Day 2 Lecture 1

Deep Belief Networks (DBN)

[Elisa Sayrol](#)

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Signal Theory and Communications
Image Processing Group

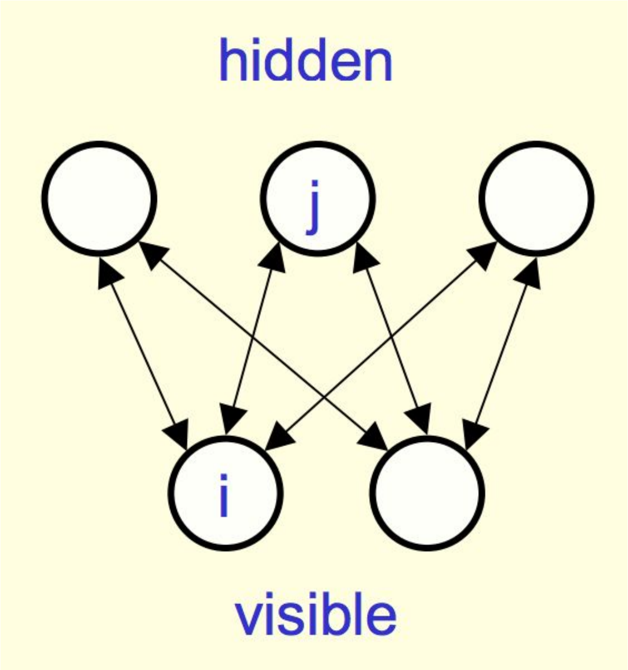
Overview

Restrictive Boltzmann Machine (RBM)

Training. Contrastive Divergence (CD)

Deep Belief Networks (DBN)

Restricted Boltzmann Machine (RBM)



- Shallow two-layer net.
- Restricted=No two nodes in a layer share a connection
- Bipartite graph.
- Bidirectional graph
  - Shared weights.
  - Different biases.

Figure: Geoffrey Hinton (2013)

Salakhutdinov, Ruslan, Andriy Mnih, and Geoffrey Hinton. "Restricted Boltzmann machines for collaborative filtering." Proceedings of the 24th international conference on Machine learning. ACM, 2007.

Restricted Boltzmann Machine (RBM)

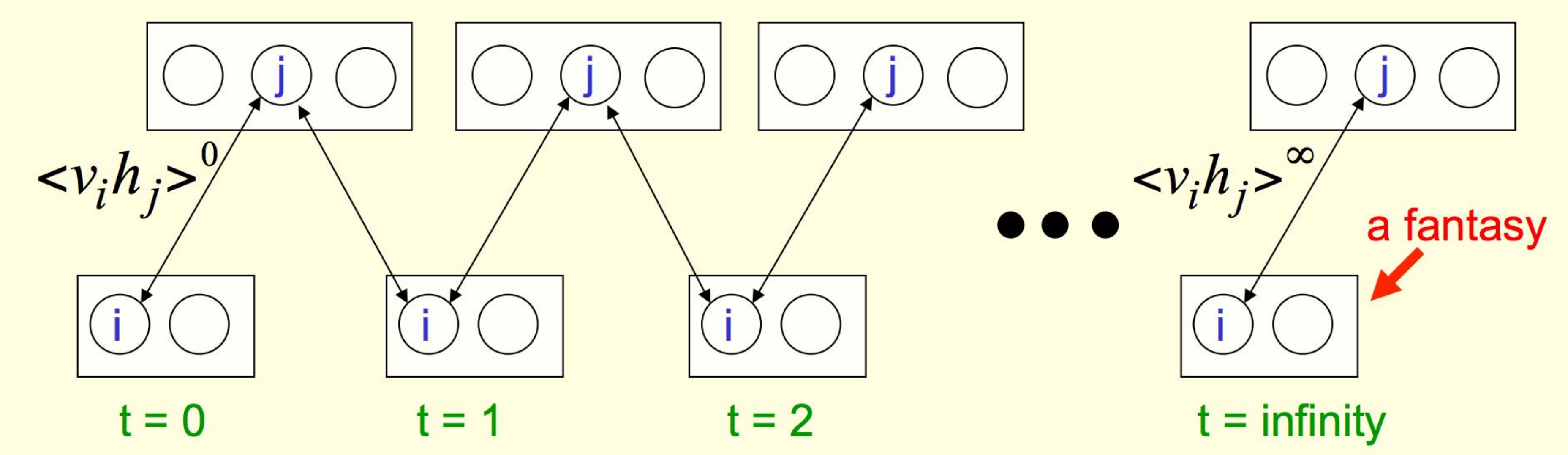
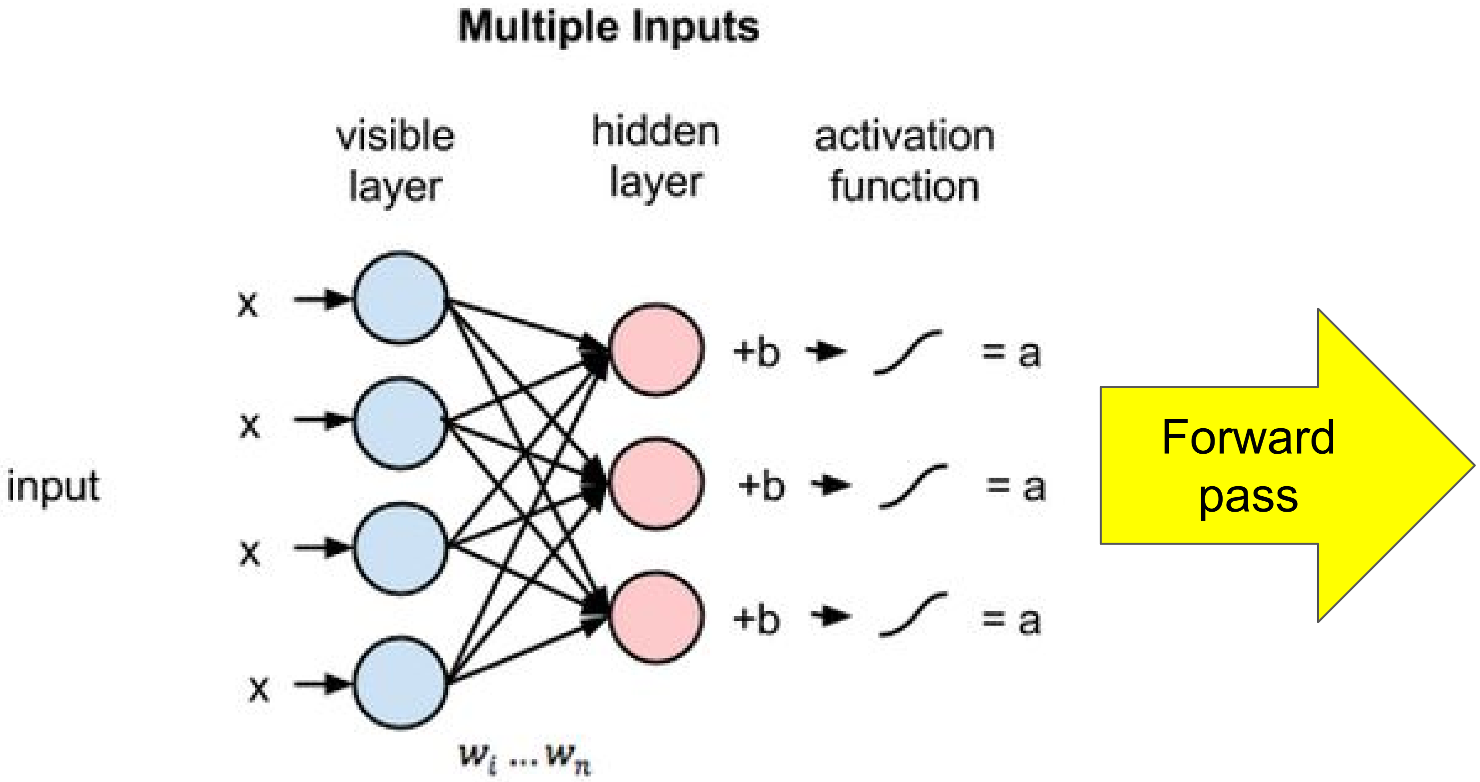


