Architektury uczenia głębokiego

na przykładzie Deep Belief Network

Wojciech Stokowiec

Marzec, 2016

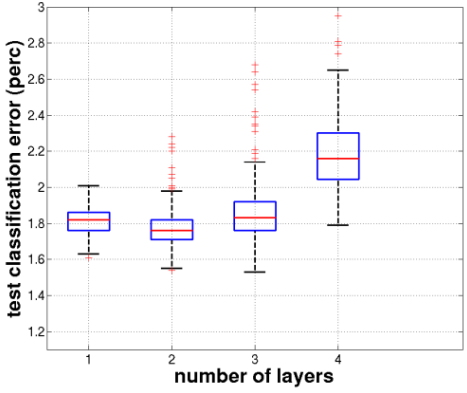
Plan Prezentacji

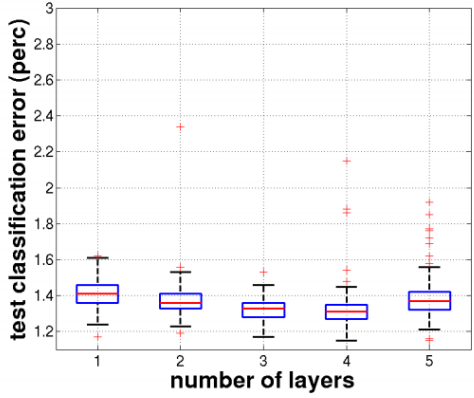
Wprowadzenie

RBM

DBN

Przykład agitujący



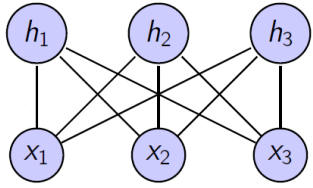


Rysunek: Wykres zaczerpnięty z pracy Erhan D., *"The Difficulty of training deep architectures and the effect of unsupervised pre-training"*

Ograniczona maszyna Boltzmanna (RBM)

Jest to nieskierowany model grafowy opisujący dwie zmienne losowe:

- **oberwowalne** $\mathbf{x} \in \{0,1\}^D$
- **ukryte** $\mathbf{h} \in \{0,1\}^H$



Funkcja energii: $E(\mathbf{x}, \mathbf{h}) = -\mathbf{h}^T \mathbf{W} \mathbf{x} - \mathbf{c}^T \mathbf{x} - \mathbf{b}^T \mathbf{h}$

$$= -\sum_j \sum_k W_{j,k} h_j x_k - \sum_k c_k x_k - \sum_j b_j h_j$$

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

Rozkłady warunkowe (Proste)

$$p(\mathbf{h}|\mathbf{x}) = \prod_j p(h_j|\mathbf{x})$$

$$p(h_j = 1|\mathbf{x}) = \frac{1}{1 + \exp(-(b_j + \mathbf{W}_{j\cdot}\mathbf{x}))}$$

$$= \text{sigm}(b_j + \mathbf{W}_{j\cdot}\mathbf{x})$$

j-ty wiersz macierzy

$$p(\mathbf{x}|\mathbf{h}) = \prod_k p(x_k|\mathbf{h})$$

$$p(x_k = 1|\mathbf{x}) = \frac{1}{1 + \exp(-(c_k + \mathbf{h}^T\mathbf{W}_{\cdot k}))}$$

$$= \text{sigm}(b_j + \mathbf{h}^T\mathbf{W}_{\cdot k})$$

k-ta kolumna macierzy

$p(\mathbf{h}|\mathbf{x})$

$p(\mathbf{h}|\mathbf{x}) = p(\mathbf{x}, \mathbf{h})/p(\mathbf{x})$

$$p(\mathbf{h}|\mathbf{x}) = p(\mathbf{x}, \mathbf{h})/p(\mathbf{x}) = p(\mathbf{x}, \mathbf{h})/\sum_{h'} p(\mathbf{x}, h')$$

$$p(\mathbf{h}|\mathbf{x}) = p(\mathbf{x},\mathbf{h})/p(\mathbf{x}) = p(\mathbf{x},\mathbf{h})/\sum_{h'} p(\mathbf{x},h')$$

$$= \frac{\exp(\mathbf{h}^T\mathbf{W}\mathbf{x} + \mathbf{c}^T\mathbf{x} + \mathbf{b}^T\mathbf{h})/Z}{\sum_{\mathbf{h}'\in\{0,1\}^H} \exp(\mathbf{h}'^T\mathbf{W}\mathbf{x} + \mathbf{c}^T\mathbf{x} + \mathbf{b}^T\mathbf{h}')/Z}$$

$$p(\mathbf{h}|\mathbf{x}) = p(\mathbf{x},\mathbf{h})/p(\mathbf{x}) = p(\mathbf{x},\mathbf{h})/\sum_{h'} p(\mathbf{x},h')$$

