Deep Learning

Restricted Boltzmann Machines (RBM)

Ali Ghodsi

University of Waterloo

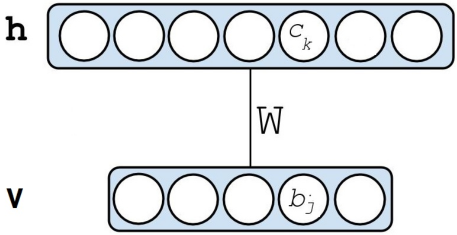
December 15, 2015

Slides are partially based on Book in preparation, Deep Learning

by Bengio, Goodfellow, and Aaron Courville, 2015

Restricted Boltzmann Machines

Restricted Boltzmann machines are some of the most common building blocks of deep probabilistic models. They are undirected probabilistic graphical models containing a layer of observable variables and a single layer of latent variables.



Restricted Boltzmann Machines

$$p(\mathbf{v},\mathbf{h}) = \frac{1}{Z} exp\{-E(\mathbf{v},\mathbf{h})\}.$$

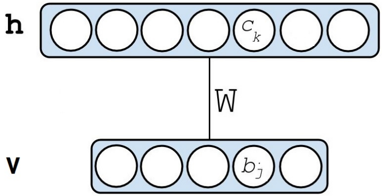
Where $E(\mathbf{v},\mathbf{h})$ is the energy function.

$$E(\mathbf{v},\mathbf{h}) = -\mathbf{b}^T\mathbf{v} - \mathbf{c}^T\mathbf{h} - \mathbf{v}^T W\mathbf{h},$$

Z is the normalizing constant partition function:

$$Z = \sum_{\mathbf{v}}\sum_{\mathbf{h}} exp\{-E(\mathbf{v},\mathbf{h})\}.$$

Restricted Boltzmann Machine (RBM)



Energy function:

$$
\begin{aligned}
E(\mathbf{v}, \mathbf{h}) &= -\mathbf{b}^T\mathbf{v} - \mathbf{c}^T\mathbf{h} - \mathbf{v}^T W\mathbf{h} \\
&= -\sum_k b_k v_k - \sum_j c_j h_j - \sum_j \sum_k W_{jk} h_j v_k
\end{aligned}
$$

Distribution: $p(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} exp\{-E(\mathbf{v}, \mathbf{h})\}$

Partition function: $Z = \sum_{\mathbf{v}} \sum_{\mathbf{h}} exp\{-E(\mathbf{v}, \mathbf{h})\}$

Conditional Distributions

The partition function Z is intractable.

Therefore the joint probability distribution is also intractable.

But $P(\mathbf{h}|\mathbf{v})$ is simple to compute and sample from.

Deriving the conditional distributions from the joint distribution.

$$
\begin{aligned}
p(\mathbf{h}|\mathbf{v}) &= \frac{p(\mathbf{h},\mathbf{v})}{p(\mathbf{v})} \\
&= \frac{1}{p(\mathbf{v})}\frac{1}{Z} exp\{\mathbf{b}^T\mathbf{v} + \mathbf{c}^T\mathbf{h} + \mathbf{v}^T W\mathbf{h}\} \\
&= \frac{1}{Z'} exp\{\mathbf{c}^T\mathbf{h} + \mathbf{v}^T W\mathbf{h}\} \\
&= \frac{1}{Z'} exp\left\{\sum_{j=1}^{n} c_j h_j + \sum_{j=1}^{n} \mathbf{v}^T W_{:j} h_j\right\} \\
&= \frac{1}{Z'} \prod_{j=1}^{n} exp\{c_j h_j + \mathbf{v}^T W_{:j} h_j\}
\end{aligned}
$$

Deriving the conditional distributions from the joint distribution.

