNFA, Raiko, 2001)

Background (2/2)

- NFA and HNFA are directed graphical models
- Recently it was shown that undirected models can yield better representations (Hinton, 2006)
- Since 2006, learning hierarchical representations is known as deep learning and it is a hot topic

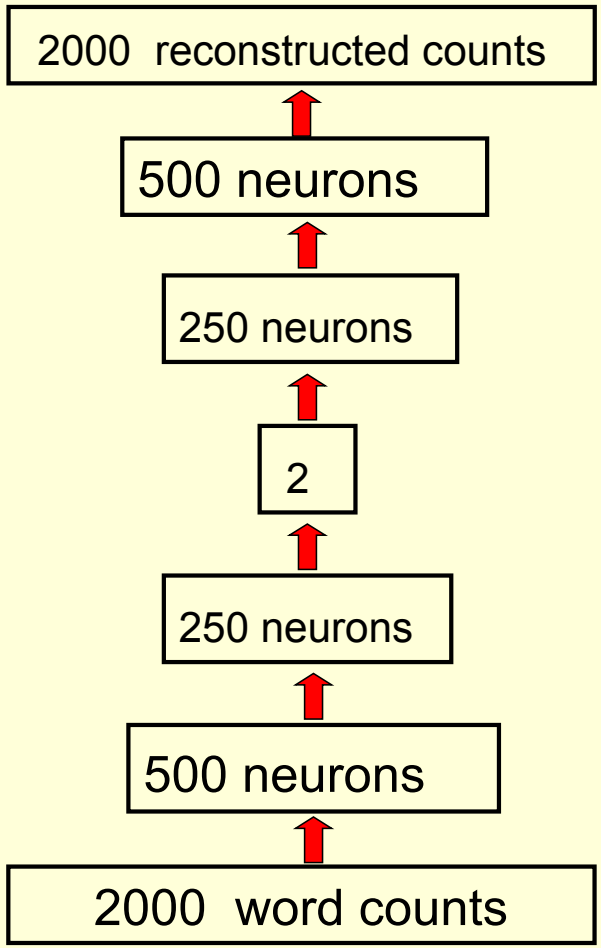
Analysing Documents by Word Counts (Hinton 1/5)

First compress all documents to 2 numbers using a type of PCA
Then use different colors for different document categories



Autoencoder (Hinton 2/5)