$$= \frac{\exp(\mathbf{h}^T\mathbf{W}\mathbf{x} + \mathbf{c}^T\mathbf{x} + \mathbf{b}^T\mathbf{h})/Z}{\sum_{\mathbf{h}'\in\{0,1\}^H} \exp(\mathbf{h}'^T\mathbf{W}\mathbf{x} + \mathbf{c}^T\mathbf{x} + \mathbf{b}^T\mathbf{h}')/Z}$$

$$= \frac{\exp(\sum_j h_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j)}{\sum_{h'_1\in\{0,1\}} \cdots \sum_{h'_H\in\{0,1\}} \exp(\sum_j h'_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j\prime)}$$

$$
\begin{aligned}
p(\mathbf{h}|\mathbf{x}) &= p(\mathbf{x},\mathbf{h})/p(\mathbf{x}) = p(\mathbf{x},\mathbf{h})/\sum_{h'} p(\mathbf{x},h') \\
&= \frac{\exp(\mathbf{h}^T\mathbf{W}\mathbf{x} + \mathbf{c}^T\mathbf{x} + \mathbf{b}^T\mathbf{h})/Z}{\sum_{\mathbf{h}'\in\{0,1\}^H} \exp(\mathbf{h'}^T\mathbf{W}\mathbf{x} + \mathbf{c}^T\mathbf{x} + \mathbf{b}^T\mathbf{h}')/Z} \\
&= \frac{\exp(\sum_j h_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j)}{\sum_{h'_1\in\{0,1\}} \cdots \sum_{h'_H\in\{0,1\}} \exp(\sum_j h'_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j\prime)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j)}{\sum_{h'_1\in\{0,1\}} \cdots \sum_{h'_H\in\{0,1\}} \prod_j \exp(h'_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j\prime)}
\end{aligned}
$$

$$
\begin{aligned}
p(\mathbf{h}|\mathbf{x}) &= p(\mathbf{x},\mathbf{h})/p(\mathbf{x}) = p(\mathbf{x},\mathbf{h})/\sum_{h'} p(\mathbf{x},h') \\
&= \frac{\exp(\mathbf{h}^T\mathbf{W}\mathbf{x} + \mathbf{c}^T\mathbf{x} + \mathbf{b}^T\mathbf{h})/Z}{\sum_{\mathbf{h'}\in\{0,1\}^H} \exp(\mathbf{h'}^T\mathbf{W}\mathbf{x} + \mathbf{c}^T\mathbf{x} + \mathbf{b}^T\mathbf{h'})/Z} \\
&= \frac{\exp(\sum_j h_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j)}{\sum_{h'_1\in\{0,1\}} \cdots \sum_{h'_H\in\{0,1\}} \exp(\sum_j h'_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j\prime)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j)}{\sum_{h'_1\in\{0,1\}} \cdots \sum_{h'_H\in\{0,1\}} \prod_j \exp(h'_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j\prime)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j)}{\left(\sum_{h_1{}'\in\{0,1\}} \exp(h_1{}'\mathbf{W_1}.\mathbf{x} + b_1 h_1{}')\right) \cdots \left(\sum_{h_H{}'\in\{0,1\}} \exp(h_H{}'\mathbf{W_H}.\mathbf{x} + b_H h_H{}')\right)}
\end{aligned}
$$

$$
\begin{aligned}
p(\mathbf{h}|\mathbf{x}) &= p(\mathbf{x},\mathbf{h})/p(\mathbf{x}) = p(\mathbf{x},\mathbf{h})/\sum_{h'} p(\mathbf{x},h') \\
&= \frac{\exp(\mathbf{h}^T\mathbf{W}\mathbf{x} + \mathbf{c}^T\mathbf{x} + \mathbf{b}^T\mathbf{h})/Z}{\sum_{\mathbf{h'}\in\{0,1\}^H} \exp(\mathbf{h'}^T\mathbf{W}\mathbf{x} + \mathbf{c}^T\mathbf{x} + \mathbf{b}^T\mathbf{h'})/Z} \\
&= \frac{\exp(\sum_j h_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j)}{\sum_{h'_1\in\{0,1\}} \cdots \sum_{h'_H\in\{0,1\}} \exp(\sum_j h'_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j\prime)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j)}{\sum_{h'_1\in\{0,1\}} \cdots \sum_{h'_H\in\{0,1\}} \prod_j \exp(h'_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j\prime)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j)}{\left(\sum_{h_1{}'\in\{0,1\}} \exp(h_1{}'\mathbf{W_1}.\mathbf{x} + b_1 h_1{}')\right) \cdots \left(\sum_{h_H{}'\in\{0,1\}} \exp(h_H{}'\mathbf{W_H}.\mathbf{x} + b_H h_H{}')\right)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j)}{\prod_j \left(\sum_{h_j{}'\in\{0,1\}} \exp(h_j{}'\mathbf{W_1}.\mathbf{x} + b_j h_j{}')\right)}
\end{aligned}
$$