Figure: Geoffrey Hinton (2013)

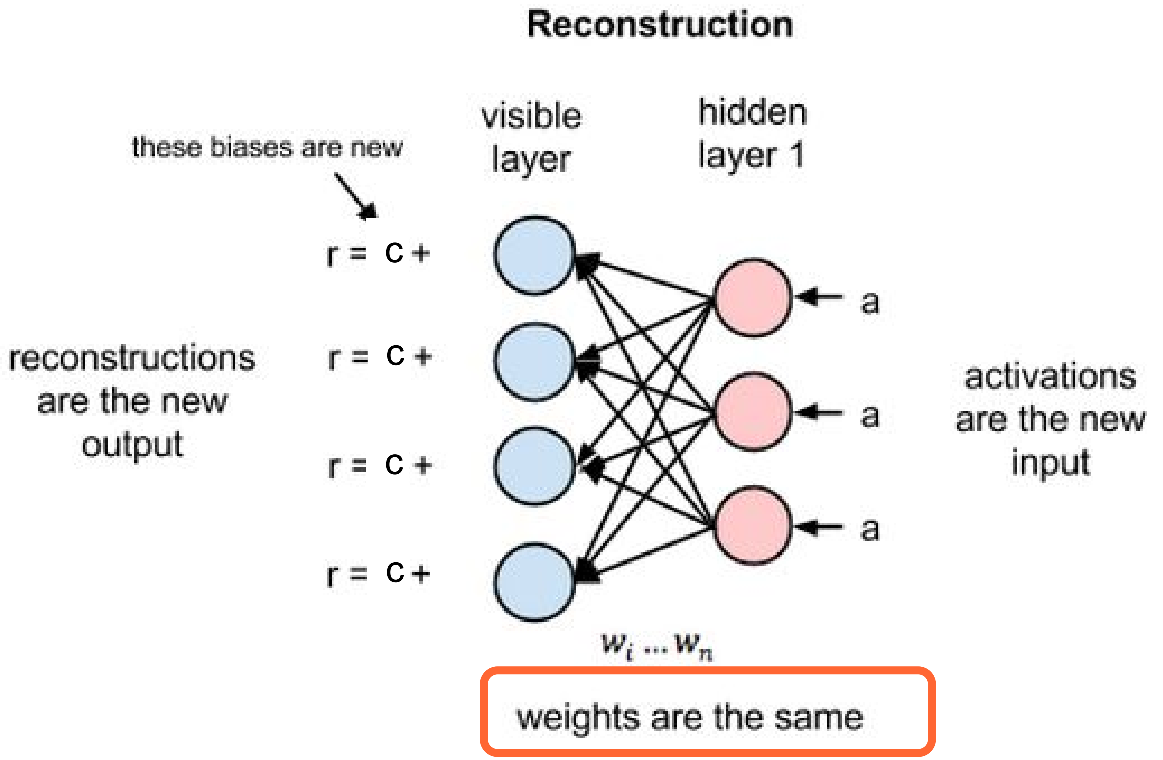
Salakhutdinov, Ruslan, Andriy Mnih, and Geoffrey Hinton. "Restricted Boltzmann machines for collaborative filtering." Proceedings of the 24th international conference on Machine learning. ACM, 2007.