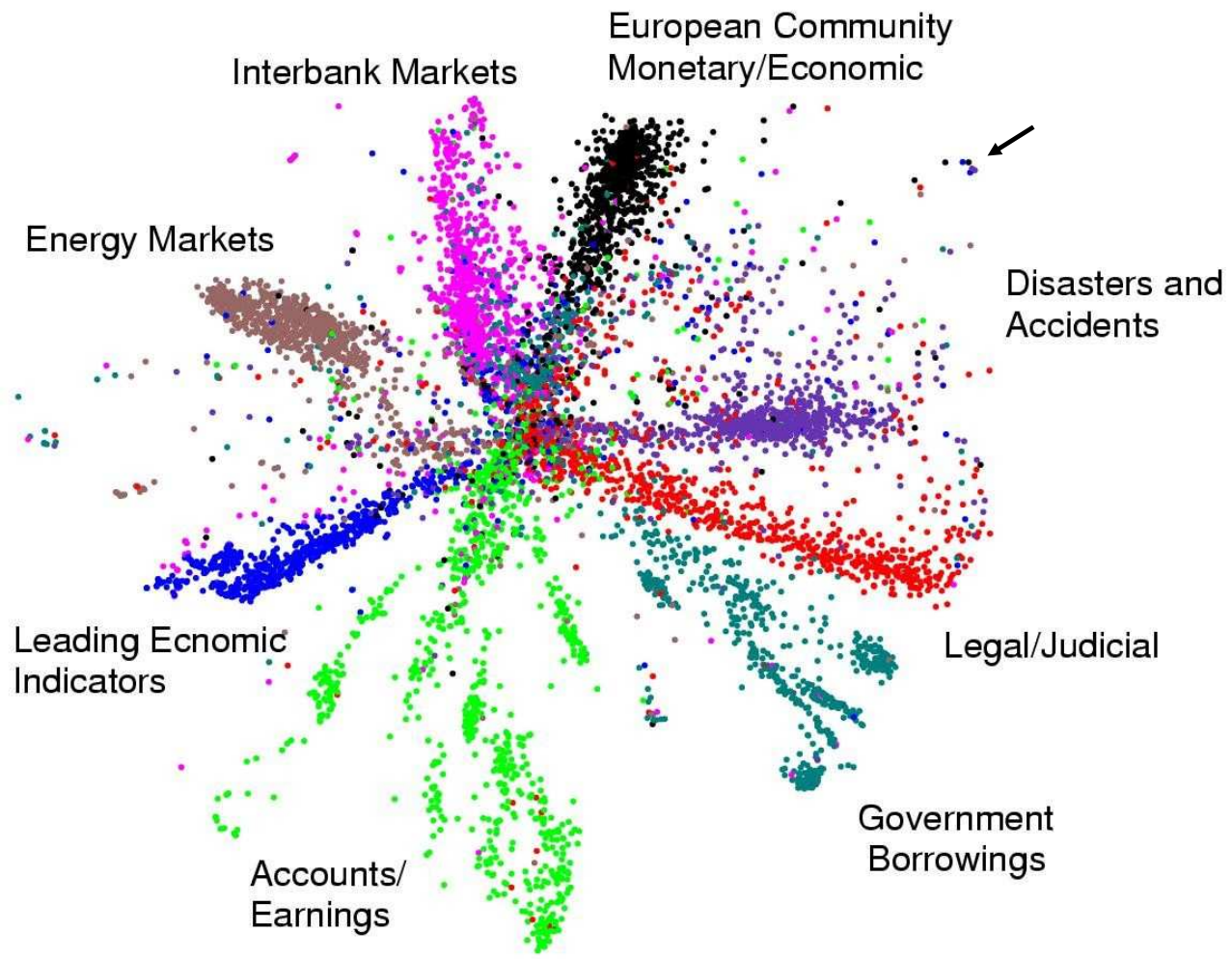
How to compress document count vectors



- We train the autoencoder to reproduce its input vector as its output
- This forces it to compress as much information as possible into the 2 real numbers in the central bottleneck.
- These 2 numbers are then a good way to visualize documents.

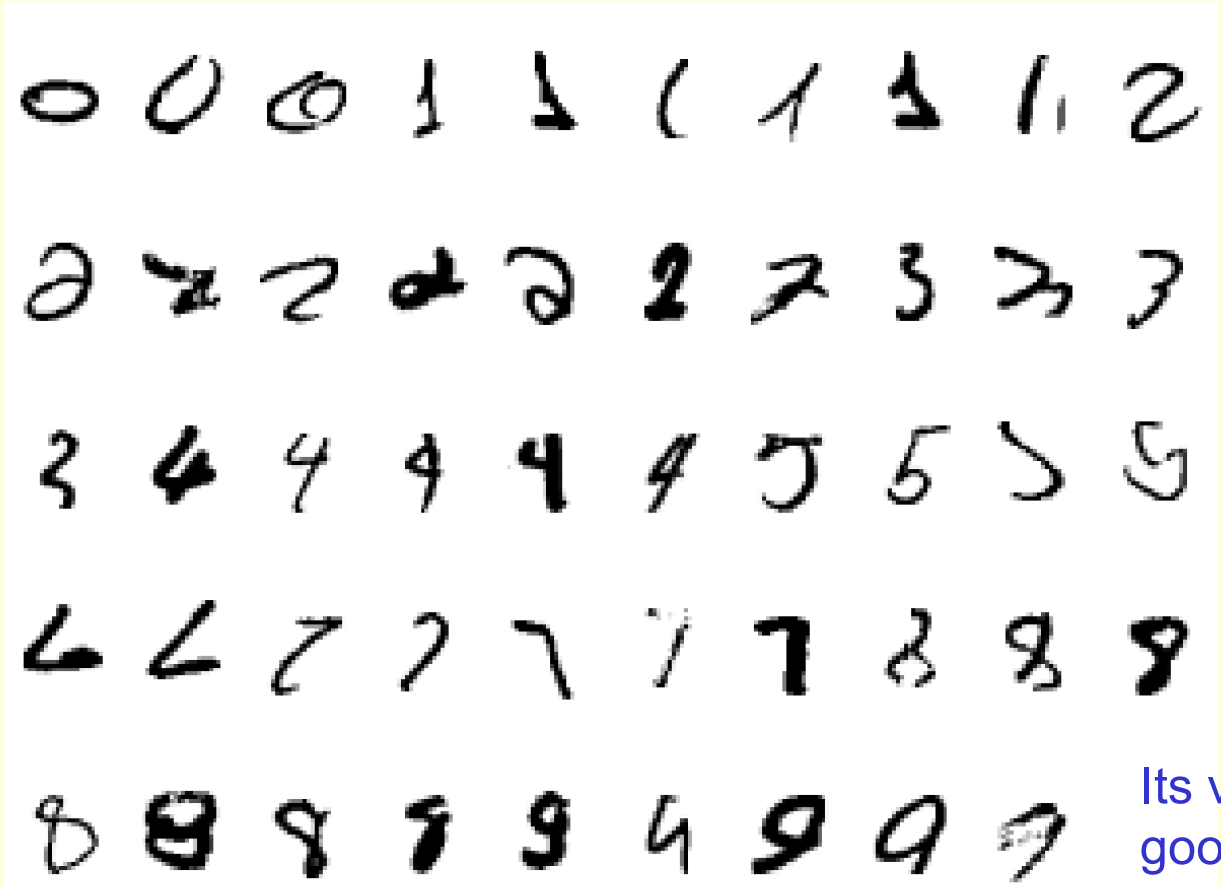
Autoencoding Documents (Hinton 3/5)