$$
\begin{aligned}
p(\mathbf{h}|\mathbf{x}) &= p(\mathbf{x},\mathbf{h})/p(\mathbf{x}) = p(\mathbf{x},\mathbf{h})/\sum_{h'} p(\mathbf{x},h') \\
&= \frac{\exp(\mathbf{h}^T\mathbf{W}\mathbf{x} + \mathbf{c}^T\mathbf{x} + \mathbf{b}^T\mathbf{h})/Z}{\sum_{\mathbf{h}'\in\{0,1\}^H} \exp(\mathbf{h'}^T\mathbf{W}\mathbf{x} + \mathbf{c}^T\mathbf{x} + \mathbf{b}^T\mathbf{h'})/Z} \\
&= \frac{\exp(\sum_j h_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j)}{\sum_{h'_1\in\{0,1\}} \cdots \sum_{h'_H\in\{0,1\}} \exp(\sum_j h'_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j\prime)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j)}{\sum_{h'_1\in\{0,1\}} \cdots \sum_{h'_H\in\{0,1\}} \prod_j \exp(h'_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j\prime)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j)}{\left(\sum_{h_1{}'\in\{0,1\}} \exp(h_1{}'\mathbf{W_1}.\mathbf{x} + b_1 h_1{}')\right) \cdots \left(\sum_{h_H{}'\in\{0,1\}} \exp(h_H{}'\mathbf{W_H}.\mathbf{x} + b_H h_H{}')\right)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j)}{\prod_j \left(\sum_{h_j{}'\in\{0,1\}} \exp(h_j{}'\mathbf{W_1}.\mathbf{x} + b_j h_j{}')\right)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j)}{\prod_j (1 + \exp(b_j + \mathbf{W_j}.\mathbf{x}))}
\end{aligned}
$$

$$
\begin{aligned}
p(\mathbf{h}|\mathbf{x}) &= p(\mathbf{x},\mathbf{h})/p(\mathbf{x}) = p(\mathbf{x},\mathbf{h})/\sum_{h'} p(\mathbf{x},h') \\
&= \frac{\exp(\mathbf{h}^T\mathbf{W}\mathbf{x} + \mathbf{c}^T\mathbf{x} + \mathbf{b}^T\mathbf{h})/Z}{\sum_{\mathbf{h}'\in\{0,1\}^H} \exp(\mathbf{h'}^T\mathbf{W}\mathbf{x} + \mathbf{c}^T\mathbf{x} + \mathbf{b}^T\mathbf{h}')/Z} \\
&= \frac{\exp(\sum_j h_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j)}{\sum_{h'_1\in\{0,1\}} \cdots \sum_{h'_H\in\{0,1\}} \exp(\sum_j h'_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j\prime)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j)}{\sum_{h'_1\in\{0,1\}} \cdots \sum_{h'_H\in\{0,1\}} \prod_j \exp(h'_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j\prime)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j)}{\left(\sum_{h_1{}'\in\{0,1\}} \exp(h_1{}'\mathbf{W_1}.\mathbf{x} + b_1 h_1{}')\right) \cdots \left(\sum_{h_H{}'\in\{0,1\}} \exp(h_H{}'\mathbf{W_H}.\mathbf{x} + b_H h_H{}')\right)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j)}{\prod_j \left(\sum_{h_j{}'\in\{0,1\}} \exp(h_j{}'\mathbf{W_1}.\mathbf{x} + b_j h_j{}')\right)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W_j}\cdot\mathbf{x} + b_j h_j)}{\prod_j (1 + \exp(b_j + \mathbf{W_j}.\mathbf{x}))} \\
&= \prod_j p(h_j|\mathbf{x})
\end{aligned}
$$

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

$$p(h_j = 1|\mathbf{x}) = \frac{\exp(b_j + W_{j\cdot}\mathbf{x})}{1 + \exp(b_j + W_{j\cdot\mathbf{x}})}$$

$$
\begin{aligned}
p(h_j = 1|\mathbf{x}) &= \frac{\exp(b_j + W_{j.}\mathbf{x})}{1 + \exp(b_j + W_{j.\mathbf{x}})} \\
&= \frac{1}{1 + \exp(-b_j - W_{j.}\mathbf{x}}) \\
&= \mathsf{sigm}(b_j + W_{j.\mathbf{x}})
\end{aligned}
$$

Uczenie RBM (trudne)

- W celu znalezienia optymalnych wartości parametów modelu dla danego zbioru posługujemy się metodą największej wiarogodności:

