深度学习第一次作业

刘旭鑫
2021214058
软硕 211

日期：2022 年 3 月 26 日

1 Block One

1.1 Gradient of BatchNormalization Layer

$$\begin{cases} \frac{\partial y_i}{\partial \gamma} = \hat{x}_i = \frac{x_i - \mu_\beta}{\sqrt{\sigma_\beta^2 + \epsilon}}, \text{其中}\mu_\beta = \frac{1}{m}\sum_{i=1}^{m} x_i, \sigma_\beta^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_\beta)^2 \\ \frac{\partial y_i}{\partial \beta} = \mathbf{1} \end{cases}$$

1.2 Gradient of Dropout Layer

从题目中的描述可以知道，对于一个概率 p，Dropout 的过程可以转换成一个概率矩阵 $\mathbf{M}$ 对输入的点积，即

$$y = \mathbf{M} \odot x$$

其中，

$$\mathbf{M}_j = \begin{cases} 0, & r_j < p, \\ 1/(1-p), & r_j \geq p \end{cases} \text{where } 1 \leq j \leq \text{x's size}$$

因此梯度 $\frac{\partial y}{\partial x} = \mathbf{M}$

1.3 Gradient of Softmax Function

对于 Softmax 函数，有

$$y_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}, \text{其中}n\text{ 为输出的向量长度}$$

考虑 $\frac{\partial y_i}{\partial x_i}$，根据求导的除法法则有

$$\begin{aligned} \frac{\partial y_i}{\partial x_i} &= \frac{e^{x_i}\sum_{j=1}^{n} e^{x_j} - e^{2x_i}}{(\sum_{j=1}^{n} e^{x_j})^2} \\ &= \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}(1 - \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}) \\ &= y_i(1 - y_i) \end{aligned}$$

再考虑 $\frac{\partial y_i}{\partial x_j}(i \neq j)$，求导有

$$\begin{aligned}\frac{\partial y_i}{\partial x_j} &= -\frac{e^{x_i}e^{x_j}}{(\sum_{k=1}^{n}e^{x_k})^2} \\ &= -\frac{e^{x_i}}{\sum_{k=1}^{n}e^{x_k}} \times \frac{e^{x_j}}{\sum_{k=1}^{n}e^{x_k}} \\ &= -y_i y_j\end{aligned}$$

2 Block Two

2.1 Feed-forward

先考虑 $\hat{\mathbf{y}}_A$：

- FC_{1A} 的输出 $\mathbf{z}_{1A} = sin(\theta_{1A}\mathbf{x} + \mathbf{b}_{1A})$
- 假设 DP 层对应的概率矩阵为 $\mathbf{M}$，那么其输出为 $\mathbf{z}_{DP} = \mathbf{M} \odot \mathbf{z}_{1A}$
- FC_{2A} 的输出 $\mathbf{z}_{2A} = \theta_{2A}\mathbf{z}_{DP} + \mathbf{b}_{2A}$

因此有

$$\hat{\mathbf{y}}_A = \mathbf{z}_{2A} = \theta_{2A}\mathbf{M} \odot sin(\theta_{1A}\mathbf{x} + \mathbf{b}_{1A}) + \mathbf{b}_{2A}$$

考虑 $\hat{\mathbf{y}}_B$:

- FC_{1B} 的输出 $\mathbf{z}_{1B} = \theta_{1B}\mathbf{x} + \mathbf{b}_{1B}$
- BN 层的输出为 $\mathbf{z}_{BN} = \mathbf{N} \odot (\mathbf{z}_{1B} - \mu + \mathbf{b}_{1B})$，其中 $\mathbf{N}$ 为符号向量，$N_i = 1$ if $x_i > 0$ else 0，$\mu = \frac{1}{m}\sum_{i=1}^{m}\mathbf{z}_{1B}^i$
- FC_{2B} 的输入为 $\mathbf{x}_{2B} = \mathbf{z}_{BN} + \mathbf{y}_A$，输出为 $\mathbf{z}_{2B} = Softmax(\theta_{2B}\mathbf{x}_{2B} + \mathbf{b}_{2B})$

所以有 $\hat{\mathbf{y}}_B = Softmax\left(\theta_{2B}(\mathbf{N} \odot ((\theta_{1B}\mathbf{x} + \mathbf{b}_{1B}) - \mu + \mathbf{b}_{1B}) + \theta_{2A}\mathbf{M} \odot sin(\theta_{1A}\mathbf{x} + \mathbf{b}_{1A}) + \mathbf{b}_{2A}) + \mathbf{b}_{2B}\right)$

2.2 Backpropagation

损失函数为

$$\mathcal{L} = \frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{2}||\hat{\mathbf{y}}_A^i - \mathbf{y}_A^i||_2^2 - \sum_{k=1}^{b}\mathbf{y}_{B,k}^i log\hat{\mathbf{y}}_{B,k}^i\right]$$