First compress all documents to 2 numbers.
Then use different colors for different document categories



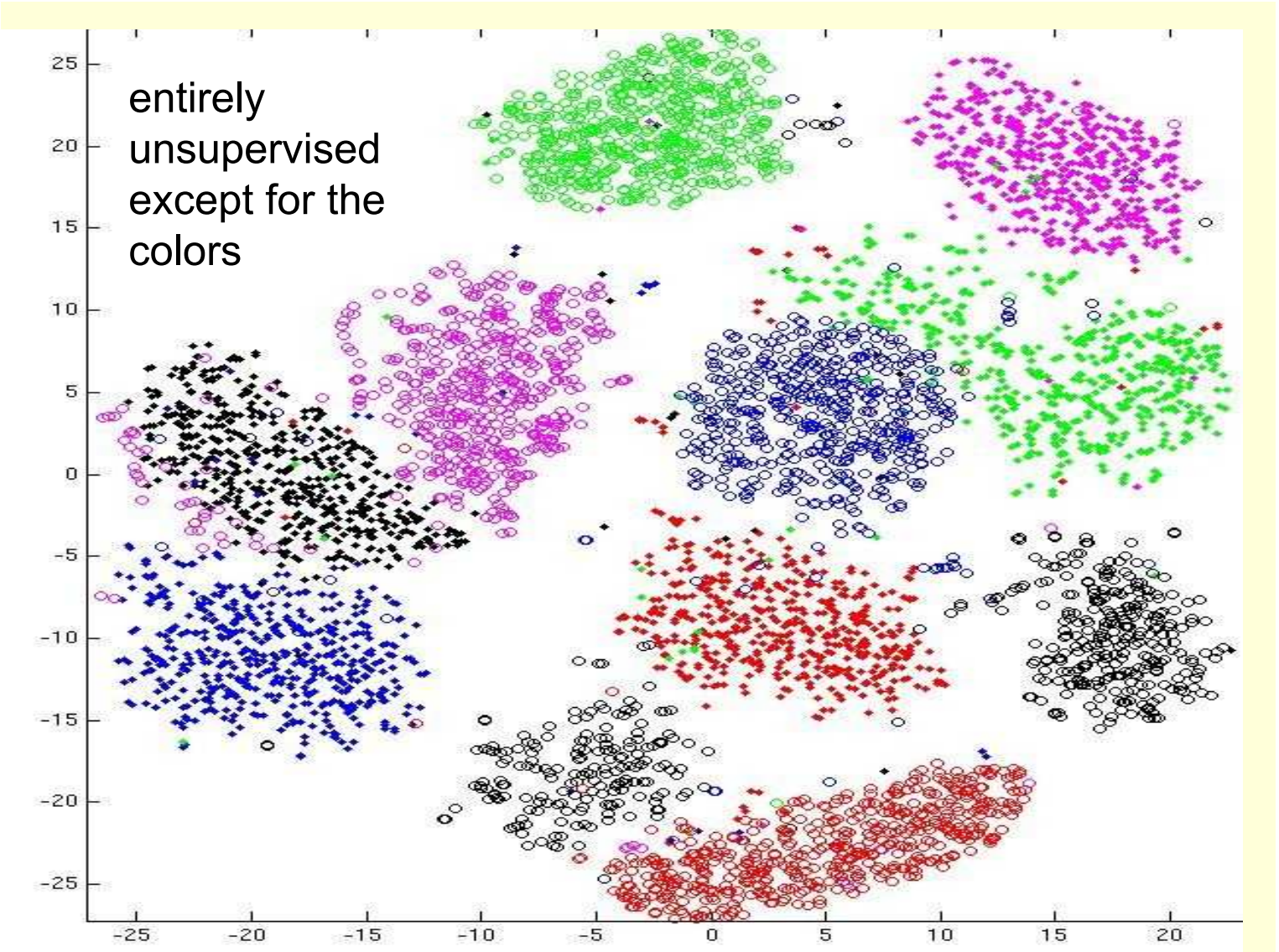
Handwritten Digits (Hinton 4/5)