$$
\begin{aligned}
p(\mathbf{h}|\mathbf{v}) &= \frac{p(\mathbf{h},\mathbf{v})}{p(\mathbf{v})} \\
&= \frac{1}{p(\mathbf{v})}\frac{1}{Z} exp\{\mathbf{b}^T\mathbf{v} + \mathbf{c}^T\mathbf{h} + \mathbf{v}^T W\mathbf{h}\} \\
&= \frac{1}{Z'} exp\{\mathbf{c}^T\mathbf{h} + \mathbf{v}^T W\mathbf{h}\} \\
&= \frac{1}{Z'} exp\left\{\sum_{j=1}^{n} c_j h_j + \sum_{j=1}^{n} \mathbf{v}^T W_{:j} h_j\right\} \\
&= \frac{1}{Z'} \prod_{j=1}^{n} exp\{c_j h_j + \mathbf{v}^T W_{:j} h_j\}
\end{aligned}
$$

Deriving the conditional distributions from the joint distribution.

$$
\begin{aligned}
p(\mathbf{h}|\mathbf{v}) &= \frac{p(\mathbf{h},\mathbf{v})}{p(\mathbf{v})} \\
&= \frac{1}{p(\mathbf{v})}\frac{1}{Z} exp\{\mathbf{b}^T\mathbf{v} + \mathbf{c}^T\mathbf{h} + \mathbf{v}^T W\mathbf{h}\} \\
&= \frac{1}{Z'} exp\{\mathbf{c}^T\mathbf{h} + \mathbf{v}^T W\mathbf{h}\} \\
&= \frac{1}{Z'} exp\left\{\sum_{j=1}^{n} c_j h_j + \sum_{j=1}^{n} \mathbf{v}^T W_{:j} h_j\right\} \\
&= \frac{1}{Z'} \prod_{j=1}^{n} exp\{c_j h_j + \mathbf{v}^T W_{:j} h_j\}
\end{aligned}
$$

Deriving the conditional distributions from the joint distribution.

$$
\begin{aligned}
p(\mathbf{h}|\mathbf{v}) &= \frac{p(\mathbf{h},\mathbf{v})}{p(\mathbf{v})} \\
&= \frac{1}{p(\mathbf{v})}\frac{1}{Z} exp\{\mathbf{b}^T\mathbf{v} + \mathbf{c}^T\mathbf{h} + \mathbf{v}^T W\mathbf{h}\} \\
&= \frac{1}{Z'} exp\{\mathbf{c}^T\mathbf{h} + \mathbf{v}^T W\mathbf{h}\} \\
&= \frac{1}{Z'} exp\left\{\sum_{j=1}^{n} c_j h_j + \sum_{j=1}^{n} \mathbf{v}^T W_{:j} h_j\right\} \\
&= \frac{1}{Z'} \prod_{j=1}^{n} exp\{c_j h_j + \mathbf{v}^T W_{:j} h_j\}
\end{aligned}
$$

Deriving the conditional distributions from the joint distribution.

$$
\begin{aligned}
p(\mathbf{h}|\mathbf{v}) &= \frac{p(\mathbf{h},\mathbf{v})}{p(\mathbf{v})} \\
&= \frac{1}{p(\mathbf{v})}\frac{1}{Z} exp\{\mathbf{b}^T\mathbf{v} + \mathbf{c}^T\mathbf{h} + \mathbf{v}^T W\mathbf{h}\} \\
&= \frac{1}{Z'} exp\{\mathbf{c}^T\mathbf{h} + \mathbf{v}^T W\mathbf{h}\} \\
&= \frac{1}{Z'} exp\left\{\sum_{j=1}^{n} c_j h_j + \sum_{j=1}^{n} \mathbf{v}^T W_{:j} h_j\right\} \\
&= \frac{1}{Z'} \prod_{j=1}^{n} exp\{c_j h_j + \mathbf{v}^T W_{:j} h_j\}
\end{aligned}
$$

