Homework1 for Deep Learning

Shijie Wang 2016010539

Block One:

(i)

$y_i = BN_{\gamma,\beta}(x_i) = \gamma\hat{x}_i + \beta$

$\frac{\partial y_i}{\partial \gamma} = \hat{x}_i,\ \frac{\partial y_i}{\partial \beta} = 1$

(ii)

y is the output of dropout layer and x is the input.

$\frac{\partial y_j}{\partial x_j} = M_j = \begin{cases} 0, & r_j < p, \\ 1/(1-p), & r_j \geq p \end{cases}$

(iii)

a is the output of softmax function and z is the input.

$$a = \frac{1}{\Sigma_{j=1}^{k} exp(z_j)} \begin{bmatrix} exp(z_1) \\ exp(z_2) \\ \vdots \\ exp(z_k) \end{bmatrix}$$

if $j \neq i$:

$$\begin{aligned} \frac{\partial a_j}{\partial z_i} &= \frac{\partial}{\partial z_i}(\frac{exp(z_j)}{C_1 + exp(z_i)}) \\ &= -\frac{exp(z_j) \cdot exp(z_i)}{(C_1 + exp(z_i))^2} \\ &= -a_i \cdot a_j \end{aligned}$$

else if $j = i$:

$$\begin{aligned} \frac{\partial a_j}{\partial z_i} &= \frac{\partial}{\partial z_i}(\frac{exp(z_i)}{C_1 + exp(z_i)}) \\ &= \frac{\partial}{\partial z_i}(1 - \frac{C_1}{C_1 + exp(z_i)}) \\ &= \frac{C_1 \cdot exp(z_i)}{(C_1 + exp(z_i))^2} \\ &= a_i(1 - a_i) \end{aligned}$$

so:

$$\frac{\partial a_j}{\partial z_i} = \begin{cases} a_i(1-a_i), & i=j, \\ -a_i a_j, & i \neq j \end{cases}$$

Block Two:

(i)

batch samples $(\mathbf{x}, y_a, y_b)$

For Task A:

After passing fc-layer FC_{1A} with $ReLU$, we get a_{1a}:

$$a_{1a} = ReLU(\theta_{1a} x + b_{1a})$$

After passing dropout layer DP_{1A}, we get a_{dp}:

$$\begin{aligned} a_{dp(i)} &= \begin{cases} 0, & r_i < p, \\ a_{1a(i)}/(1-p), & r_i \geq p \end{cases} \\ &= a_{1a} \odot M \end{aligned}$$

After passing fc-layer FC_{2A}, we get $\hat{y}_a = a_{2a}$:

$$\hat{y}_a = a_{2a} = \theta_{2a} a_{dp} + b_{2a}$$

For Task B:

After passing fc-layer FC_{1B} with $ReLU$, we get a_{1b}:

$$a_{1b} = ReLU(\theta_{1b} x + b_{1b})$$

After passing BN-layer BN_{1B}, we get a_{bn} :

$$\begin{aligned} \mu_{\mathcal{B}} &\leftarrow \frac{1}{m} \sum_{i=1}^{m} a_{1b(i)} \\ \sigma_{\mathcal{B}}^2 &\leftarrow \frac{1}{m} \sum_{i=1}^{m} (a_{1b(i)} - \mu_{\mathcal{B}})^2 \\ \hat{a}_{1b(i)} &\leftarrow \frac{a_{1b(i)} - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \\ a_{bn(i)} &\leftarrow \gamma \hat{a}_{1b(i)} + \beta \end{aligned}$$

After adding the output of task A, we get a_{add} :

$$a_{add} = a_{2a} + a_{bn}$$

After passing fc-layer FC_{2B} with $Softmax$, we get $\hat{y}_b = a_{2b}$:

$$\hat{y}_b = a_{2b} = Softmax(\theta_{2b} a_{add} + b_{2b})$$

(ii)

Denote L_a, L_b, L the task A loss, task B loss and overall loss,

$$L(x, y_a, y_b; \theta) = L_a + L_b = \frac{1}{m} \sum_{i=1}^{m} \left[\frac{1}{2} \|(\hat{y}_{ai} - y_{ai})\|_2^2 - \sum_{j=1}^{n_{yb}} y_{bi}^j \log\left(\hat{y}_{bi}^j\right) \right]$$

where $\hat{y}_{ai}, \hat{y}_{bi}$ denotes the predictions of task A,B for the i-th sample .

TASK B:

The gradient with respect to z_{2b}^k, the k-th element for FC_{2B} before softmax:

$$\frac{\partial L}{\partial \hat{y}_{bi}^j} = -\frac{1}{m} \frac{y_{bi}^j}{\hat{y}_{bi}^j}$$

$$\begin{aligned}
\frac{\partial L}{\partial z_{2bi}^k} &= \sum_{j=1}^{n_{yb}} \frac{\partial L}{\partial \hat{y}_{bi}^j} \frac{\partial \hat{y}_{bi}^j}{\partial z_{2bi}^k} \\
&= (-\frac{1}{m}) \sum_{j=1}^{n_{yb}} \frac{y_{bi}^j}{\hat{y}_{bi}^j} \frac{\partial \hat{y}_{bi}^j}{\partial z_{2bi}^k} \\
&= \frac{1}{m} \sum_{j \neq k} \frac{y_{bi}^j}{\hat{y}_{bi}^j} \hat{y}_{bi}^k \hat{y}_{bi}^j - \frac{1}{m} \frac{y_{bi}^k}{\hat{y}_{bi}^k} (1 - \hat{y}_{bi}^k) \hat{y}_{bi}^k \\
&= \frac{1}{m} (\hat{y}_{bi}^k - y_{bi}^k)
\end{aligned}$$

so we get the gradient with respect to z_{2b} and θ_{2b}

$$\begin{aligned}
\frac{\partial L}{\partial z_{2b}} &= \frac{1}{m} (\hat{y}_b - y_b) \\
\frac{\partial L}{\partial \theta_{2b}} &= \frac{\partial L}{\partial z_{2b}} \frac{\partial z_{2b}}{\partial \theta_{2b}} \\
&= \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_{bi} - y_{bi}) a_{addi}^\top
\end{aligned}$$