Examples of correctly recognized handwritten digits that the neural network had never seen before



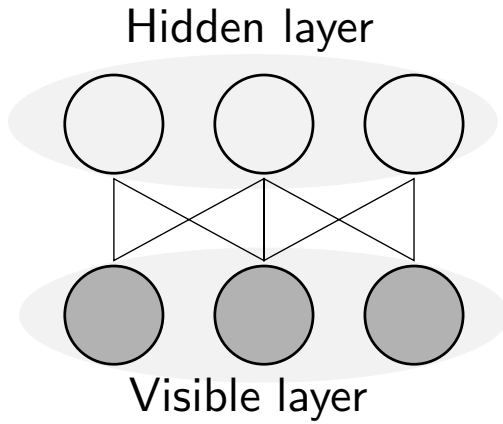
Its very good

Autoencoding Digits (Hinton 5/5)



Restricted Boltzmann Machine (RBM)

- Building block of deep belief networks
- Stochastic undirected neural network (Hinton & Sejnowski 1980s)
- Binary units in visible and hidden layers
- Can model any distribution



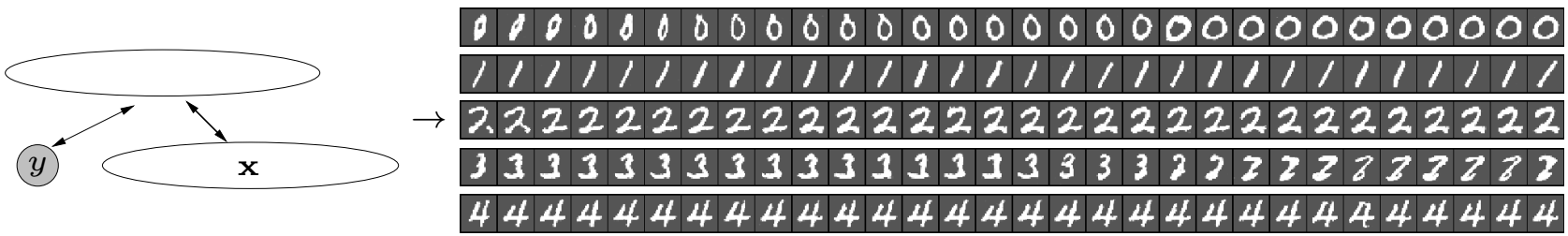
$$P(\mathbf{v}, \mathbf{h} \mid \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\mathbf{v}^T \mathbf{W} \mathbf{h} + \mathbf{b}^T \mathbf{v} + \mathbf{c}^T \mathbf{h}\right)$$

$$\Rightarrow P(v_i = 1 \mid \mathbf{h}, \boldsymbol{\theta}) = \text{sigmoid}\left(\sum_j W_{ij} h_j + b_i\right)$$

$$P(h_j = 1 \mid \mathbf{v}, \boldsymbol{\theta}) = \text{sigmoid}\left(\sum_i W_{ij} v_i + c_j\right)$$