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

- Licząc gradient (np. w celu optymalizacji metodą stochastycznego spadku wzdłuż gradientu) otrzymujemy:

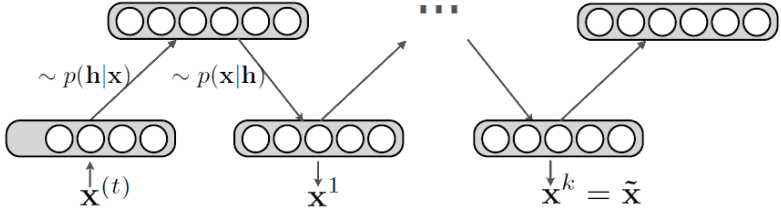
$$\frac{\partial -\log p(\mathbf{x}^{(t)})}{\partial \theta} = \underbrace{\mathbb{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x}^{(t)}, \mathbf{h})}{\partial \theta} \mid \mathbf{x}^{(t)}\right]}_{\text{faza dodatnia}} - \underbrace{\mathbb{E}_{\mathbf{x},\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial \theta}\right]}_{\text{faza ujemna}} \tag{2}$$

trudne do policzenia

Algorytm Contrastive Divergence

Główna idea:

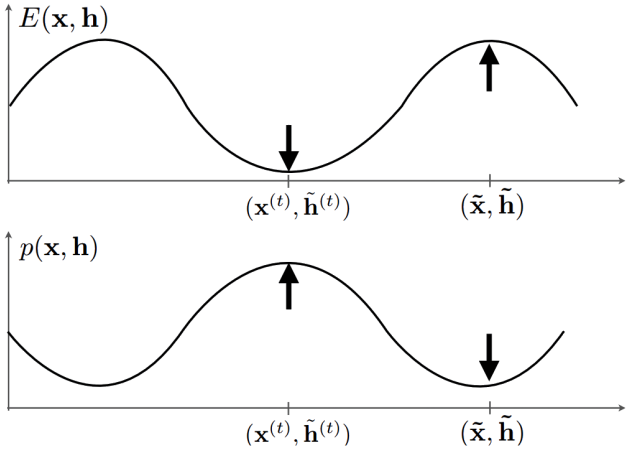
- Zatąpić wartość oczewianą estymatorem punktowym $\tilde{\mathbf{x}}$
- Punkt $\tilde{\mathbf{x}}$ otrzymać stosując próbkowanie Gibbsa
- Próbkować rozpocząć od $\mathbf{x}^{(t)}$



- Jak próbkować z rozkładu $p(\mathbf{h}|x)$?
  - Wylosować $u \sim U[0,1]$
  - $\mathbb{1}_{[u,1]}(p(h_j = 1|\mathbf{x}))$
- okazuje się, że $k = 1$ w praktyce w zupełności wystarcza.

Algorytm Contrastive Divergence - Intuicja

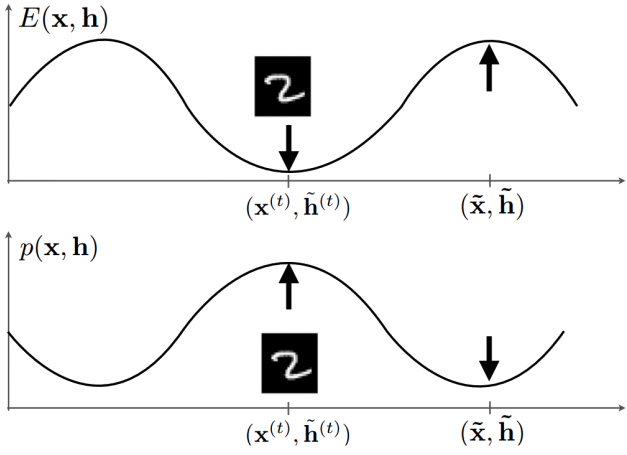
$$\mathbb{E}_{\mathbf{h}}\Big[\frac{\partial E(\mathbf{x}^{(t)},\mathbf{h})}{\partial\theta} \mid \mathbf{x}^{(t)}\Big] \approx \frac{\partial E(\mathbf{x}^{(t)},\tilde{\mathbf{h}}^{(t)})}{\partial\theta} \qquad \mathbb{E}_{\mathbf{x},\mathbf{h}}\Big[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial\theta}\Big] \approx \frac{\partial E(\tilde{\mathbf{x}},\tilde{\mathbf{h}})}{\partial\theta}$$