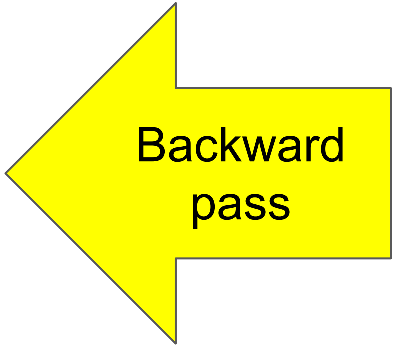
Restricted Boltzmann Machine (RBM)



DeepLearning4j, “A Beginner’s Tutorial for Restricted Boltzmann Machines”.

Restricted Boltzmann Machine (RBM)

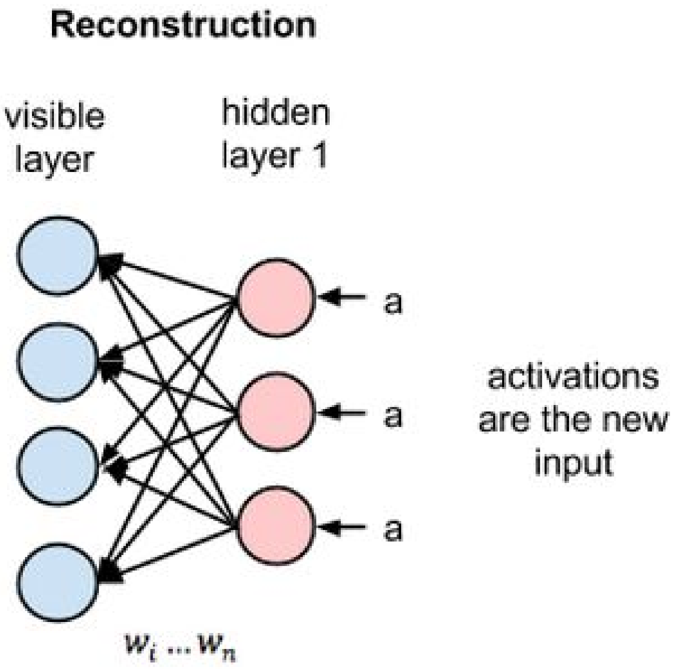




DeepLearning4j, [“A Beginner’s Tutorial for Restricted Boltzmann Machines”.]

Restricted Boltzmann Machine (RBM)

The reconstructed values at the visible layer are compared with the actual ones with the KL Divergence.



Backward pass

DeepLearning4j, “A Beginner’s Tutorial for Restricted Boltzmann Machines”.

What are the Maths behind RBMs?

(Estimation of the parameters)

Geoffrey Hinton, ["Introduction to Deep Learning & Deep Belief Nets"]() (2012)
Geoorey Hinton, ["Tutorial on Deep Belief Networks"](). NIPS 2007.

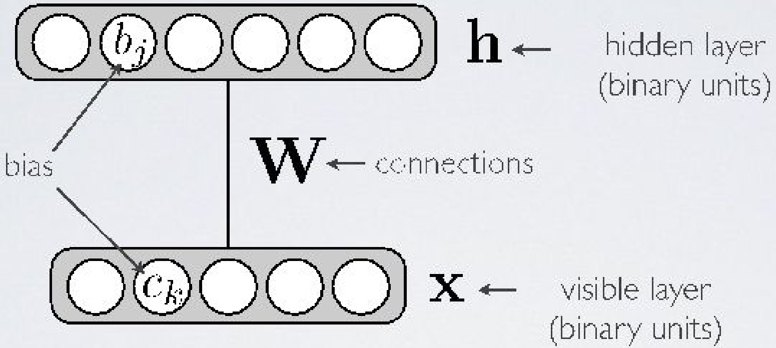
What are the Maths behind RBMs?

Other references:

Deeplearning.net: Restricted Boltzmann Machines (with Theano functions and concepts)
Hugo Larochelle: Course on NN
Let's take a look at some of his slides on RBM….

RESTRICTED BOLTZMANN MACHINE

Topics: RBM, visible layer, hidden layer, energy function



Energy function:

$$\begin{aligned} E(\mathbf{x},\mathbf{h}) &= -\mathbf{h}^\top \mathbf{W}\mathbf{x} - \mathbf{c}^\top \mathbf{x} - \mathbf{b}^\top \mathbf{h} \\ &= -\sum_j \sum_k W_{j,k} h_j x_k - \sum_k c_k x_k - \sum_j b_j h_j \end{aligned}$$

Distribution: $p(\mathbf{x},\mathbf{h}) = \exp(-E(\mathbf{x},\mathbf{h}))/Z$

partition function
(intractable)

Hugo Larochelle Slides

TRAINING

Topics: training objective

- To train an RBM, we'd like to minimize the average negative log-likelihood (NLL)

$$\frac{1}{T}\sum_t l(f(\mathbf{x}^{(t)})) = \frac{1}{T}\sum_t -\log p(\mathbf{x}^{(t)})$$