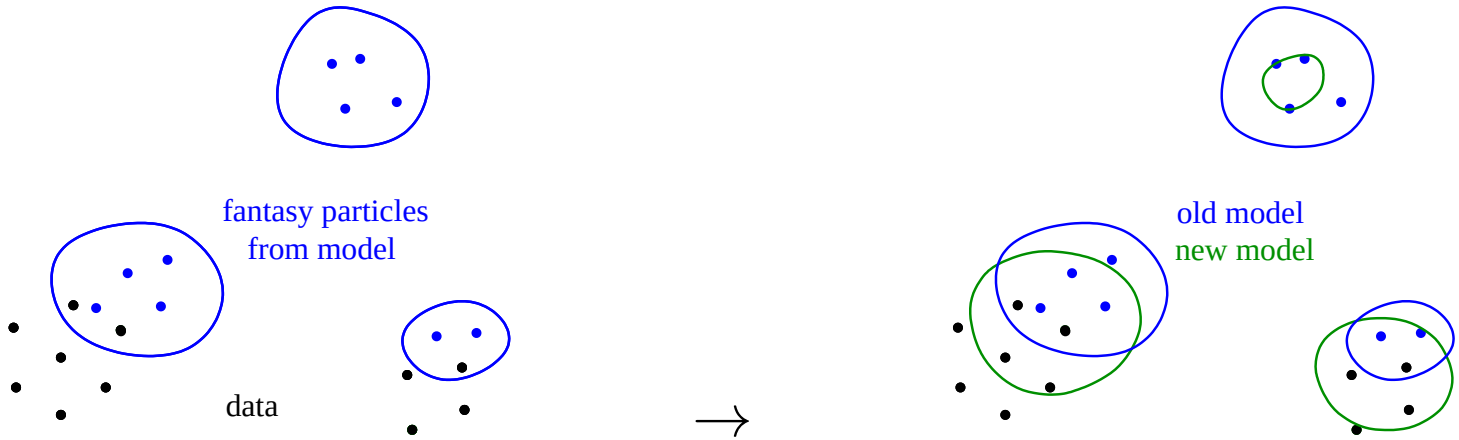
RBM Example

- Visible layer $\mathbf{v}$ contains handwritten digits $\mathbf{x}$ and their labels y
- Generated samples $p(\mathbf{x} \mid y)$ from the RBM:



- Classification accuracy based on $p(y \mid \mathbf{x})$: 97.06%

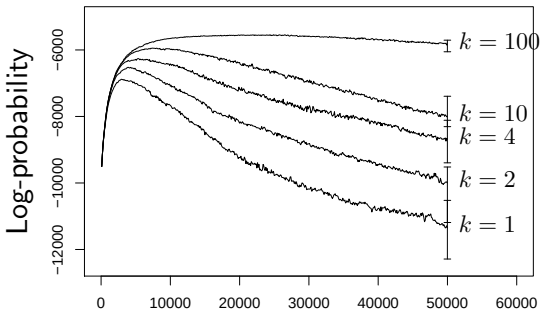
Training RBMs



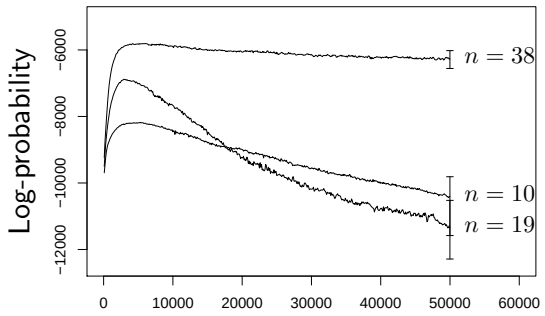
- Stochastic gradient-based maximum likelihood learning:
  Gradient = statistics of data - statistics of model samples
- Likelihood exponentially hard to compute
  ($\rightarrow$ Difficult to evaluate the goodness)

Difficulties in Training

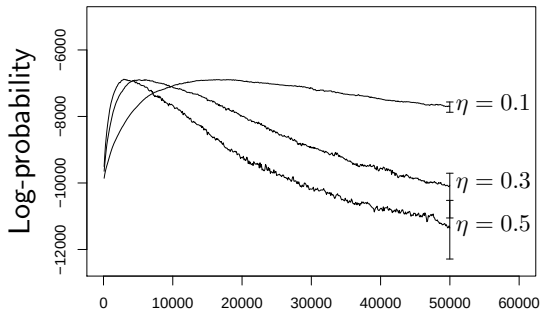
- Despite many success stories of deep networks, training even an RBM is rather difficult (Fischer & Igel, 2010)



