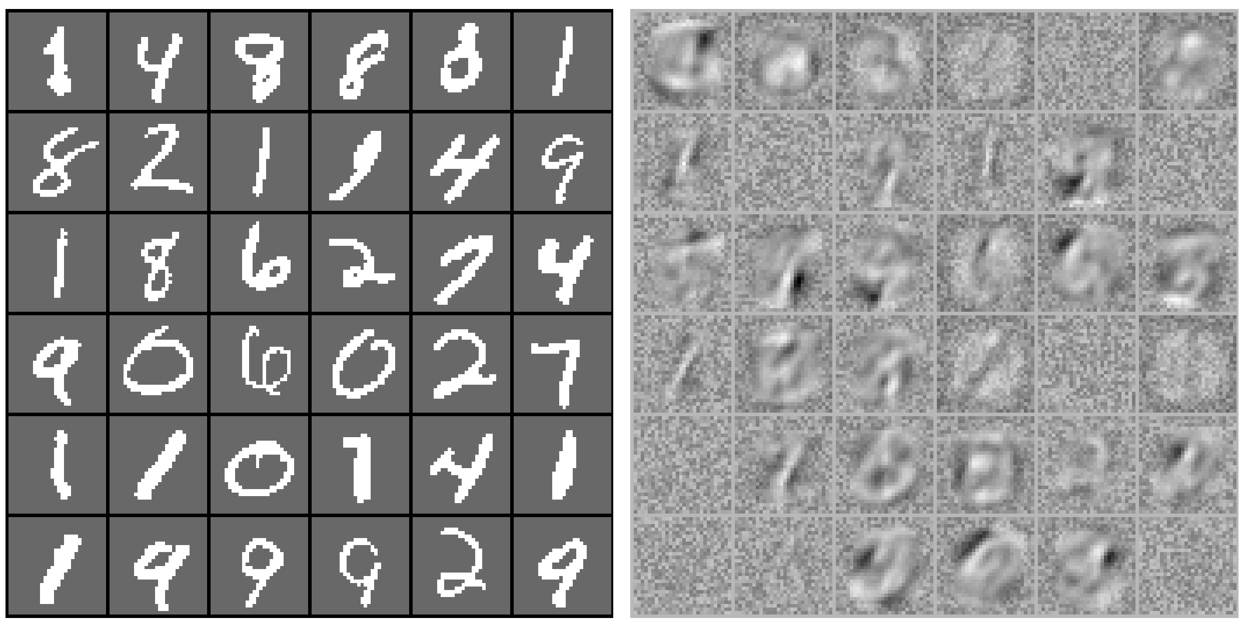
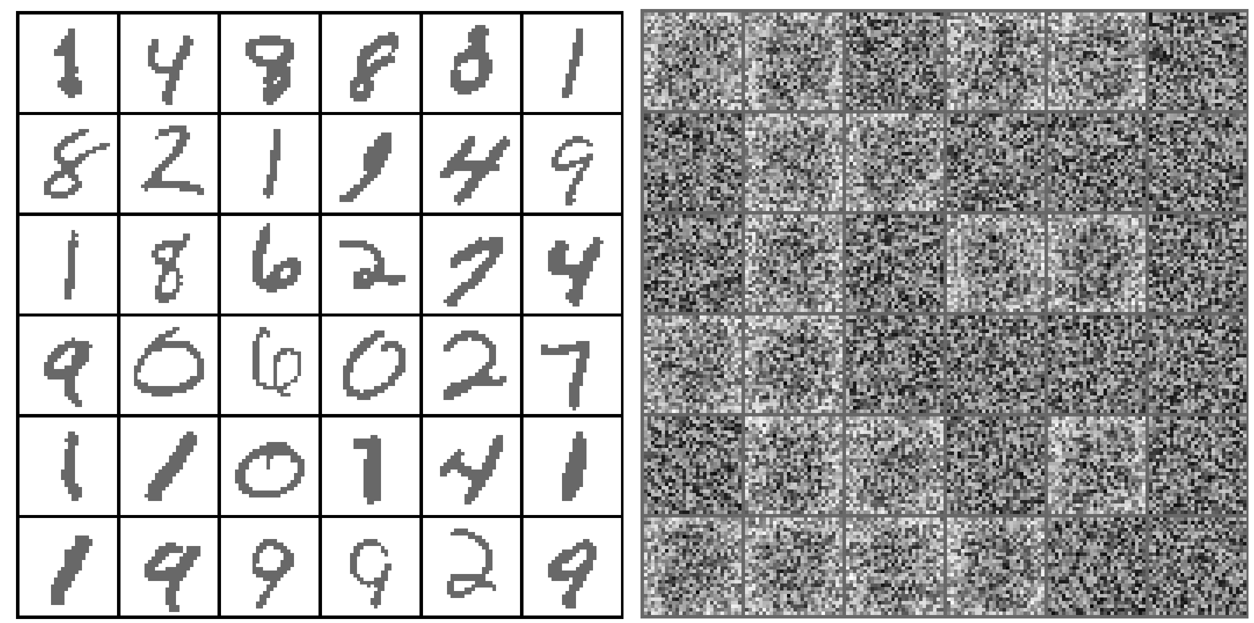
My Background (Tapani Raiko)