- We'd like to proceed by stochastic gradient descent

hard to compute

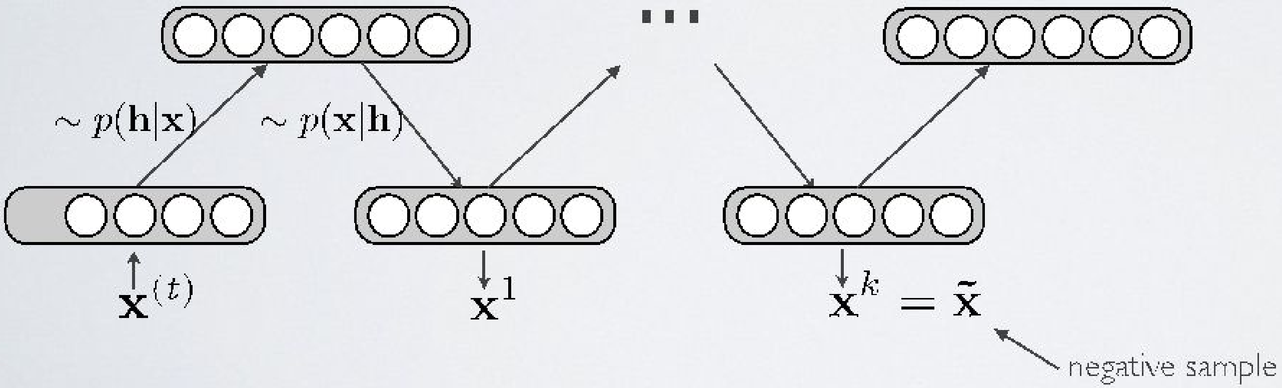
$$\frac{\partial -\log p(\mathbf{x}^{(t)})}{\partial \theta} = \underbrace{\mathrm{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x}^{(t)}, \mathbf{h})}{\partial \theta} \middle| \mathbf{x}^{(t)}\right]}_{\text{positive phase}} - \underbrace{\boxed{\mathrm{E}_{\mathbf{x},\mathbf{h}}\left[\frac{\partial E(\mathbf{x}, \mathbf{h})}{\partial \theta}\right]}}_{\text{negative phase}}$$

CONTRASTIVE DIVERGENCE (CD)

(HINTON, NEURAL COMPUTATION, 2002)

Topics: contrastive divergence, negative sample

- Idea:
  1. replace the expectation by a point estimate at $\tilde{\mathbf{x}}$
  2. obtain the point $\tilde{\mathbf{x}}$ by Gibbs sampling
  3. start sampling chain at $\mathbf{x}^{(t)}$



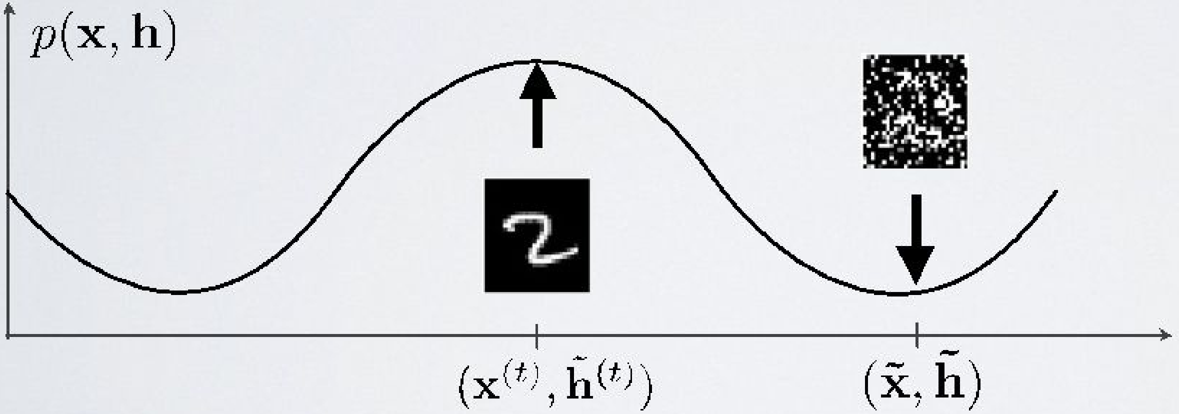
Hugo Larochelle Slides

CONTRASTIVE DIVERGENCE (CD)

(HINTON, NEURAL COMPUTATION, 2002)

Topics: contrastive divergence, negative sample

$$\mathrm{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x}^{(t)},\mathbf{h})}{\partial\theta}\middle|\mathbf{x}^{(t)}\right] \approx \frac{\partial E(\mathbf{x}^{(t)},\tilde{\mathbf{h}}^{(t)})}{\partial\theta} \qquad \mathrm{E}_{\mathbf{x},\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial\theta}\right] \approx \frac{\partial E(\tilde{\mathbf{x}},\tilde{\mathbf{h}})}{\partial\theta}$$



Hugo Larochelle Slides

DERIVATION OF THE LEARNING RULE

Topics: contrastive divergence

- Derivation of $\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial \theta}$ for $\theta = W_{jk}$

$$\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial W_{jk}} = \frac{\partial}{\partial W_{jk}}\left(-\sum_{jk} W_{jk} h_j x_k - \sum_k c_k x_k - \sum_j b_j h_j\right)$$

$$= -\frac{\partial}{\partial W_{jk}} \sum_{jk} W_{jk} h_j x_k$$

$$= -h_j x_k$$

$$\nabla_{\mathbf{W}} E(\mathbf{x},\mathbf{h}) = -\mathbf{h}\,\mathbf{x}^\top$$

Hugo Larochelle Slides

DERIVATION OF THE LEARNING RULE

Topics: contrastive divergence

- Derivation of $\mathrm{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial \theta}\Big|\mathbf{x}\right]$ for $\theta = W_{jk}$

$$\mathrm{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial W_{jk}}\Big|\mathbf{x}\right] = \mathrm{E}_{\mathbf{h}}\left[-h_j x_k \Big|\mathbf{x}\right] = \sum_{h_j \in \{0,1\}} -h_j x_k p(h_j|\mathbf{x})$$

$$= -x_k p(h_j = 1|\mathbf{x})$$

$$\mathbf{h}(\mathbf{x}) \stackrel{\text{def}}{=} \begin{pmatrix} p(h_1 = 1|\mathbf{x}) \\ \dots \\ p(h_H = 1|\mathbf{x}) \end{pmatrix} = \mathrm{sigm}(\mathbf{b} + \mathbf{W}\mathbf{x})$$

$$\mathrm{E}_{\mathbf{h}}\left[\nabla_{\mathbf{W}} E(\mathbf{x},\mathbf{h})|\mathbf{x}\right] = -\mathbf{h}(\mathbf{x})\,\mathbf{x}^{\top}$$

