10-707: Deep Learning, Fall 2017

Homework 2

Lecturer: Russ Salakhutdinov *Name: Yuan Liu(yuanl4)*

1 Problem 1

- The input image is a 2-D matrix X, the size is $d \times d$.
- Convolutional Layer: the parameter of feature map $i \in \{1, 2\}$ is W^i. They are two 2-D matrixes, the size of each matrix is $k \times k$.

  The output of the convolutional layer is two 2-D matrixes $Y^i, i \in \{1, 2\}$, the size of each matrix is $\frac{d-k+1}{s} \times \frac{d-k+1}{s}$

  $$Y^i_{m,n} = \sum_{p=1}^{k} \sum_{q=1}^{k} X_{(m-1)s+p,(n-1)s+q} W^i_{p,q}$$

  Then we can get:

  $$\frac{\partial Y^i_{m,n}}{\partial W^i_{p,q}} = X_{(m-1)s+p,(n-1)s+q}$$

  The derivation input of the convolutional layer is two 2-D matrix $\frac{\partial Loss}{\partial Y^i}, i \in \{1, 2\}$, the size of each matrix is $\frac{d-k+1}{s} \times \frac{d-k+1}{s}$. Then we can get:

  $$\frac{\partial Loss}{\partial W^i_{p,q}} = \sum_{m,n=1}^{\frac{d-k+1}{s}} \frac{\partial Loss}{\partial Y^i}_{m,n} \frac{\partial Y^i_{m,n}}{\partial W^i_{p,q}} = \sum_{m,n=1}^{\frac{d-k+1}{s}} \frac{\partial Loss}{\partial Y^i}_{m,n} X_{(m-1)s+p,(n-1)s+q}$$

- Pooling Layer: The output of pooling layer is two 2-D matrixes $P^i, i \in \{1, 2\}$, the size of each matrix is $(\frac{d-k+1}{s} - p + 1) \times (\frac{d-k+1}{s} - p + 1)$.

  $$P^k_{m,n} = \max_{i,j \in \{1,\ldots,p\}} Y^k_{m+i-1,n+j-1}$$

  The derivation input of the pooling layer is two 2-D matrix $\frac{\partial Loss}{\partial P^i}, i \in \{1, 2\}$, the size of each matrix is $(\frac{d-k+1}{s} - p + 1) \times (\frac{d-k+1}{s} - p + 1)$. Then we gan get the derivation output is:

  $$\frac{\partial Loss}{\partial Y^k}_{m,n} = \sum_{i,j=1}^{p} \frac{\partial Loss}{\partial P^k}_{m+i-1,n+j-1} \mathbb{1}(i, j = \arg\max_{i,j \in \{1,\ldots,p\}} Y^k_{m+i-1,n+j-1})$$

- Flatten Layer, which convert the two 2-dimension matrixes into a vector. The length of this vector is $2(\frac{d-k+1}{s} - p + 1)^2$.

  $$F_i = P^k_{m,n}$$

  $$k = \lceil \frac{i}{(\frac{d-k+1}{s} - p + 1)^2} \rceil$$

$$m = \lceil \frac{i - (\frac{d-k+1}{s} - p + 1)^2(k-1)}{\frac{d-k+1}{s} - p + 1} \rceil$$

$$n = i - (\frac{d-k+1}{s} - p + 1)^2(k-1) - (\lceil \frac{i - (\frac{d-k+1}{s} - p + 1)^2(k-1)}{\frac{d-k+1}{s} - p + 1} \rceil - 1)(\frac{d-k+1}{s} - p + 1)$$

$$\frac{\partial Loss}{\partial P^k_{m,n}} = \frac{\partial Loss}{\partial F_{(k-1)*(\frac{d-k+1}{s}-p+1)^2+(m-1)*(\frac{d-k+1}{s}-p+1)+n}}$$

- Softmax Layer

2 Problem 2

Because the model is a directed graphical model, so it is a directed acyclic graph. Then we can find an order $\{I_i\}_{i=1}^K$, that $pa_{I_i} \subset \{x_{I_j}\}_{j>i}$. For simplicity, we can just assume that $\{x_i\}_{i=1}^K$ satisfies this order, which means $pa_{x_i} \subset \{x_j\}_{j>i}$.

$$\int p(x)dx = \int \prod_{k=1}^{K} p(x_k|pa_k)dx_1...dx_k = \int p(x_1|pa_1)dx_1 \int \prod_{k=2}^{K} p(x_k|pa_k)dx_2...dx_k$$