- 2001 Master's thesis on deep learning (see JMLR 2007)
  - Gaussian prior over a hidden representation
  - Stochastic decoder network (no encoder)
  - Unsupervised layer-wise training with variational Bayes
  - Extension to variance modelling (=heteroscedastic)
  - Computationally heavy, not-so-great performance
- 2002-2009 Relational & other latent variable models
- 2009- Concentrating on deep learning again
- (wish me luck for becoming an assoc. prof. next week)

Some lessons learned



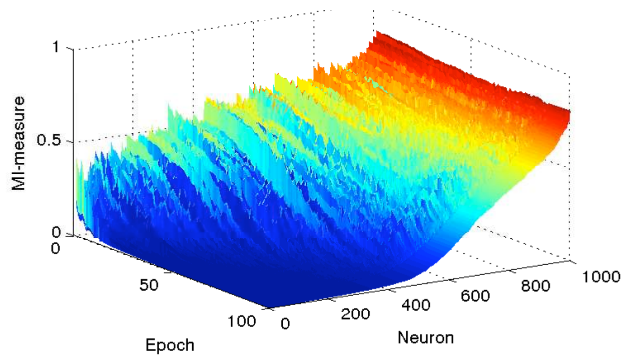
MNIST



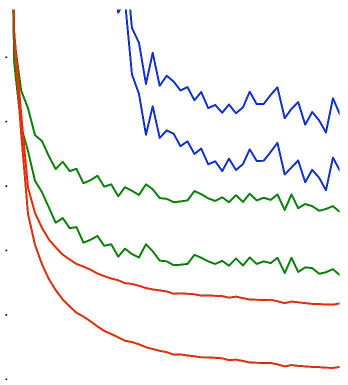
1-MNIST

Some lessons learned

- Boltzmann machines (used to) require wizardry
- Standard gradient is not invariant to representation (Cho&al Neural Computation 2013)
- Monitoring and visualizing is important (Berglund&al ICONIP 2013)



- Even back-propagation of MLPs can be improved (Vatanen&al ICONIP 2013)



- For broader use, there is a need for usable software
- However, methods are still developing (too) fast

Why unsupervised pre-training?

- Deep neural networks have lots of representational power
  - It is easy to (over)fit to training data
  - Learning will converge close to the initialization
- Another way to lessen overfitting is to use stochastisity

- Unsupervised learning is useful for handling missing values
  - E.g. handling a partially occluded image

Future: Making HMM obsolete?

- Memory capacity of a standard Hidden Markov Model is one discrete variable (no generalization)
- Memory capacity of a recurrent neural network is a continuous valued vector (interpolation&extrapolation)
- => Huge potential for improvement
- E.g. speech recognition systems are built around HMM and improved gradually for dozens of years
- Now we see hybrid approaches (part of an existing system replaced with deep learning)

Future: Towards even bigger models

- Unique type of an optimization problem
  (huge dimensionality, pathological curvature, ...)
  - Lots of room for improvement in optimization
- Limitation: Model should be stored and trained on a single computer?
  - Store parts of models on different computers
    - Requires lots of communication
  - Train approximate copies of model on different comp.
    - How to minimize communication?
    - How to guarantee good behavior?

Future: Modeling Relations

- Deep learning from relational data
(Raiko ICANN 2005, Lodhi ICONIP 2013, Liu et al. ICONIP 2013)
- Finding relations, e.g. image understanding
"Cat chases a mouse"
  - Currently state-of-the-art finds "cat" & "mouse" but not "chase" (and especially not the whole sentence)
- Big innovations awaited in
  - segmentation
(for separating individual objects, sound sources etc.)
  - understanding relations (third-order connections etc.)

Future: AI & Robotics

- For something to be considered intelligent by people, lots of basic understanding of the real world is required
- I think browsing the internet is not enough
(Who writes "you empty a mug by turning it over"?
"white swan": 3M hits, "black swan": 11M hits)
- You learn segmentation easily by pushing objects around (things that move together belong to the same object)
- I think learning to interact with the world and with people requires embodiment in a robot and a "childhood" of playing around and actively studying stuff