The distributions over the individual binary h_j

$$
\begin{aligned}
P(h_j = 1|\mathbf{v}) &= \frac{P(h_j = 1, \mathbf{v})}{P(h_j = 0, \mathbf{v}) + P(h_j = 1, \mathbf{v})} \\
&= \frac{\exp\{c_j + \mathbf{v}^T W_{:j}\}}{exp\{0\} + exp\{c_j + \mathbf{v}^T W_{:j}\}} \\
&= sigmoid(c_j + \mathbf{v}^T W_{:j})
\end{aligned}
$$

$$
P(\mathbf{h}|\mathbf{v}) = \prod_{j=1}^{n} sigmoid(c_j + \mathbf{v}^T W_{:j})
$$

$$
P(\mathbf{v}|\mathbf{h}) = \prod_{i=1}^{d} sigmoid(b_i + W_{i:}\mathbf{h})
$$

RBM Gibbs Sampling

Step1: Sample $\mathbf{h}^{(l)} \sim P(\mathbf{h}|\mathbf{v}^{(l)})$.

We can simultaneously and independently sample from all the elements of $\mathbf{h}^{(l)}$ given $\mathbf{v}^{(l)}$.

Step 2: Sample $\mathbf{v}^{(l+1)} \sim P(\mathbf{v}|\mathbf{h}^{(l)})$.

We can simultaneously and independently sample from all the elements of $\mathbf{v}^{(l+1)}$ given $\mathbf{h}^{(l)}$.

Training Restricted Boltzmann Machines

The log-likelihood is given by:

$$\begin{aligned}
\ell(W, \mathbf{b}, \mathbf{c}) &= \sum_{t=1}^{n} logP(\mathbf{v}^{(t)}) \\
&= \sum_{t=1}^{n} log \sum_{\mathbf{h}} P(\mathbf{v}^{(t)}, \mathbf{h}) \\
&= \sum_{t=1}^{n} log \sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)}, \mathbf{h})\} \ ) - n \ logZ \\
&= \sum_{t=1}^{n} log \sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)}, \mathbf{h})\} \ ) - n \ log \sum_{\mathbf{v},\mathbf{h}} exp\{-E(\mathbf{v}, \mathbf{h})\}
\end{aligned}$$

Training Restricted Boltzmann Machines

The log-likelihood is given by:

$$\begin{aligned}
\ell(W,\mathbf{b},\mathbf{c}) &= \sum_{t=1}^{n} logP(\mathbf{v}^{(t)}) \\
&= \sum_{t=1}^{n} log \sum_{\mathbf{h}} P(\mathbf{v}^{(t)},\mathbf{h}) \\
&= \sum_{t=1}^{n} log \sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)},\mathbf{h})\}\ ) - n\ logZ \\
&= \sum_{t=1}^{n} log \sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)},\mathbf{h})\}\ ) - n\ log \sum_{\mathbf{v},\mathbf{h}} exp\{-E(\mathbf{v},\mathbf{h})\}
\end{aligned}$$

Training Restricted Boltzmann Machines

The log-likelihood is given by:

$$
\begin{aligned}
\ell(W, \mathbf{b}, \mathbf{c}) &= \sum_{t=1}^{n} log P(\mathbf{v}^{(t)}) \\
&= \sum_{t=1}^{n} log \sum_{\mathbf{h}} P(\mathbf{v}^{(t)}, \mathbf{h}) \\
&= \sum_{t=1}^{n} log \sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)}, \mathbf{h})\}\ ) - n\ log Z \\
&= \sum_{t=1}^{n} log \sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)}, \mathbf{h})\}\ ) - n\ log \sum_{\mathbf{v},\mathbf{h}} exp\{-E(\mathbf{v}, \mathbf{h})\}
\end{aligned}
$$

Training Restricted Boltzmann Machines

The log-likelihood is given by:

$$\begin{aligned}
\ell(W, \mathbf{b}, \mathbf{c}) &= \sum_{t=1}^{n} log P(\mathbf{v}^{(t)}) \\
&= \sum_{t=1}^{n} log \sum_{\mathbf{h}} P(\mathbf{v}^{(t)}, \mathbf{h}) \\
&= \sum_{t=1}^{n} log \sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)}, \mathbf{h})\}\ ) - n\ logZ \\
&= \sum_{t=1}^{n} log \sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)}, \mathbf{h})\}\ ) - n\ log \sum_{\mathbf{v},\mathbf{h}} exp\{-E(\mathbf{v}, \mathbf{h})\}
\end{aligned}$$

