深度学习
2020 年春季学期

作业 1

徐荣琛 2019214518 软件学院 2020 年 3 月 16 日

1.1. **Block One: Gradients of some basic layers (30 points)**

(i) Given a BatchNorm layer, please calculate the gradients of the output $y_i = \mathbf{BN}_{\gamma,\beta}(x_i)$ with respect to the parameters of γ, β shown in Figure 4. **(10 points)**

(ii) Given a dropout layer, please calculate the gradients of **the output of a dropout layer** with respect to **the input of a dropout layer**. **(10 points)**

(iii) Given a Softmax function, please calculate the gradients of **the output of a Softmax function** with respect to **the input of a Softmax function**. **(10 points)**

解. (i) $\frac{\partial y_i}{\partial \gamma} = \hat{x}_i$, $\frac{\partial y_i}{\partial \beta} = 1$

(ii) 设 dropout 层的输入为 $\mathcal{DI}$（n_{1a} 维），其中第 i 个值为 $\mathcal{DI}_i$，设 dropout 层的输出为 $\mathcal{DO}$（n_{1a} 维），其中第 j 个值为 $\mathcal{DO}_j$，则：

$$\frac{\partial \mathcal{DO}_j}{\partial \mathcal{DI}_i} = \begin{cases} 0 & i \neq j \\ 0 & i = j \wedge r_i < p \\ 1/(1-p) & i = j \wedge r_i \geq p \end{cases}$$

(iii) 设 Softmax 函数的输入为 $\mathcal{SI}$（n_{yb} 维），其中第 i 个值为 $\mathcal{SI}_i$，设 Softmax 函数的输出为 $\mathcal{SO}$（n_{yb} 维），其中第 j 个值为 $\mathcal{SO}_j$，则：

- 若 $i = j$：

$$\begin{aligned} \frac{\partial \mathcal{SO}_j}{\partial \mathcal{SI}_i} = \frac{\partial \frac{e^{\mathcal{SI}_i}}{\sum_k e^{\mathcal{SI}_k}}}{\partial \mathcal{SI}_i} &= \frac{e^{\mathcal{SI}_i} \cdot (\sum_k e^{\mathcal{SI}_k}) - e^{\mathcal{SI}_i} \cdot e^{\mathcal{SI}_i}}{(\sum_k e^{\mathcal{SI}_k})^2} \\ &= \mathcal{SO}_j \cdot (1 - \mathcal{SO}_j) \end{aligned}$$

- 若 $i \neq j$：

$$\frac{\partial \mathcal{SO}_j}{\partial \mathcal{SI}_i} = \frac{\partial \frac{e^{\mathcal{SI}_j}}{\sum_k e^{\mathcal{SI}_k}}}{\partial \mathcal{SI}_i} = \frac{-e^{\mathcal{SI}_j} \cdot e^{\mathcal{SI}_i}}{(\sum_k e^{\mathcal{SI}_k})^2} = -\mathcal{SO}_i \cdot \mathcal{SO}_j$$

1.2. **Block Two: Feed-forward and back-propagation of the multi-task network (30 points)**

(i) Finish the detailed **feed-forward computations** of a batch samples $(\boldsymbol{x}, y_a, y_b)$ during a training iteration, coming with final predictions $(\hat{y}_a, \hat{y}_b)$ of Task A, Task B. **(10 points)**

(ii) Use the back-propagation algorithm we have learned in class and give **the gradients of the overall loss function with respect to the parameters at each layer** corresponding to a batch of samples. **(20 points)**

解. (i) 由于是全连接层，FC_{1A} 层的输入：

$$Z^{FC_{1A}} = \theta_{1a}\boldsymbol{x} + b_{1a}$$

对应 FC_{1A} 层的激活：

$$a^{FC_{1A}} = \mathbf{ReLU}(Z^{FC_{1A}})$$

由于是全连接层，但受到 DP_{1A} 层 dropout 的影响，且 FC_{2A} 层没有激活函数，FC_{2A} 层的输入即预测的 $\hat{y}_a$：

$$\hat{y}_a = Z^{FC_{2A}} = \theta_{2a}\left(a^{FC_{1A}} \odot \mathbf{M}\right) + b_{2a}$$

其中 $\mathbf{M}$ 是 random mask 向量，运算 $\odot$ 是向量的逐元素乘法；
类似 FC_{1A} 层，FC_{1B} 层的输入：

$$Z^{FC_{1B}} = \theta_{1b}\boldsymbol{x} + b_{1b}$$

对应 FC_{1B} 层的激活：

$$a^{FC_{1B}} = \mathbf{ReLU}(Z^{FC_{1B}})$$

经过 **BN** 层以及随后的逐元素加运算 $\oplus$，FC_{2B} 层的输入：

$$Z^{FC_{2B}} = \theta_{2b}\left(\mathbf{BN}_{\gamma,\beta}(a^{FC_{1B}}) \oplus \hat{y}_a\right) + b_{2b}$$

其中 γ, β 是 Batch Normalize 的参数；
对应 FC_{2B} 层的激活即预测的 $\hat{y}_b$：

$$\hat{y}_b = a^{FC_{2B}} = \mathbf{Softmax}(Z^{FC_{2B}})$$

(ii) 损失函数 L：

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

FC_{2B} 的残余 $\delta_{FC_{2B}}$：

$$\delta_{FC_{2B}} = \frac{\partial L}{\partial z^{FC_{2B}}} = \frac{\partial L}{\partial a^{FC_{2B}}} \cdot \frac{\partial a^{FC_{2B}}}{\partial z^{FC_{2B}}}$$