The gradient with respect to γ and β in layer BN_{1B}:

$$\frac{\partial L}{\partial a_{ddi}} = \frac{\partial L}{\partial a_{bni}}$$
$$= \frac{1}{m}(\hat{y}_{bi} - y_{bi})\theta_{2b}$$

$$\frac{\partial L}{\partial \gamma} = \frac{\partial L}{\partial add}\frac{\partial add}{\partial \gamma} = \frac{\partial L}{\partial add}\frac{\partial a_{bn}}{\partial \gamma}$$
$$= \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_{bi} - y_{bi})\theta_{2b}\hat{a}_{1bi}$$

$$\frac{\partial L}{\partial \beta} = \frac{\partial L}{\partial add}\frac{\partial add}{\partial \beta} = \frac{\partial L}{\partial add}\frac{\partial a_{bn}}{\partial \beta}$$
$$= \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_{bi} - y_{bi})\theta_{2b}$$

The gradient with respect to a_{1b}:

$$\frac{\partial L}{\partial a_{1bi}} = \frac{\partial L}{\partial a_{bni}}\frac{\partial a_{bni}}{\partial a_{1bi}} + \frac{\partial L}{\partial \sigma_{\mathcal{B}}^2}\frac{\partial \sigma_{\mathcal{B}}^2}{\partial a_{1bi}} + \frac{\partial L}{\partial \mu_{\mathcal{B}}}\frac{\partial \mu_{\mathcal{B}}}{\partial a_{1bi}}$$
$$= \frac{\partial L}{\partial a_{bni}} \cdot \frac{\gamma}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} + \frac{\partial L}{\partial \sigma_{\mathcal{B}}^2} \cdot \frac{2(a_{1bi} - \mu_{\mathcal{B}})}{m} + \frac{\partial L}{\partial \mu_{\mathcal{B}}} \cdot \frac{1}{m}$$

where:

$$\frac{\partial L}{\partial \sigma_{\mathcal{B}}^2} = \sum_{i=1}^{m}\frac{\partial L}{\partial a_{bni}}\gamma\left(a_{1bi} - \mu_{\mathcal{B}}\right) \cdot \frac{-1}{2}\left(\sigma_{\mathcal{B}}^2 + \epsilon\right)^{-3/2}$$
$$\frac{\partial L}{\partial \mu_{\mathcal{B}}} = \left(\sum_{i=1}^{m}\frac{\partial L}{\partial a_{bni}} \cdot \frac{-\gamma}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}\right) - \frac{\partial L}{\partial \sigma_{\mathcal{B}}^2} \cdot \frac{2\sum_{i=1}^{m}\left(a_{bni} - \mu_{\mathcal{B}}\right)}{m}$$

The gradient with respect to a_{1b}:

$$\frac{\partial L}{\partial z_{1bi}} = \frac{\partial L}{\partial a_{1bi}}\frac{\partial a_{1bi}}{\partial z_{1bi}}$$
$$= \frac{\partial L}{\partial a_{1bi}} \odot sgn(z_{1bi})$$

$$\frac{\partial L}{\partial \theta_{1b}} = \sum_{i=1}^{m}\frac{\partial L}{\partial z_{1bi}}\frac{\partial z_{1bi}}{\partial \theta_{1b}}$$
$$= \sum_{i=1}^{m}\frac{\partial L}{\partial z_{1bi}}x_i$$

TASK A:

The gradient with respect to a_{2a}:

$$\frac{\partial L}{\partial a_{2ai}} = \frac{1}{m}(\hat{y}_{ai} - y_{ai}) + \frac{\partial L}{\partial a_{addi}}$$

The gradient with respect to θ_{2a}:

$$\frac{\partial L}{\partial \theta_{2a}} = \sum_{i=1}^{m} \frac{\partial L}{\partial a_{2ai}} a_{dpi}$$

The gradient with respect to a_{dp}:

$$\frac{\partial L}{\partial a_{dp}} = \theta_{2a} \frac{\partial L}{\partial a_{2ai}}$$

The gradient with respect to a_{1a}:

$$\begin{aligned}\frac{\partial L}{\partial a_{1ai}} &= \frac{\partial L}{\partial a_{dpi}} \frac{\partial a_{dpi}}{\partial a_{1ai}} \\ &= \frac{\partial L}{\partial a_{dpi}} \odot M\end{aligned}$$

The gradient with respect to z_{1a}:

$$\frac{\partial L}{\partial z_{1ai}} = \frac{\partial L}{\partial a_{dpi}} \odot M \odot sgn(z_{1ai})$$

The gradient with respect to θ_{1a}:

$$\begin{aligned}\frac{\partial L}{\partial \theta_{1a}} &= \sum_{i=1}^{m} \frac{\partial L}{\partial z_{1ai}} \frac{\partial z_{1ai}}{\partial \theta_{1a}} \\ &= \sum_{i=1}^{m} \frac{\partial L}{\partial z_{1ai}} x_i\end{aligned}$$

That all, thank!