Maximizing the likelihood

$\theta = \{\mathbf{b}, \mathbf{c}, W\}$:

$$\ell(\theta) = \sum_{t=1}^{n} log \sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)}, \mathbf{h})\} \;) - n \; log \sum_{\mathbf{v},\mathbf{h}} exp\{-E(\mathbf{v}, \mathbf{h})\}$$

$$\begin{aligned}
\nabla_\theta \ell(\theta) &= \nabla_\theta \sum_{t=1}^{n} log \sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)}, \mathbf{h})\} \;) - n \; \nabla_\theta log \sum_{\mathbf{v},\mathbf{h}} exp\{-E(\mathbf{v}, \mathbf{h})\} \\
&= \sum_{t=1}^{n} \frac{\sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)}, \mathbf{h})\} \nabla \theta - E(\mathbf{v}^{(t)}, \mathbf{h})}{\sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)}, \mathbf{h})\}} \\
&\quad - n \frac{\sum_{\mathbf{v},\mathbf{h}} exp\{-E(\mathbf{v}, \mathbf{h})\} \nabla \theta - E(\mathbf{v}, \mathbf{h})}{\sum_{\mathbf{v},\mathbf{h}} exp\{-E(\mathbf{v}, \mathbf{h})\}} \\
&= \sum_{t=1}^{n} \mathbb{E}_{P(\mathbf{h}|\mathbf{v}^{(t)})}[\nabla_\theta - E(\mathbf{v}^{(t)}, \mathbf{h})] - n \; \mathbb{E}_{P(\mathbf{h},\mathbf{v})}[\nabla_\theta - E(\mathbf{v}, \mathbf{h})]
\end{aligned}$$

Maximizing the likelihood

$\theta = \{\mathbf{b}, \mathbf{c}, W\}$:

$$\ell(\theta) = \sum_{t=1}^{n} log \sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)}, \mathbf{h})\} \;) - n \; log \sum_{\mathbf{v},\mathbf{h}} exp\{-E(\mathbf{v}, \mathbf{h})\}$$

$$\begin{aligned}
\nabla_\theta \ell(\theta) &= \nabla_\theta \sum_{t=1}^{n} log \sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)}, \mathbf{h})\} \;) - n \; \nabla_\theta log \sum_{\mathbf{v},\mathbf{h}} exp\{-E(\mathbf{v}, \mathbf{h})\} \\
&= \sum_{t=1}^{n} \frac{\sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)}, \mathbf{h})\} \nabla\theta - E(\mathbf{v}^{(t)}, \mathbf{h})}{\sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)}, \mathbf{h})\}} \\
&\quad - n \frac{\sum_{\mathbf{v},\mathbf{h}} exp\{-E(\mathbf{v}, \mathbf{h})\} \nabla\theta - E(\mathbf{v}, \mathbf{h})}{\sum_{\mathbf{v},\mathbf{h}} exp\{-E(\mathbf{v}, \mathbf{h})\}} \\
&= \sum_{t=1}^{n} \mathbb{E}_{P(\mathbf{h}|\mathbf{v}^{(t)})}[\nabla_\theta - E(\mathbf{v}^{(t)}, \mathbf{h})] - n \; \mathbb{E}_{P(\mathbf{h},\mathbf{v})}[\nabla_\theta - E(\mathbf{v}, \mathbf{h})]
\end{aligned}$$