Rysunek: Wykres za Hugo Larochelle

Algorytm Contrastive Divergence - Intuicja

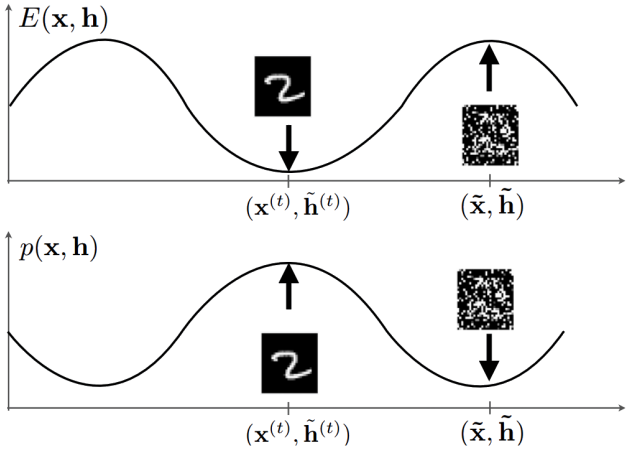
$$\mathbb{E}_{\mathbf{h}}\Big[\frac{\partial E(\mathbf{x}^{(t)},\mathbf{h})}{\partial\theta} \mid \mathbf{x}^{(t)}\Big] \approx \frac{\partial E(\mathbf{x}^{(t)},\tilde{\mathbf{h}}^{(t)})}{\partial\theta} \qquad \mathbb{E}_{\mathbf{x},\mathbf{h}}\Big[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial\theta}\Big] \approx \frac{\partial E(\tilde{\mathbf{x}},\tilde{\mathbf{h}})}{\partial\theta}$$



Rysunek: Wykres za Hugo Larochelle

Algorytm Contrastive Divergence - Intuicja

$$\mathbb{E}_{\mathbf{h}}\Big[\frac{\partial E(\mathbf{x}^{(t)},\mathbf{h})}{\partial\theta} \mid \mathbf{x}^{(t)}\Big] \approx \frac{\partial E(\mathbf{x}^{(t)},\tilde{\mathbf{h}}^{(t)})}{\partial\theta} \qquad \mathbb{E}_{\mathbf{x},\mathbf{h}}\Big[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial\theta}\Big] \approx \frac{\partial E(\tilde{\mathbf{x}},\tilde{\mathbf{h}})}{\partial\theta}$$



Rysunek: Wykres za Hugo Larochelle

Algorytm Contrastive Divergence - Gradienty

Wyprowadźmy $\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial \theta}$ dla $\theta = W_{jk}$

$$\begin{aligned}
\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial W_{jk}} &= \frac{\partial}{\partial W_{jk}}\Big(-\sum_{jk} W_{jk} h_j x_k - \sum_k c_k x_k - \sum_j b_j h_j \Big) \\
&= -\frac{\partial}{\partial W_{jk}} \sum_{jk} W_{jk} h_j x_k \\
&= -h_j x_k
\end{aligned}$$

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

Algorytm Contrastive Divergence - Gradienty

Wyprowadźmy $\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial \theta}$ dla $\theta = W_{jk}$

$$\begin{aligned}\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial W_{jk}} &= \frac{\partial}{\partial W_{jk}}\Big(-\sum_{jk} W_{jk} h_j x_k - \sum_k c_k x_k - \sum_j b_j h_j \Big) \\ &= -\frac{\partial}{\partial W_{jk}} \sum_{jk} W_{jk} h_j x_k \\ &= -h_j x_k\end{aligned}$$

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

iloczyn diadyczny (ang. *outer product*)

Algorytm Contrastive Divergence - Gradienty

Wyprowadźmy $\mathbb{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial \theta} \mid \mathbf{x}\right]$ dla $\theta = W_{jk}$

$$\begin{aligned}\mathbb{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial \theta} \mid \mathbf{x}\right] &= \mathbb{E}_{\mathbf{h}}\left[- h_j x_k \mid \mathbf{x}\right] \\ &= \sum_{h_j \in \{0,1\}} -h_j x_k p(h_j|\mathbf{x}) \\ &= -x_k p(h_j = 1|\mathbf{x})\end{aligned}$$

$$\mathbb{E}_{\mathbf{h}}\left[\nabla_{\mathbf{W}} E(\mathbf{x},\mathbf{h}) \mid \mathbf{x}\right] = - \mathbf{h}(\mathbf{x})\, \mathbf{x}^T, \quad \mathbf{h}(\mathbf{x}) \stackrel{\text{def}}{=} \text{sigm}(\mathbf{b} + \mathbf{W}\mathbf{x})$$