DERIVATION OF THE LEARNING RULE

Topics: contrastive divergence

- Given $\mathbf{x}^{(t)}$ and $\tilde{\mathbf{x}}$ the learning rule for $\theta = \mathbf{W}$ becomes

$$
\begin{aligned}
\mathbf{W} \quad & \Longleftarrow \quad \mathbf{W} - \alpha \left(\nabla_{\mathbf{W}} - \log p(\mathbf{x}^{(t)}) \right) \\
& \Longleftarrow \quad \mathbf{W} - \alpha \left(\mathrm{E}_{\mathbf{h}} \left[\nabla_{\mathbf{W}} E(\mathbf{x}^{(t)}, \mathbf{h}) \,\middle|\, \mathbf{x}^{(t)} \right] - \mathrm{E}_{\mathbf{x},\mathbf{h}} \left[\nabla_{\mathbf{W}} E(\mathbf{x}, \mathbf{h}) \right] \right) \\
& \Longleftarrow \quad \mathbf{W} - \alpha \left(\mathrm{E}_{\mathbf{h}} \left[\nabla_{\mathbf{W}} E(\mathbf{x}^{(t)}, \mathbf{h}) \,\middle|\, \mathbf{x}^{(t)} \right] - \mathrm{E}_{\mathbf{h}} \left[\nabla_{\mathbf{W}} E(\tilde{\mathbf{x}}, \mathbf{h}) \,|\, \tilde{\mathbf{x}} \right] \right) \\
& \Longleftarrow \quad \mathbf{W} + \alpha \left(\mathbf{h}(\mathbf{x}^{(t)}) \, {\mathbf{x}^{(t)}}^{\top} - \mathbf{h}(\tilde{\mathbf{x}}) \, \tilde{\mathbf{x}}^{\top} \right)
\end{aligned}
$$

CD-K: PSEUDOCODE

Topics: contrastive divergence

1. For each training example $\mathbf{x}^{(t)}$
   i. generate a negative sample $\tilde{\mathbf{x}}$ using k steps of Gibbs sampling, starting at $\mathbf{x}^{(t)}$
   ii. update parameters

$$\mathbf{W} \Longleftarrow \mathbf{W} + \alpha \left(\mathbf{h}(\mathbf{x}^{(t)})\, {\mathbf{x}^{(t)}}^\top - \mathbf{h}(\tilde{\mathbf{x}})\, \tilde{\mathbf{x}}^\top \right)$$

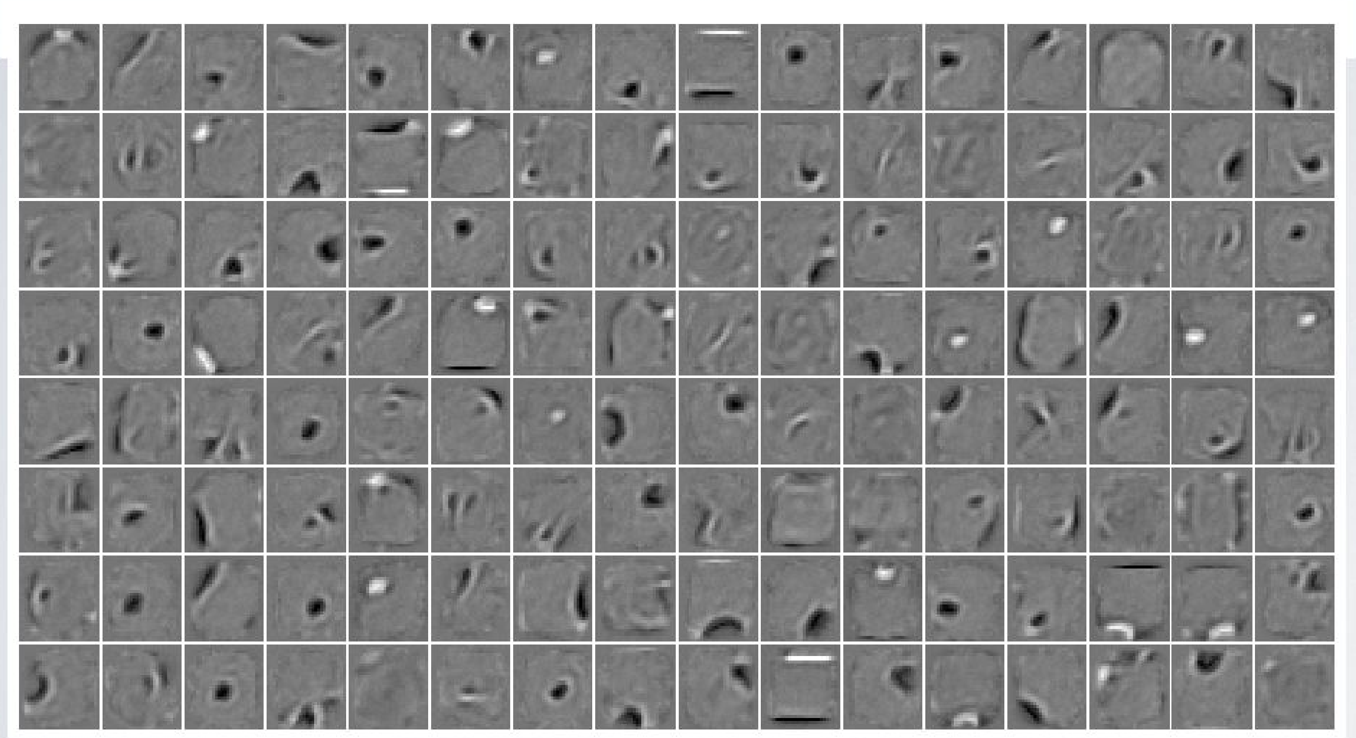
$$\mathbf{b} \Longleftarrow \mathbf{b} + \alpha \left(\mathbf{h}(\mathbf{x}^{(t)}) - \mathbf{h}(\tilde{\mathbf{x}}) \right)$$