CDk

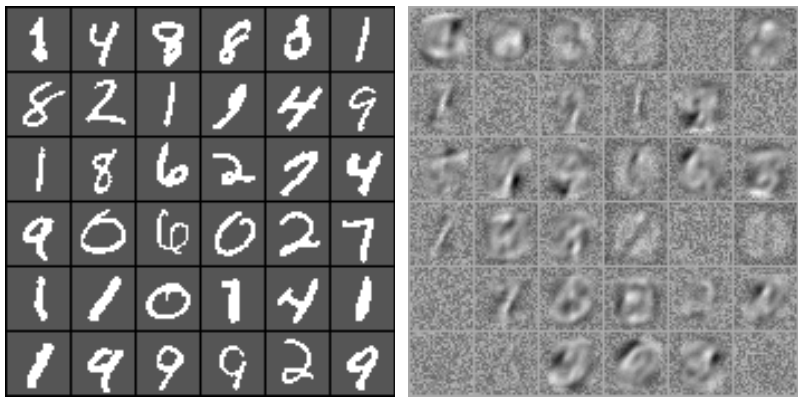


No. of hidden neurons

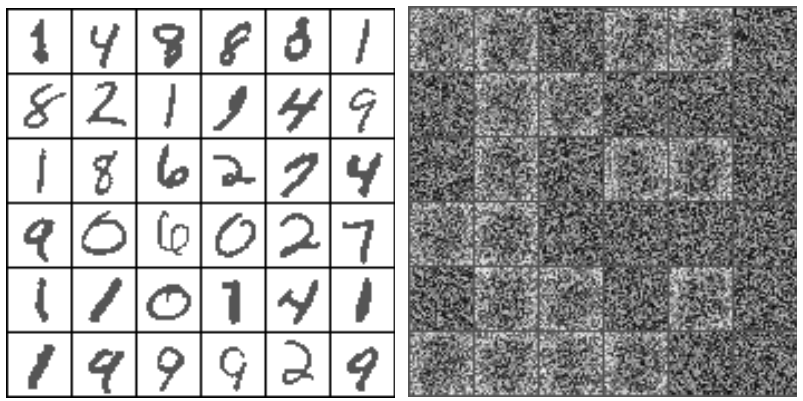


Learning rate

- Same data, but flipping zeroes and ones: learning fails!



MNIST

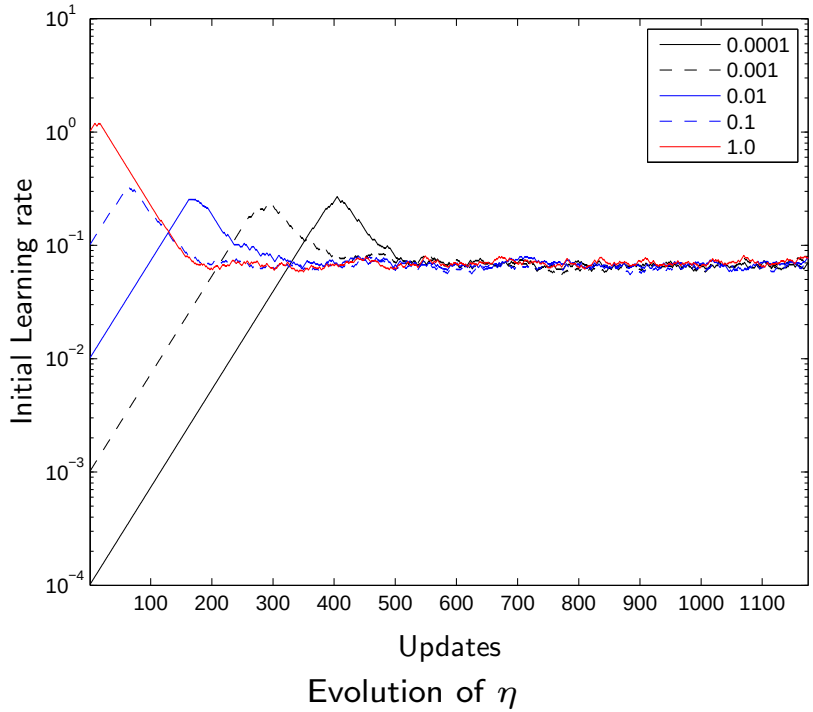


1-MNIST

Improved Learning Algorithm (Cho, Raiko & Ilin, ICML 2011)