Algorytm Contrastive Divergence - Aktualizacja parametrów

Mając $\mathbf{x}^{(t)}$ oraz $\tilde{\mathbf{x}}$ wagi dla $\theta = \mathbf{W}$ aktualizujemy:

$$
\begin{aligned}
\mathbf{W} &\leftarrow \mathbf{W} - \alpha\Big(- \nabla_{\mathbf{w}} \log p(\mathbf{x}^{(t)})\Big) \\
&\leftarrow \mathbf{W} - \alpha\Big(\mathbb{E}_{\mathbf{h}}\Big[\nabla_{\mathbf{W}} E(\mathbf{x}^{(t)}, \mathbf{h}) \mid \mathbf{x}^{(t)}\Big] - \mathbb{E}_{\mathbf{x},\mathbf{h}}[\nabla_{\mathbf{W}} E(\mathbf{x}, \mathbf{h})]\Big) \\
&\leftarrow \mathbf{W} - \alpha\Big(\mathbb{E}_{\mathbf{h}}\Big[\nabla_{\mathbf{W}} E(\mathbf{x}^{(t)}, \mathbf{h}) \mid \mathbf{x}^{(t)}\Big] - \mathbb{E}_{\mathbf{h}}\Big[\nabla_{\mathbf{W}} E(\tilde{\mathbf{x}}, \mathbf{h} \mid \tilde{\mathbf{x}})\Big]\Big) \\
&\leftarrow \mathbf{W} - \alpha\Big(\mathbf{h}(\mathbf{x}^{(t)})\mathbf{x}^{(t)^T} - \mathbf{h}(\tilde{\mathbf{x}})\tilde{\mathbf{x}}^T\Big)
\end{aligned}
$$

Algorytm Contrastive Divergence - Aktualizacja parametrów

Mając $\mathbf{x}^{(t)}$ oraz $\tilde{\mathbf{x}}$ wagi dla $\theta = \mathbf{W}$ aktualizujemy:

$$
\begin{aligned}
\mathbf{W} &\leftarrow \mathbf{W} - \alpha\Big(- \nabla_{\mathbf{w}} \log p(\mathbf{x}^{(t)})\Big) \\
&\leftarrow \mathbf{W} - \alpha\Big(\mathbb{E}_{\mathbf{h}}\Big[\nabla_{\mathbf{W}} E(\mathbf{x}^{(t)}, \mathbf{h}) \mid \mathbf{x}^{(t)}\Big] - \mathbb{E}_{\mathbf{x},\mathbf{h}}[\nabla_{\mathbf{W}} E(\mathbf{x}, \mathbf{h})]\Big) \\
&\leftarrow \mathbf{W} - \alpha\Big(\mathbb{E}_{\mathbf{h}}\Big[\nabla_{\mathbf{W}} E(\mathbf{x}^{(t)}, \mathbf{h}) \mid \mathbf{x}^{(t)}] - \mathbb{E}_{\mathbf{h}}\Big[\nabla_{\mathbf{W}} E(\tilde{\mathbf{x}}, \mathbf{h} \mid \tilde{\mathbf{x}})\Big]\Big) \\
&\leftarrow \mathbf{W} - \alpha\Big(\mathbf{h}(\mathbf{x}^{(t)})\mathbf{x}^{(t)^T} - \mathbf{h}(\tilde{\mathbf{x}})\tilde{\mathbf{x}}^T\Big)
\end{aligned}
$$

Uwaga

Wypdowadzenie reguł aktualizacji parametów $\mathbf{b}$, $\mathbf{c}$ jest analogiczne. Pozostawiamy jako zadanie domowe dla słuchaczy ;-).

Algorytm Contrastive Divergence - Pseudokod

1. Dla każdego elementu $\mathbf{x}^{(t)}$ ze zbioru treningowego:
   - wyznacz estymator punktowy $\tilde{\mathbf{x}}$ stosując k krokowe próbowanie Gibbsa, rozpoczynając z punktu $\mathbf{x}^{(t)}$
   - zaktualizuj parametry modelu:

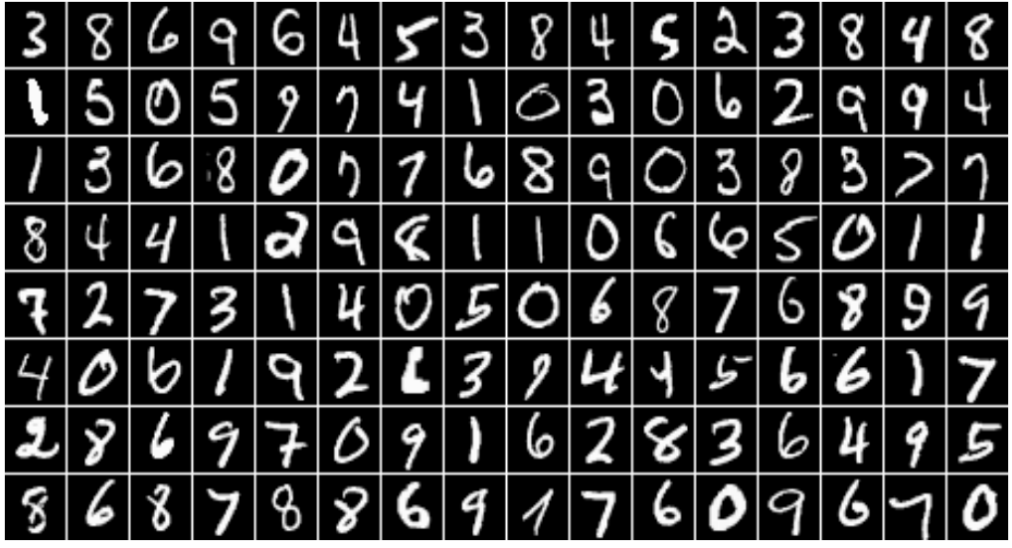
$$\mathbf{W} \leftarrow \mathbf{W} - \alpha\Big(\mathbf{h}(\mathbf{x}^{(t)})\mathbf{x}^{(t)^T} - \mathbf{h}(\tilde{\mathbf{x}})\tilde{\mathbf{x}}^T\Big)$$