$$\mathbf{c} \Longleftarrow \mathbf{c} + \alpha \left(\mathbf{x}^{(t)} - \tilde{\mathbf{x}} \right)$$

2. Go back to 1 until stopping criteria

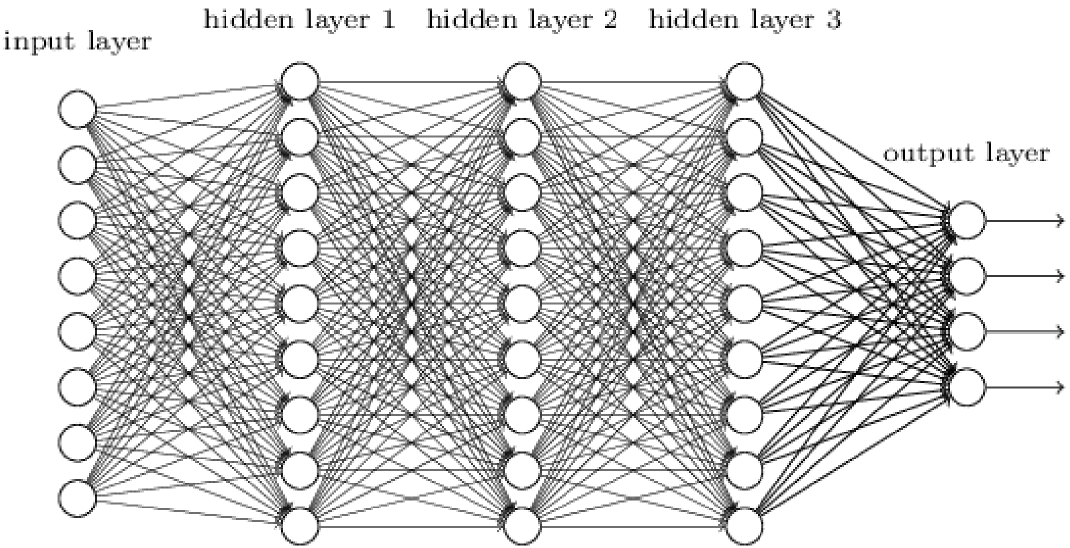
FILTERS

(LAROCHELLE ET AL., JMLR2009)



Hugo Larochelle Slides

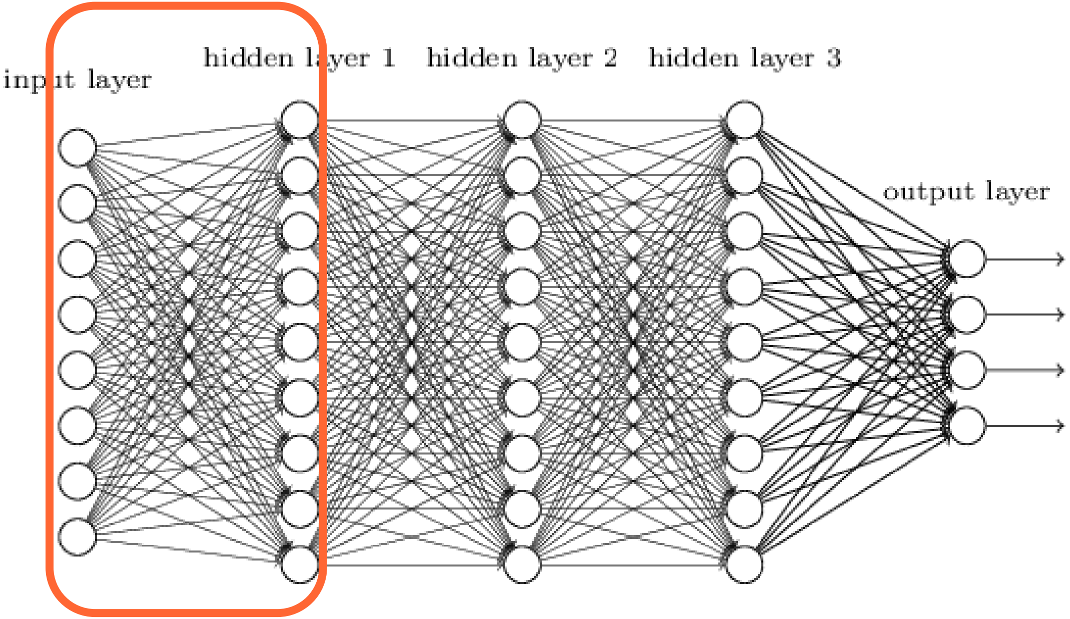
Deep Belief Networks (DBN)



- Architecture like an MLP.
- Training as a stack of RBMs.

Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." Neural computation 18, no. 7 (2006): 1527-1554.