Maximizing the likelihood

$\theta = \{\mathbf{b}, \mathbf{c}, W\}$:

$$\ell(\theta) = \sum_{t=1}^{n} log \sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)}, \mathbf{h})\}\ ) - n\ log \sum_{\mathbf{v},\mathbf{h}} exp\{-E(\mathbf{v}, \mathbf{h})\}$$

$$\begin{aligned}
\nabla_\theta \ell(\theta) &= \nabla_\theta \sum_{t=1}^{n} log \sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)}, \mathbf{h})\}\ ) - n\ \nabla_\theta log \sum_{\mathbf{v},\mathbf{h}} exp\{-E(\mathbf{v}, \mathbf{h})\} \\
&= \sum_{t=1}^{n} \frac{\sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)}, \mathbf{h})\} \nabla\theta - E(\mathbf{v}^{(t)}, \mathbf{h})}{\sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)}, \mathbf{h})\}} \\
&\quad - n \frac{\sum_{\mathbf{v},\mathbf{h}} exp\{-E(\mathbf{v}, \mathbf{h})\} \nabla\theta - E(\mathbf{v}, \mathbf{h})}{\sum_{\mathbf{v},\mathbf{h}} exp\{-E(\mathbf{v}, \mathbf{h})\}} \\
&= \sum_{t=1}^{n} \mathbb{E}_{P(\mathbf{h}|\mathbf{v}^{(t)})}[\nabla_\theta - E(\mathbf{v}^{(t)}, \mathbf{h})] - n\ \mathbb{E}_{P(\mathbf{h},\mathbf{v})}[\nabla_\theta - E(\mathbf{v}, \mathbf{h})]
\end{aligned}$$

Maximizing the likelihood

$\theta = \{\mathbf{b}, \mathbf{c}, W\}$:

$$\ell(\theta) = \sum_{t=1}^{n} log \sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)}, \mathbf{h})\}\ ) - n\ log \sum_{\mathbf{v},\mathbf{h}} exp\{-E(\mathbf{v}, \mathbf{h})\}$$

$$\begin{aligned}
\nabla_\theta \ell(\theta) &= \nabla_\theta \sum_{t=1}^{n} log \sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)}, \mathbf{h})\}\ ) - n\ \nabla_\theta log \sum_{\mathbf{v},\mathbf{h}} exp\{-E(\mathbf{v}, \mathbf{h})\} \\
&= \sum_{t=1}^{n} \frac{\sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)}, \mathbf{h})\} \nabla\theta - E(\mathbf{v}^{(t)}, \mathbf{h})}{\sum_{\mathbf{h}} exp\{-E(\mathbf{v}^{(t)}, \mathbf{h})\}} \\
&\quad - n \frac{\sum_{\mathbf{v},\mathbf{h}} exp\{-E(\mathbf{v}, \mathbf{h})\} \nabla\theta - E(\mathbf{v}, \mathbf{h})}{\sum_{\mathbf{v},\mathbf{h}} exp\{-E(\mathbf{v}, \mathbf{h})\}} \\
&= \sum_{t=1}^{n} \mathbb{E}_{P(\mathbf{h}|\mathbf{v}^{(t)})}[\nabla_\theta - E(\mathbf{v}^{(t)}, \mathbf{h})] - n\ \mathbb{E}_{P(\mathbf{h},\mathbf{v})}[\nabla_\theta - E(\mathbf{v}, \mathbf{h})]
\end{aligned}$$

The gradient of the negative energy function

$$\begin{aligned}\nabla_W - E(\mathbf{v},\mathbf{h}) &= \frac{\partial}{\partial W}(\mathbf{b}^T\mathbf{v} + \mathbf{c}^T h + \mathbf{v}^T W\mathbf{h}) \\ &= \mathbf{h}\mathbf{v}^T\end{aligned}$$

$$\begin{aligned}\nabla_{\mathbf{b}} - E(\mathbf{v},\mathbf{h}) &= \frac{\partial}{\partial \mathbf{b}}(\mathbf{b}^T\mathbf{v} + \mathbf{c}^T h + \mathbf{v}^T W\mathbf{h}) \\ &= \mathbf{v}\end{aligned}$$

$$\begin{aligned}\nabla_{\mathbf{c}} - E(\mathbf{v},\mathbf{h}) &= \frac{\partial}{\partial \mathbf{c}}(\mathbf{b}^T\mathbf{v} + \mathbf{c}^T h + \mathbf{v}^T W\mathbf{h}) \\ &= \mathbf{h}\end{aligned}$$