We can do this calculation, because $x_1 \notin \cap_{k=2}^{K} pa_k$. We also know $\int p(x_1|pa_1)dx_1 = 1$. Then we can know:

$$\int p(x)dx = \int \prod_{k=1}^{K} p(x_k|pa_k)dx_1...dx_k = \int \prod_{k=2}^{K} p(x_k|pa_k)dx_2...dx_k$$

By the same way, we can finally get

$$\int p(x)dx = \int p(x_K|pa_K)dx_K$$

Because x_K is the last one in the node list $x_1, ..., x_K$, so $pa_K = \emptyset$, $\int p(x_K|pa_K)dx_K = \int p(x_k)dx_K = 1$. Finally we get:

$$\int p(x)dx = 1$$

3 Problem 3

$$p_\theta(v,h) = \frac{1}{Z} exp(v^T W h + v^T b + h^T a)$$

$$
\begin{aligned}
p_\theta(h|v) &= \frac{p(v,h)}{p(h)} = \frac{\frac{1}{Z}exp(v^TWh + v^Tb + h^Ta)}{\sum_h \frac{1}{Z}exp(v^TWh + v^Tb + h^Ta)} \\
&= \frac{\exp(v^TWh + h^Ta)}{\sum_h exp(v^TWh + h^Ta)} \\
&= \frac{\prod_{i=1}^P exp(h_i(W^Tv+a)_i)}{\sum_{h_1} exp(h_1(W^Tv+a)_1) \times \sum_{h_2} exp(h_2(W^Tv+a)_2) \times ... \times \sum_{h_P} exp(h_P(W^Tv+a)_P)} \\
&= \prod_{i=1}^P \frac{exp(h_i(W^Tv+a)_i)}{\sum_{h_i\in\{0,1\}} exp(h_i(W^Tv+a)_i)} \\
&= \prod_{j=1}^P \sigma(h_i(W^Tv+a)_i)
\end{aligned}
$$

$$
\begin{aligned}
p_\theta(h_j=1|v) &= \sum_{h_j=1,h_{i\neq j}\in\{0,1\}} p_\theta(h|v) = \sum_{h_j=1,h_{i\neq j}\in\{0,1\}} \prod_{j=1}^P \sigma(h_i(W^Tv+a)_i) \\
&= \sigma((W^Tv+a)_j) \sum_{h_{i\neq j}\in\{0,1\}} \prod_{j\in\{1,..,P\}-\{j\}} \sigma(h_i(W^Tv+a)_i) \\
&= \sigma((W^Tv+a)_j) \prod_{j\in\{1,..,P\}-\{j\}} \sum_{h_{i\neq j}\in\{0,1\}} \sigma(h_i(W^Tv+a)_i) \\
&= \sigma((W^Tv+a)_j)
\end{aligned}
$$

By this formula we can know $p_\theta(h_j|v) = \sigma(h_j(W^Tv+1)_j)$. Thus

$$p_\theta(v,h) = \prod_{j=1}^P p_\theta(h_j|v)$$

4 Problem 4

4.1

$$
\begin{aligned}
E(x_m=1, x_{i\neq m}, y) &= h\sum_{i\neq m} x_i + h - \beta \sum_{i\neq m, j\neq m} x_ix_j - \beta \sum_{i\in local(m)} x_i - \eta \sum_{i\neq m} x_iy_i - \eta y_m \\
E(x_m=-1, x_{i\neq m}, y) &= h\sum_{i\neq m} x_i - h - \beta \sum_{i\neq m, j\neq m} x_ix_j + \beta \sum_{i\in local(m)} x_i - \eta \sum_{i\neq m} x_iy_i + \eta y_m
\end{aligned}
$$

Then we can get:

$$E(x_m=1, x_{i\neq m}, y) - E(x_m=-1, x_{i\neq m}, y) = 2h - 2\beta \sum_{i\in local(m)} x_i - 2\eta y_m$$

So the difference in the value of energy depends only on quantities that are local to x_m in the graph.

4.2

If $\beta = h = 0$, we can get: $E(x, y) = -\eta \sum_i x_i y_i$. If we want to minimize the energy, we need to maximize $\sum_i x_i y_i$. Because $x_i \in \{-1, +1\}, y_i \in \{-1, +1\}$, so the maximum of $x_i y_i$ is 1, which can be got by $x_i = y_i$. So the most probable configuration of the latent variables is given by $x_i = y_i$ for all i.