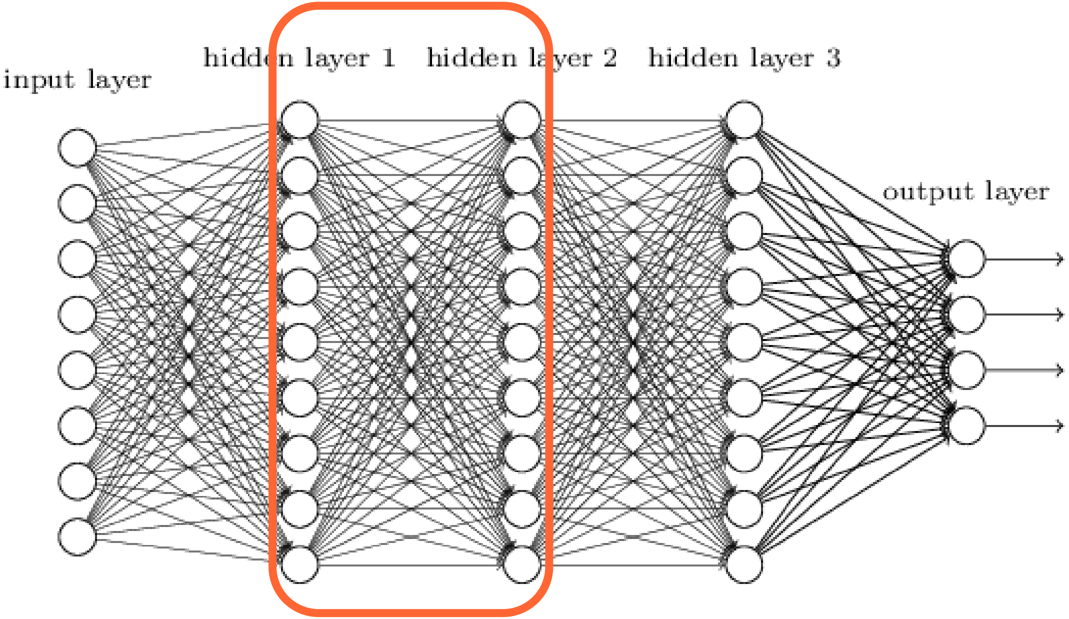
Deep Belief Networks (DBN)



- Architecture like an MLP.
- Training as a stack of RBMs.

Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." Neural computation 18, no. 7 (2006): 1527-1554.

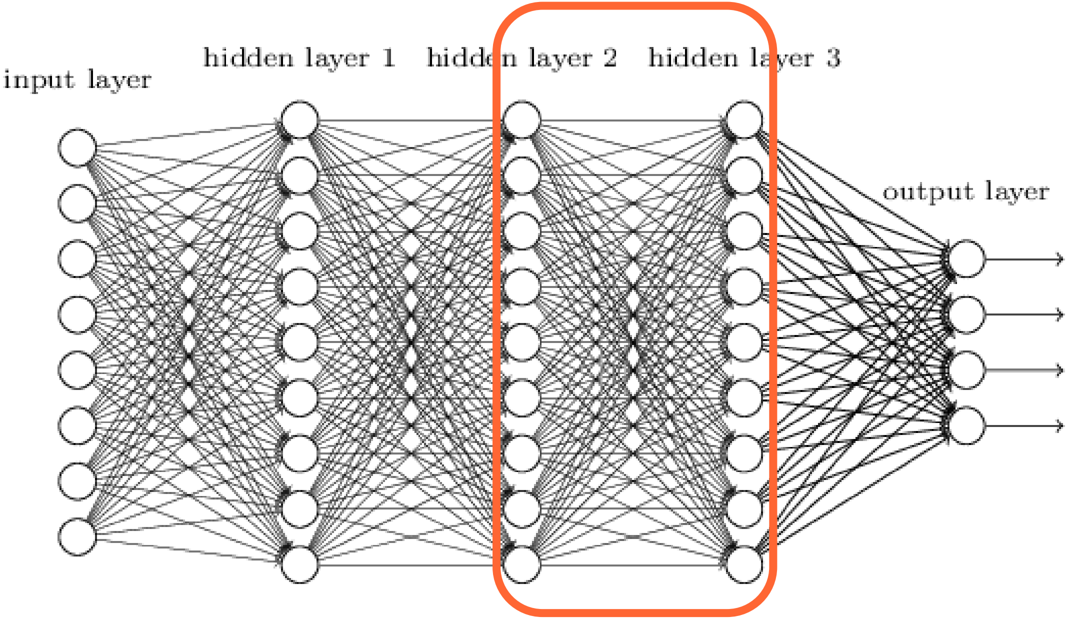
Deep Belief Networks (DBN)



- Architecture like an MLP.
- Training as a stack of RBMs.

Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." Neural computation 18, no. 7 (2006): 1527-1554.