- Adaptive Learning Rate
- Enhanced Gradient

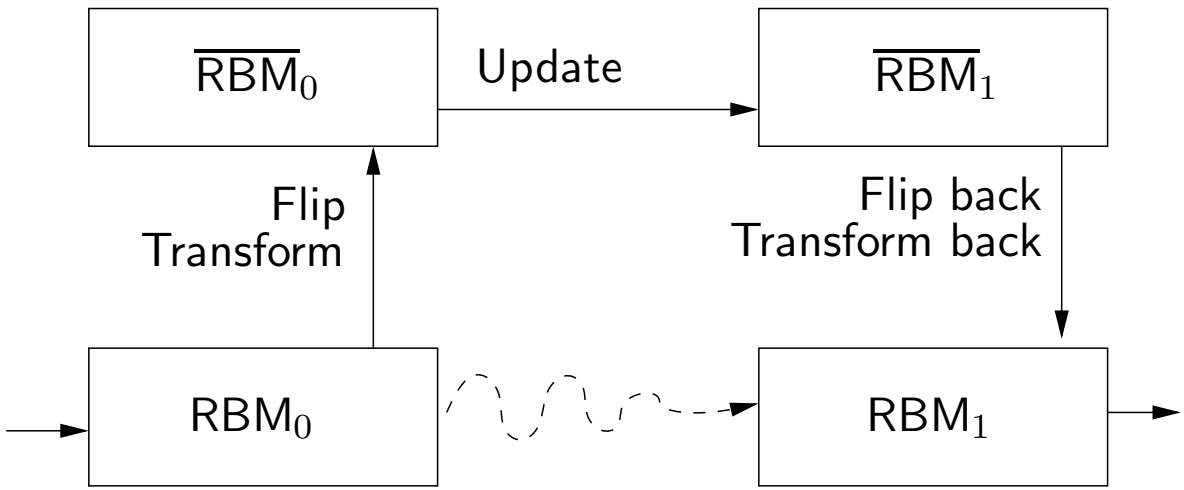
Adaptive Learning Rate



- Approximately compare increasing and decreasing learning rate
- Likelihood ratios can be estimated from samples

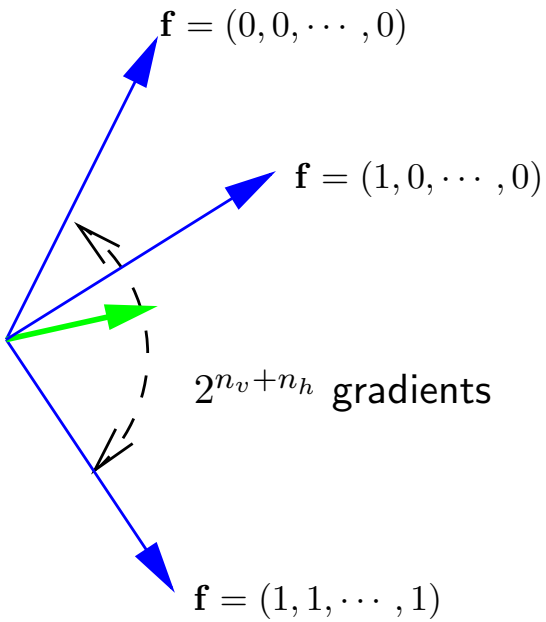
Enhanced Gradient (1/2)

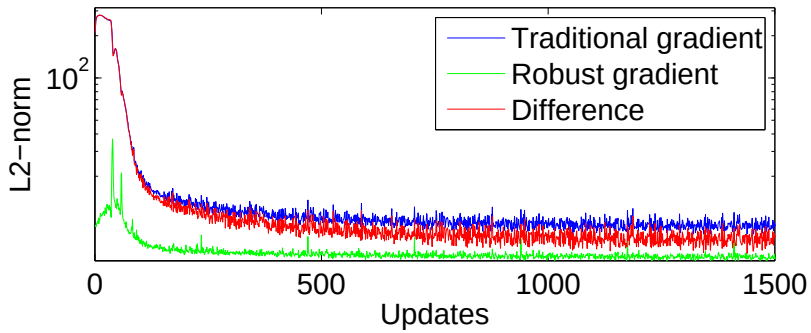
- Flip some neurons $(0 \leftrightarrow 1)$
- Equivalent RBM model can be constructed by transforming parameters



- Update: *Transform*, *update*, and *transform back*
- $2^{n_v+n_h}$ well-founded ML updates exist

