

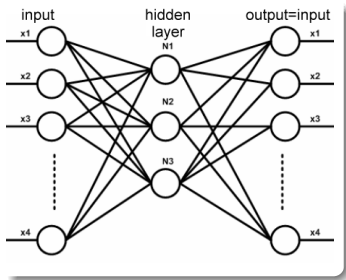
Deep Networks

most slides by Christian Osendorfer

auto-encoder networks

idea: find compact representation of inputs (unsupervised!) by

1. letting the network re-create its own inputs, i.e., $\mathbf{z}_n \equiv \mathbf{x}_n$, and
2. creating a bottleneck by using fewer hidden units than inputs.



- ▶ activations of hidden units = compact code for \mathbf{x}
- ▶ usually many ways of encoding the same input and output

these networks make a compact representation of data (“dimensionality reduction”). however, you cannot control the representation!

Caveat: autoencoders do little more than PCA!

Why Deep Networks?

Insufficiently deep architectures can be exponentially inefficient Functions compactly represented with k layers may require exponential size with $k - 1$ layers

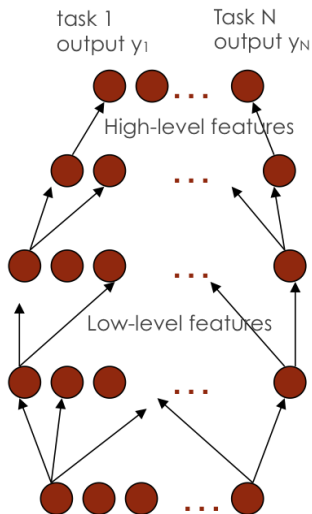
Multiple levels of latent variables allow combinatorial sharing of statistical strength.

Different tasks can share the same features

from: *Understanding and Improving Deep Learning Algorithms*, Yoshua Bengio, ML Google Distinguished Lecture, 2010

Feature Sharing

Different high-level features can be built from the same set of lower-level features (from: *Understanding and Improving Deep Learning Algorithms*, Yoshua Bengio, ML Google Distinguished Lecture, 2010)



Problem

(It seems that) Back-propagation breaks down for multiple hidden layers due to the vanishing gradient.

Recent research [Hinton et al., 2006] uses *unsupervised pre-training* by Restricted Boltzmann Machines to get into vicinity of the minimum.

Various kinds of unsupervised pre-training:

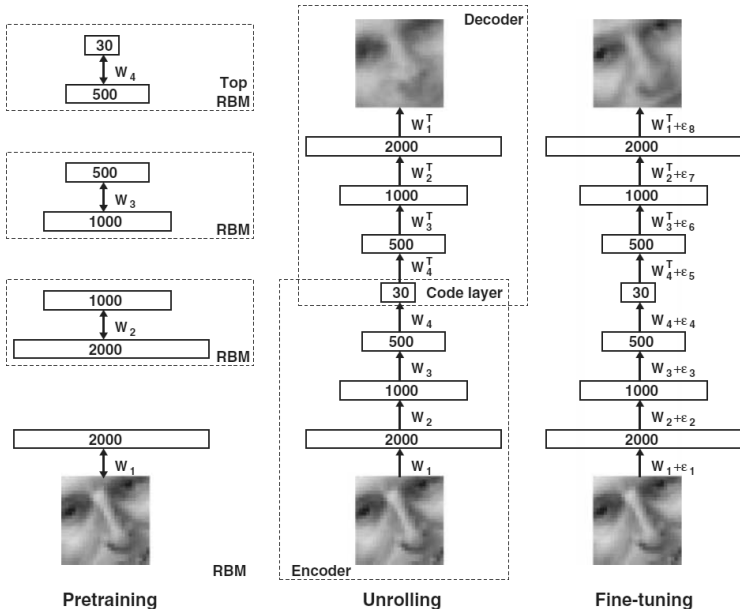
- ▶ RBM (Restricted Boltzmann Machines)
- ▶ (tied) Autoencoders and variants (denoising AE, sparse AE)
- ▶ Sparse coding
- ▶ ...

Very recent developments

Deep Learning via Hessian-Free optimisation (J. Martens, ICML 2010)

Krylov Subspace Descent for Deep Learning (O. Vinyals and D. Povey, AISTATS 2012)

Deep autoencoder Example: Face Compression



Deep autoencoder Example: Text Categorisation

Fig. 4. (A) The fraction of retrieved documents in the same class as the query when a query document from the test set is used to retrieve other test set documents, averaged over all 402,207 possible queries. (B) The codes produced by two-dimensional LSA. (C) The codes produced by a 2000-500-250-125-2 autoencoder.