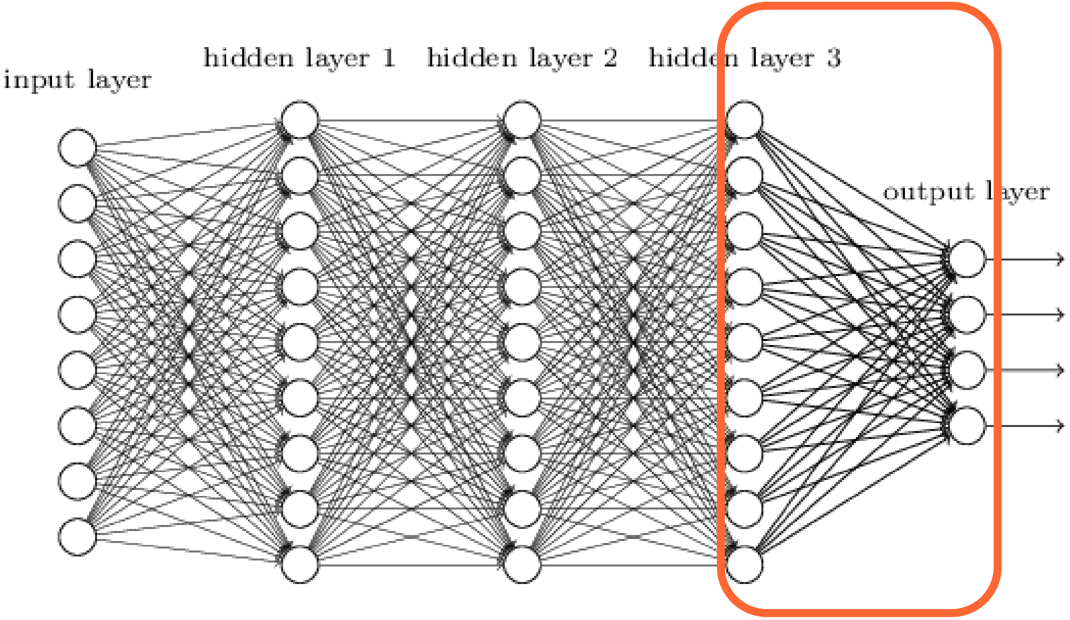
Deep Belief Networks (DBN)



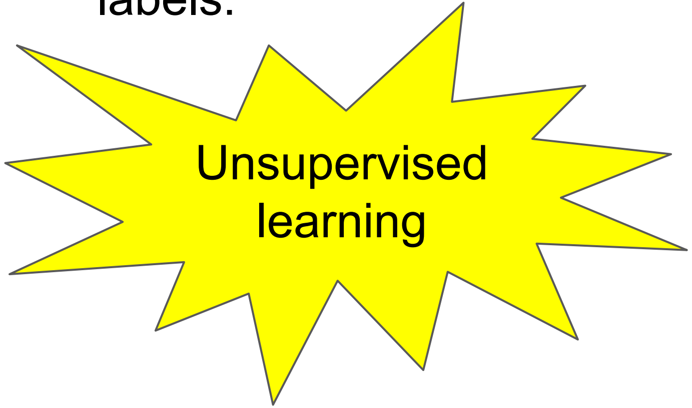
- Architecture like an MLP.
- Training as a stack of RBMs.

Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." Neural computation 18, no. 7 (2006): 1527-1554.

Deep Belief Networks (DBN)

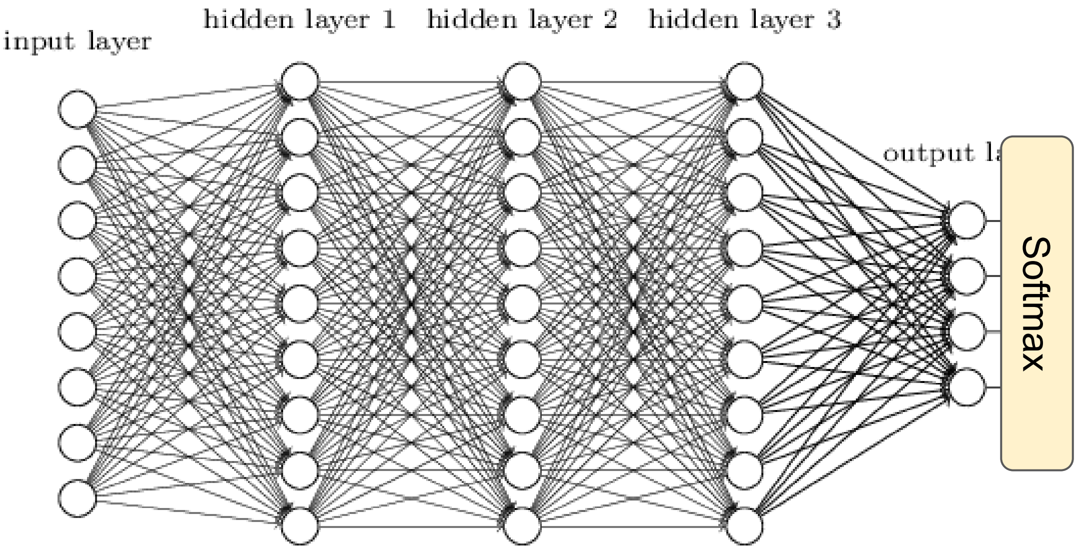


- Architecture like an MLP.
- Training as a stack of RBMs…
- ...so they do not need labels:

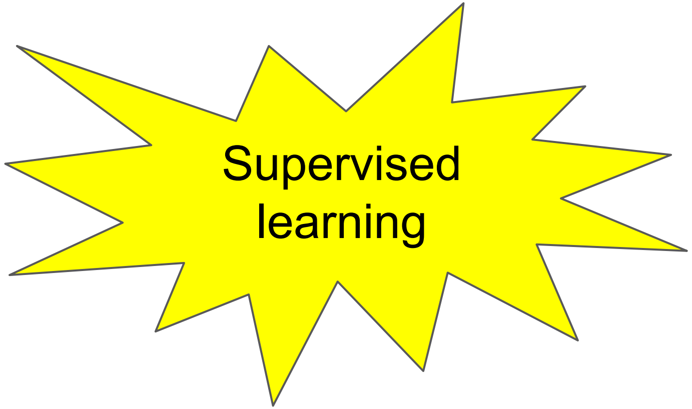


Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." Neural computation 18, no. 7 (2006): 1527-1554.

Deep Belief Networks (DBN)



After the DBN is trained, it can be fine-tuned with a **reduced amount of labels** to solve a supervised task with superior performance.



Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." Neural computation 18, no. 7 (2006): 1527-1554.

Thank You!