$$\nabla_\theta \ell(\theta) = \sum_{t=1}^{n} \mathbb{E}_{P(\mathbf{h}|\mathbf{v}^{(t)})}[\nabla_\theta - E(\mathbf{v}^{(t)}, \mathbf{h})] - n\ \mathbb{E}_{P(\mathbf{h},\mathbf{v})}[\nabla_\theta - E(\mathbf{v}, \mathbf{h})]$$

$$\begin{aligned}
\nabla_W \ell(W, \mathbf{b}, \mathbf{c}) &= \sum_{t=1}^{n} \hat{\mathbf{h}}^{(t)} \mathbf{v}^{(t)^T} - n\mathbb{E}_{P(\mathbf{v},\mathbf{h})}[\mathbf{h}\mathbf{v}^T] \\
\nabla_{\mathbf{b}} \ell(W, \mathbf{b}, \mathbf{c}) &= \sum_{t=1}^{n} \mathbf{v}^{(t)^T} - n\mathbb{E}_{P(\mathbf{v},\mathbf{h})}[\mathbf{v}] \\
\nabla_{\mathbf{c}} \ell(W, \mathbf{b}, c) &= \sum_{t=1}^{n} \hat{\mathbf{h}}^{(t)} - n\mathbb{E}_{P(\mathbf{v},\mathbf{h})}[\mathbf{h}]
\end{aligned}$$

where

$$\hat{\mathbf{h}}^{(t)} = \mathbb{E}_{P(\mathbf{h},\mathbf{v}^{(t)})}[\mathbf{h}] = sigmoid(\mathbf{c} + \mathbf{v}^{(t)} W).$$

it is impractical to compute the exact log-likelihood gradient.

Contrastive Divergence

Idea:

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

$$\mathbb{E}_{P(\mathbf{h},\mathbf{v})}[\nabla_\theta - E(\mathbf{v},\mathbf{h})] \approx \nabla_\theta - E(\mathbf{v},\mathbf{h})|_{\mathbf{v}=\tilde{\mathbf{v}},\mathbf{h}=\tilde{\mathbf{h}}}$$

Set $\in$, the step size, to a small positive number
Set k, the number of Gibbs steps, high enough to allow a Markov chain of $p(\mathbf{v};\theta)$ to mix when initialized from p_{data}. Perhaps 1-20 to train an RBM on a small image patch.
while Not converged do
Sample a mini batch of m examples from the training set $\{\mathbf{v}^{(1)},\ldots,\mathbf{v}^{(m)}\}$.
$\nabla_W \leftarrow \frac{1}{m}\sum_{t=1}^{m}\mathbf{v}^{(t)}\hat{\mathbf{h}}^{(t)}\,T$
$\nabla_{\mathbf{b}} \leftarrow \frac{1}{m}\sum_{t=1}^{m}\mathbf{v}^{(t)}$
$\nabla_{\mathbf{c}} \leftarrow \frac{1}{m}\sum_{t=1}^{m}\hat{\mathbf{h}}^{(t)}$
for $t = 1$ to m do
$\hat{\mathbf{v}}^{(t)} \leftarrow \mathbf{v}^{(t)}$
end for
for $\ell = 1$ to k do
for $t = 1$ to m do

$\hat{\mathbf{h}}^{(t)}$ sampled from $\prod_{j=1}^{n} sigmoid(c_j + \hat{\mathbf{v}}^{(t)}TW_{:,j})$.
$\hat{\mathbf{v}}^{(t)}$ sampled from $\prod_{i=1}^{d} sigmoid(\mathbf{b}_j + W_{i,:}\hat{\mathbf{h}}^{(t)})$.
end for
end for
$\hat{\mathbf{h}}^{(t)} \leftarrow sigmoid(\mathbf{c} + \hat{\mathbf{v}}^{(t)}TW)$
$\nabla_W \leftarrow \nabla_W - \frac{1}{m}\sum_{t=1}^{m} \mathbf{v}^{(t)}\hat{\mathbf{h}}^{(t)}T$
$\nabla_{\mathbf{b}} \leftarrow \nabla_{\mathbf{b}} - \frac{1}{m}\sum_{t=1}^{m} \mathbf{v}^{(t)}$
$\nabla_{\mathbf{c}} \leftarrow \nabla_{\mathbf{c}} - \frac{1}{m}\sum_{t=1}^{m} \hat{\mathbf{h}}^{(t)}$
$W \leftarrow W + \in \nabla_W$
$\mathbf{b} \leftarrow \mathbf{b} + \in \nabla_{\mathbf{b}} \sum$
$\mathbf{c} \leftarrow \mathbf{c} + \in \nabla_{\mathbf{c}} \sum$
end while