损失函数对 $\hat{\mathbf{y}_B}^i$ 的导数为

$$\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}_{B,k}^i} = -\frac{1}{m}\mathbf{y}_{B,k}^i\frac{1}{\hat{\mathbf{y}_{B,k}^i}} \tag{1}$$

根据 1.3 节，$\hat{\mathbf{y}}_B^i$ 对 θ_{2B} 的导数为

$$
\begin{aligned}
\frac{\partial \hat{\mathbf{y}}_{B,k}^i}{\partial \theta_{2B}} &= \frac{\partial \hat{\mathbf{y}}_{B,k}^i}{\partial \mathbf{z}_{2B}^i}\frac{\partial \mathbf{z}_{2B}^i}{\partial \theta_{2B}} \\
&= \begin{bmatrix} -\hat{\mathbf{y}}_{B,k}^i\hat{\mathbf{y}}_{B,1}^i \\ \cdots \\ -\hat{\mathbf{y}}_{B,k}^i\hat{\mathbf{y}}_{B,k-1}^i \\ \hat{\mathbf{y}}_{B,k}^i(1-\hat{\mathbf{y}}_{B,k}^i) \\ -\hat{\mathbf{y}}_{B,k}^i\hat{\mathbf{y}}_{B,k+1}^i \\ \cdots \\ -\hat{\mathbf{y}}_{B,k}^i\hat{\mathbf{y}}_{B,b}^i \end{bmatrix} (\mathbf{x}_{2B}^i)^T
\end{aligned} \tag{2}
$$

再由链式法则可以求出

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \theta_{2B}} &= \sum_{i=1}^{m}\sum_{k=1}^{b}\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}_{B,k}^i}\frac{\partial \hat{\mathbf{y}}_{B,k}^i}{\partial \theta_{2B}} \\
&= \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{b}\hat{\mathbf{y}}_{B,k}^i \begin{bmatrix} \hat{\mathbf{y}}_{B,1}^i \\ \cdots \\ \hat{\mathbf{y}}_{B,k-1}^i \\ \hat{\mathbf{y}}_{B,k}^i - 1 \\ \hat{\mathbf{y}}_{B,k+1}^i \\ \cdots \\ \hat{\mathbf{y}}_{B,b}^i \end{bmatrix} (\mathbf{x}_{2B}^i)^T \\
&= \frac{1}{m}\sum_{i=1}^{m} \begin{bmatrix} \hat{\mathbf{y}}_{B,i}^i \sum_{k=1}^{b}\hat{\mathbf{y}}_{B,k}^i - \hat{\mathbf{y}}_{B,1}^i \\ \cdots \\ \hat{\mathbf{y}}_{B,i}^i \sum_{k=1}^{b}\hat{\mathbf{y}}_{B,k}^i - \hat{\mathbf{y}}_{B,b}^i \end{bmatrix} (\mathbf{x}_{2B}^i)^T \\
&= \frac{1}{m}\sum_{i=1}^{m}(\hat{\mathbf{y}}_B^i - \mathbf{y}_B^i)(\mathbf{x}_{2B}^i)^T
\end{aligned} \tag{3}
$$

$\frac{\partial \mathcal{L}}{\partial \mathbf{b}_{2B}}$ 的计算过程与上面类似，结果为 $\frac{1}{m}\sum_{i=1}^{m}(\hat{\mathbf{y}}_B^i - \mathbf{y}_B^i)$

由 1.1 到 1.3 节的推导，以及链式法则得到

$$
\frac{\partial \mathcal{L}}{\partial \mathbf{x}_{2B}} = \frac{\partial L}{\partial \mathbf{z}_{BN}^i} = \frac{1}{m}\sum_{i=1}^{m}(\theta_{2B})^T(\hat{\mathbf{y}}_B^i - \hat{\mathbf{y}}_B^i) \tag{4}
$$

$$
\frac{\partial \mathcal{L}}{\partial \mathbf{z}_{1B}^i} = \frac{1}{m}(1-\frac{1}{m})\sum_{i=1}^{m}(\theta_{2B})^T(\hat{\mathbf{y}}_B^i - \hat{\mathbf{y}}_B^i) \odot sgn(\mathbf{H}_{BN}^i) \tag{5}
$$

其中，$\mathbf{H}_{BN}^i = \mathbf{z}_{1B}^i - \mu + \mathbf{b}_{1B}$

$$
\frac{\partial \mathcal{L}}{\partial \theta_{1B}} = \frac{1}{m}(1-\frac{1}{m})\sum_{i=1}^{m}(\theta_{2B})^T(\hat{\mathbf{y}}_B^i - \mathbf{y}_B^i) \odot sgn(\mathbf{H}_{BN}^i)(\mathbf{x}^i)^T \tag{6}
$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}_{1B}} = \frac{\partial \mathcal{L}}{\partial \mathbf{H}_{BN}^i} = \frac{1}{m}\sum_{i=1}^{m}(\theta_{2B})^T(\hat{\mathbf{y}}_B^i - \mathbf{y}_B^i) \odot sgn(\mathbf{z}_{BN}^i) \tag{7}$$