$$\mathbf{b} \leftarrow \mathbf{b} - \alpha\Big(\mathbf{h}(\mathbf{x}^{(t)}) - \mathbf{h}(\tilde{\mathbf{x}})\Big)$$

$$\mathbf{c} \leftarrow \mathbf{c} - \alpha\Big(\mathbf{x}^{(t)} - \tilde{\mathbf{x}}\Big)$$

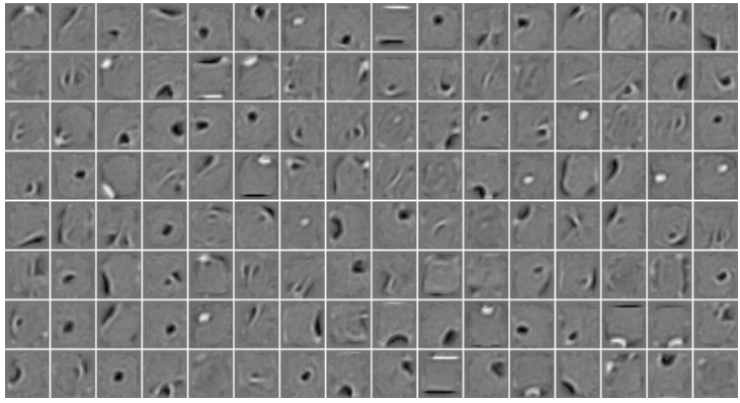
2. Wróc do punktu 1 i powtarzaj dopóki nie zpełnione zostaną kryteria zbierzności

Obowiązkowy MNIST



Rysunek: Przekładowe elementy. Czarny piksel odpowiada wartości 0, biały wartości 1

Obowiązkowy MNIST



Rysunek: Podgląd wag odpowidających losowo wybranym ukrytym wektorom uzyskanych podczas uczenia RBM na zbiorze MNIST $\mathbf{h}$. Czerne piksele odpowiadają wagom < -3, białe > 3, a szare to przeskalowane wartości wag z przedziału $[-3, 3]$

Deep Belief Networks

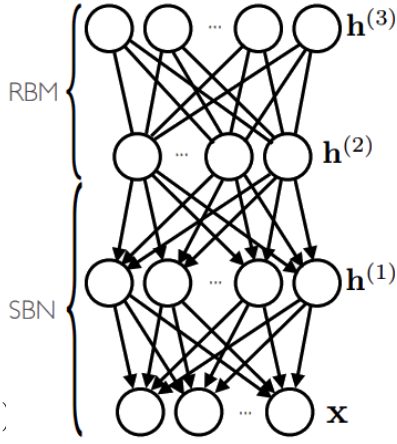
Deep Belief Networks

- Jest to model generatywny, który zawiera w sobie zarówno skierowane, jak i nieskierowane połączenia pomiędzy zmiennymi losowymi.
- Rozkład dwóch górynch warstw $p(\mathbf{h}^{(2)}, \mathbf{h}^{(2)})$ jest wyznaczony przez RBM
- Pozostałe warstwy tworzą sieć bayesowską

$$p(h_j^{(1)} = 1|\mathbf{h}^{(2)}) = \text{sigm}(\mathbf{b}^{(1)} + \mathbf{W}^{(2)^T}\mathbf{h}^{(2)})$$

$$p(x_i^{(1)} = 1|\mathbf{h}^{(1)}) = \text{sigm}(\mathbf{b}^{(0)} + \mathbf{W}^{(1)^T}\mathbf{h}^{(1)})$$