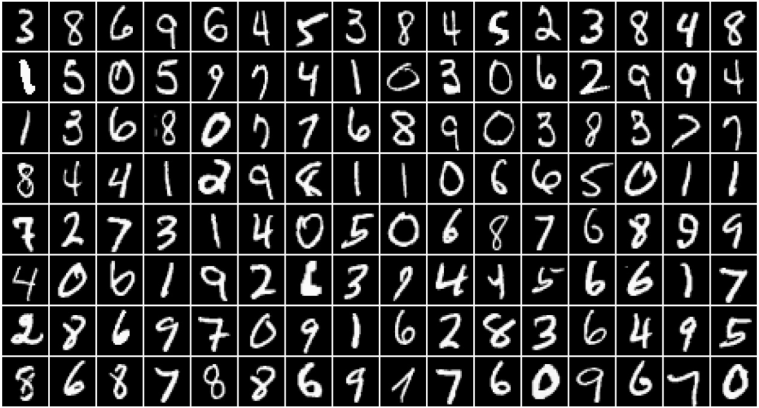
Pseudo code

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

$$W \Leftarrow W + \alpha \left(\mathbf{h}(\mathbf{v}^{(t)}) x^{(t)} - \mathbf{h}(\tilde{\mathbf{v}}) \tilde{\mathbf{v}}^{T} \right)$$
$$\mathbf{b} \Leftarrow \mathbf{b} + \alpha \left(\mathbf{h}(\mathbf{v}^{(t)}) - \mathbf{h}(\tilde{\mathbf{v}}) \right)$$
$$\mathbf{c} \Leftarrow \mathbf{c} + \alpha \left(\mathbf{v}^{(t)} - \tilde{\mathbf{v}}) \right)$$

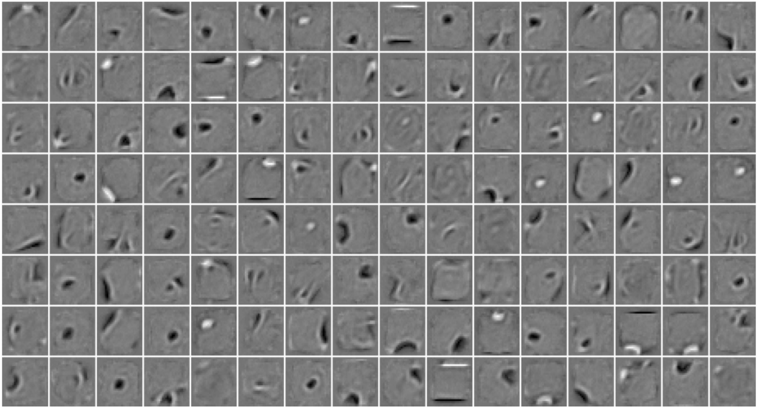
2. Go back to 1 until stopping criteria

Example



Samples from the MNIST digit recognition data set. Here, a black pixel corresponds to an input value of 0 and a white pixel corresponds to 1 (the inputs are scaled between 0 and 1).

Example



The input weights of a random subset of the hidden units. The activation of units of the first hidden layer is obtained by a dot product of such a weight "image" with the input image. In these images, a black pixel corresponds to a weight smaller than 3 and a white pixel to a weight larger than 3, with the different shades of gray corresponding to different weight values uniformly between 3 and 3.

Larochelle, et. al, JMLR2009