下面推导 Task A 路径的梯度。

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \theta_{2A}} &= \frac{\partial \mathcal{L}_{\text{taskA}}}{\partial \theta_{2A}} + \frac{\partial \mathcal{L}_{\text{taskB}}}{\partial \theta_{2A}} \\
&= \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}_A^i}\frac{\partial \hat{\mathbf{y}}_A^i}{\partial \theta_{2A}} + \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}_B^i}\frac{\partial \hat{\mathbf{y}}_B^i}{\partial \theta_{2A}} \\
&= \frac{1}{m}\sum_{i=1}^{m}(\hat{\mathbf{y}}_A^i - \mathbf{y}_A^i)(\mathbf{z}_{DP}^i)^T + \frac{\partial \mathcal{L}}{\partial \mathbf{x}_{2B}}\frac{\partial \mathbf{x}_{2B}}{\partial \hat{\mathbf{y}}_A^i}\frac{\partial \hat{\mathbf{y}}_A^i}{\partial \theta_{2A}} \\
&= \frac{1}{m}\sum_{i=1}^{m}[(\hat{\mathbf{y}}_A^i - \mathbf{y}_A^i) + (\theta_{2B})^T(\hat{\mathbf{y}}_B^i - \mathbf{y}_B^i)](\mathbf{z}_{DP}^i)^T
\end{aligned} \tag{8}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{b}_{2A}} &= \frac{\partial \mathcal{L}_{\text{taskA}}}{\partial \mathbf{b}_{2A}} + \frac{\partial \mathcal{L}_{\text{taskB}}}{\partial \mathbf{b}_{2A}} \\
&= \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}_A^i}\frac{\partial \hat{\mathbf{y}}_A^i}{\partial \mathbf{b}_{2A}} + \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}_B^i}\frac{\partial \hat{\mathbf{y}}_B^i}{\partial \mathbf{b}_{2A}} \\
&= \frac{1}{m}\sum_{i=1}^{m}(\hat{\mathbf{y}}_A^i - \mathbf{y}_A^i) + \frac{\partial \mathcal{L}}{\partial \mathbf{x}_{2B}}\frac{\partial \mathbf{x}_{2B}}{\partial \hat{\mathbf{y}}_A^i}\frac{\partial \hat{\mathbf{y}}_A^i}{\partial \mathbf{b}_{2A}} \\
&= \frac{1}{m}\sum_{i=1}^{m}[(\hat{\mathbf{y}}_A^i - \mathbf{y}_A^i) + (\theta_{2B})^T(\hat{\mathbf{y}}_B^i - \mathbf{y}_B^i)]
\end{aligned} \tag{9}$$

通过链式法则以及 1.1 到 1.3 节的结论，有

$$\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}_A^i} = \frac{1}{m}\sum_{i=1}^{m}[(\hat{\mathbf{y}}_A^i - \mathbf{y}_A^i) + \theta_{2B}^T(\hat{\mathbf{y}}_B^i - \mathbf{y}_B^i)] \tag{10}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{z}_{DP}^i} = \frac{1}{m}\sum_{i=1}^{m}\theta_{2A}^T[(\hat{\mathbf{y}}_A^i - \mathbf{y}_A^i) + \theta_{2B}^T(\hat{\mathbf{y}}_B^i - \mathbf{y}_B^i)] \tag{11}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{z}_{1A}^i} = \frac{1}{m}\sum_{i=1}^{m}\theta_{2A}^T[(\hat{\mathbf{y}}_A^i - \mathbf{y}_A^i) + \theta_{2B}^T(\hat{\mathbf{y}}_B^i - \mathbf{y}_B^i)] \odot \mathbf{M} \tag{12}$$

$$\frac{\partial \mathcal{L}}{\partial \theta_{1A}} = \frac{1}{m}\sum_{i=1}^{m}\theta_{2A}^T\left([(\hat{\mathbf{y}}_A^i - \mathbf{y}_A^i) + \theta_{2B}^T(\hat{\mathbf{y}}_B^i - \mathbf{y}_B^i)] \odot \mathbf{M} \odot cos(\mathbf{H}_{1A}^i)\right)(\mathbf{x}^i)^T \tag{13}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}_{1A}} = \frac{1}{m}\sum_{i=1}^{m}\theta_{2A}^T[(\hat{\mathbf{y}}_A^i - \mathbf{y}_A^i) + \theta_{2B}^T(\hat{\mathbf{y}}_B^i - \mathbf{y}_B^i)] \odot \mathbf{M} \odot cos(\mathbf{H}_{1A}^i) \tag{14}$$