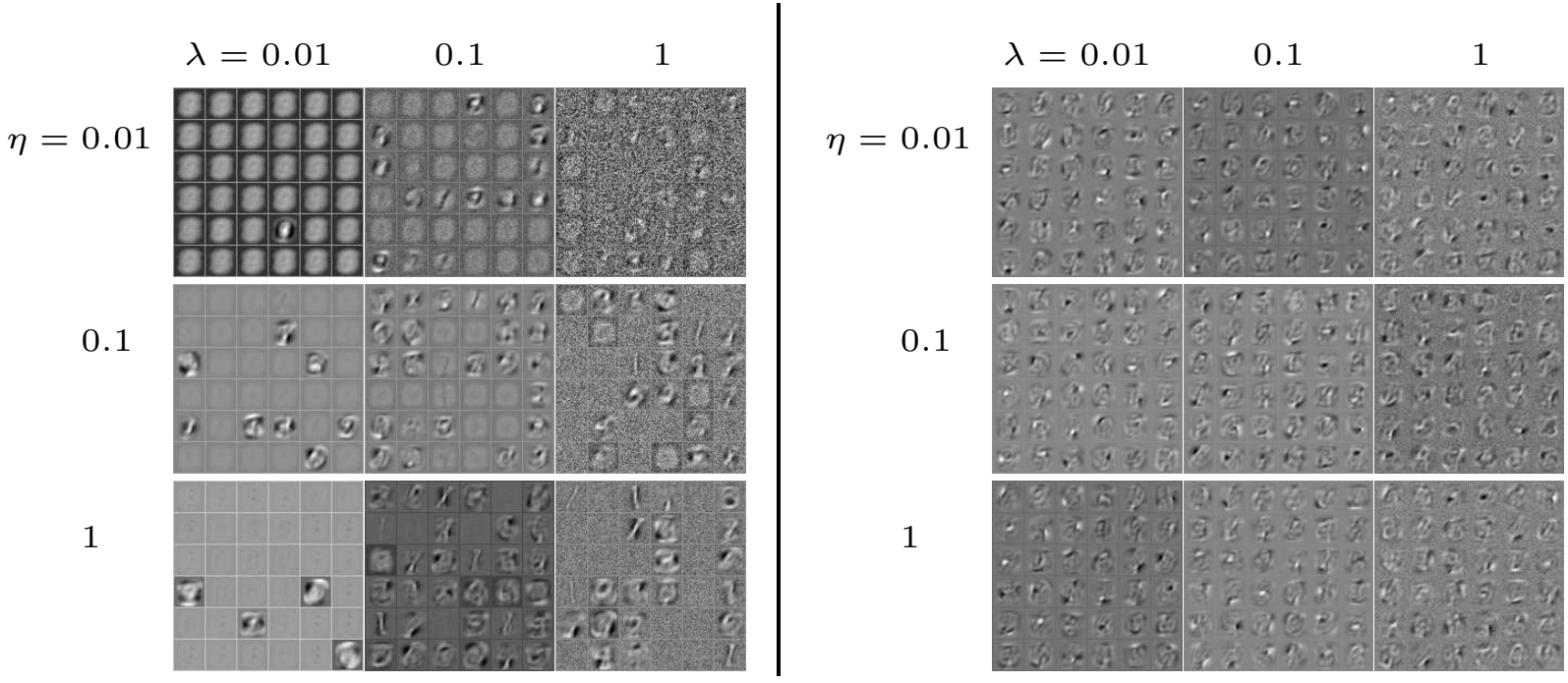
Enhanced Gradient (2/2)





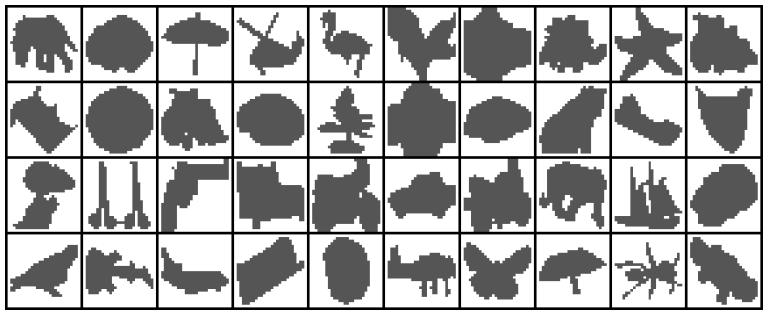
- Weighted sum of all updates
- Results in simple equations (no computational overhead)

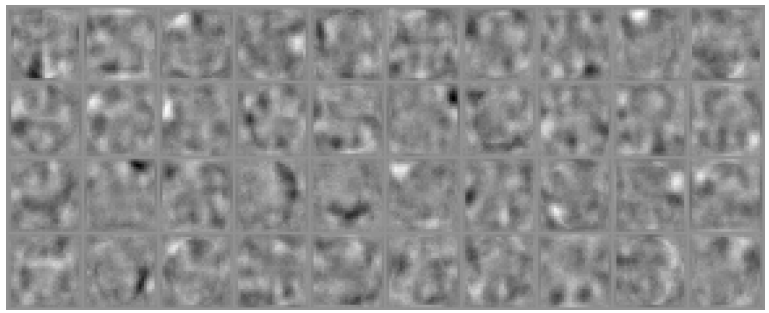
Robust Learning



- Visualization of weights after 5 epochs of learning
- Robust to setting learning parameters
  (initial learning rate η, scale of initial weights λ)
- Each hidden unit becomes useful

Experiment: Caltech 101 Silhouette Classification





Hidden neurons	Test accuracy	
	Proposed	Marlin et al. 2010
500	71.56%	65.8%
1000	**72.61%**	
2000	71.82%	

- Improved result without any laborious tuning

On-Going Work

- Continuous values, multiple layers, 3-way connections, ...
- Collaboration: speech recognition and image annotation