- DBN **nie jest** siecią neuronową typu *feed-forward*



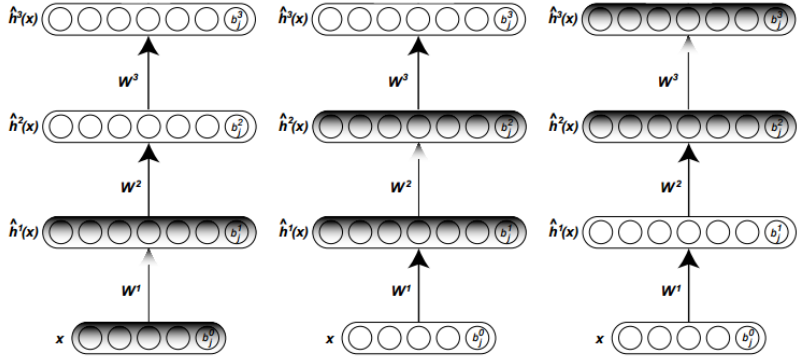
Deep Belief Networks

- Pełny rozkład DBN przedstawia się następująco:

$$p(\mathbf{x}, \mathbf{h}^{(1)}, \mathbf{h}^{(2)}, \mathbf{h}^{(3)}) = p(\mathbf{h}^{(2)}, \mathbf{h}^{(3)}) p(\mathbf{h}^{(1)} \mid \mathbf{h}^{(2)}) p(\mathbf{x} \mid \mathbf{h}^{(1)}), \tag{3}$$

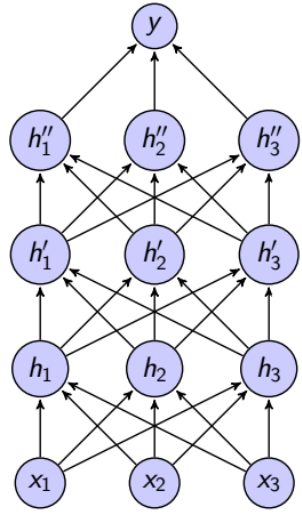
- gdzie
  - $p(\mathbf{h}^{(2)}, \mathbf{h}^{(3)}) = \exp({\mathbf{h}^{(2)}}^T \mathbf{W} \mathbf{h}^{(3)} + {\mathbf{b}^{(2)}}^T \mathbf{h}^{(2)} + {\mathbf{b}^{(3)}}^T \mathbf{h}^{(3)})/Z$
  - $p(\mathbf{h}^{(1)} \mid \mathbf{h}^{(2)}) = \prod_j p(h_j^{(1)} | \mathbf{h}^{(2)})$
  - $p(\mathbf{x} \mid \mathbf{h}^{(1)}) = \prod_i p(x_i, | \mathbf{h}^{(1)})$

DBN - Uczenie

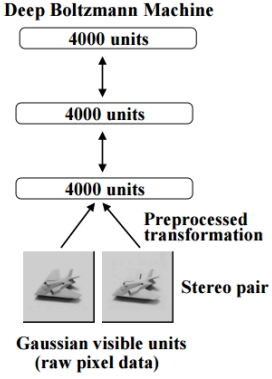


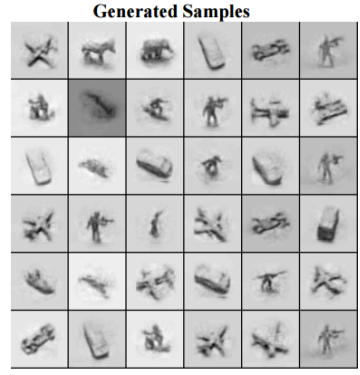
Deep Belief Networks

- Stos kilku RBM może zostać użyty do zainicjalizowania głębokiej sieci neuronowej, np. klasycznego MLP.
- Następnym krokiem jest etap douczania (ang. *fine-tuning*), np za pomocą propagacji wstecznej.
- Nie znam pracy, gdzie *fine-tuning* przeprowadzany byłby metodami optymalizacji globalnej.



Co potrafi głęboki model generatywny





Rysunek: Przykład z parcy Salakhutdinova i Hintona, *Deep Boltzmann Machines*. Po prawej stronie przedstawiono wyniki próbkowania z modelu uczonego na 20 tys. obrazków z $k = 10$ tys.

RBM i DBN podsumowanie

1. Zachłnanne uczenie warstwa po warstwie było pomysłem odpowiedzialnym za renesans głębokich architektur.
2. Uczenie wstępne, tzw. *pre-training* jest podejściem nienadzorowanym. Głównym celem jest optymalizacja względem funkcji wiarogorności na danych, a nie predykcja klasy dla danej obserwacji. (Na początku wyznaczmy $p(x)$, dopiero później $p(y, |x))$
3. Naiwna metoda uczenia RBM jest kosztowna. Rozwiązanie: algorytm **contrastive divergence**.
4. Ograniczone maszyny Boltzmana możemy albo, składać jedna na drugą otrzymując Deep Belief Network (głęboki generatywny probabilistyczny model grafowy), albo użyć parametrów do zainicjalizowania głebokiej sieci neuronowej