前者

$$\frac{\partial L}{\partial a^{FC_{2B}}} = \frac{\partial L}{\partial \hat{y}_b} = -\frac{1}{m}\sum_{i=1}^{m}\frac{y_{bi}}{\hat{y}_{bi}}$$

后者由 1.1(iii) 求得，代入：

$$\begin{aligned}
\delta_{FC_{2B}} &= -\sum_{i=1}^{m}\sum_{j=1}^{n_{yb}}\frac{y_{bi}^{j}}{\hat{y}_{bi}^{j}} \cdot \frac{\partial a^{FC_{2B}}}{\partial z^{FC_{2B}}} \\
&= \frac{1}{m}\sum_{i=1}^{m}\left((-\frac{y_{bi}^{k}}{\hat{y}_{bi}^{k}})\hat{y}_{bi}^{k}(1-\hat{y}_{bi}^{k}) + \sum_{j\neq k}\frac{y_{bi}^{j}}{\hat{y}_{bi}^{j}}\left(\hat{y}_{bi}^{j}\right)^2\right) \\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\hat{y}_{bi} - y_{bi}\right)
\end{aligned}$$

对于 θ_{2b} 的梯度：

$$\begin{aligned}
\frac{\partial L}{\partial \theta_{2b}} &= \delta_{FC_{2B}} \cdot \left(\frac{\partial z^{FC_{2B}}}{\partial \theta_{2b}}\right)^T \\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\hat{y}_{bi} - y_{bi}\right)\left(\mathbf{BN}_{\gamma,\beta}(a^{FC_{1B}}) \oplus \hat{y}_a\right)^T
\end{aligned}$$

对于 Batch Normalize 中 γ 的梯度：

$$\begin{aligned}
\frac{\partial L}{\partial \gamma} &= \left(\frac{\partial \mathbf{BN}_{\gamma,\beta}(a^{FC_{1B}})}{\partial \gamma}\right)^T \cdot \left(\frac{\partial z^{FC_{2B}}}{\partial \mathbf{BN}_{\gamma,\beta}(a^{FC_{1B}})}\right)^T \cdot \frac{\partial L}{\partial z^{FC_{2B}}} \\
&= \left(\hat{a}^{FC_{1B}}\right)^T \left(\theta_{2b}\right)^T \frac{1}{m}\sum_{i=1}^{m}\left(\hat{y}_{bi} - y_{bi}\right)
\end{aligned}$$

其中，$\hat{a}^{FC_{1B}}$ 是对 $a^{FC_{1B}}$ 的正则化。

对于 Batch Normalize 中 β 的梯度：

$$\begin{aligned}
\frac{\partial L}{\partial \beta} &= \left(\frac{\partial \mathbf{BN}_{\gamma,\beta}(a^{FC_{1B}})}{\partial \beta}\right)^T \cdot \left(\frac{\partial z^{FC_{2B}}}{\partial \mathbf{BN}_{\gamma,\beta}(a^{FC_{1B}})}\right)^T \cdot \frac{\partial L}{\partial z^{FC_{2B}}} \\
&= \sum \left(\theta_{2b}\right)^T \frac{1}{m}\sum_{i=1}^{m}\left(\hat{y}_{bi} - y_{bi}\right)
\end{aligned}$$

FC_{1B} 的残余 $\delta_{FC_{1B}}$：

$$\begin{aligned}\delta_{FC_{1B}} &= \frac{\partial L}{\partial z^{FC_{1B}}} = \left(\frac{\partial z^{FC_{2B}}}{\partial a^{FC_{1B}}}\right)^T \cdot \delta_{FC_{2B}} \cdot \frac{\partial a^{FC_{1B}}}{\partial z^{FC_{1B}}} \\ &= \left(\theta_{2b}\mathbf{BN}'_{\gamma,\beta}(a^{FC_{1B}})\right)^T \delta_{FC_{2B}} \odot \mathbf{ReLU}'(z^{FC_{1B}})\end{aligned}$$

其中，$\frac{\partial \mathbf{BN}_j}{\partial x_i}$ 满足：

$$\begin{aligned}&\frac{\partial \mathbf{BN}_j}{\partial x_i} = \frac{\partial \frac{x_j-\mu}{\sqrt{\sigma^2+\epsilon}}}{\partial x_i} \\ &= \begin{cases} -\frac{1}{m}\left(\sigma^2+\epsilon\right)^{-1/2} - \frac{1}{m}\left(\sigma^2+\epsilon\right)^{-3/2}\left(x_i-\mu\right)\left(x_j-\mu\right) & i \neq j \\ \left(1-\frac{1}{m}\right)\left(\sigma^2+\epsilon\right)^{-1/2} - \frac{1}{m}\left(\sigma^2+\epsilon\right)^{-3/2}\left(x_i-\mu\right)\left(x_j-\mu\right) & i = j \end{cases}\end{aligned}$$

对于 θ_{1b} 的梯度：

$$\begin{aligned}\frac{\partial L}{\partial \theta_{1b}} &= \delta_{FC_{1B}} \cdot \left(\frac{\partial z^{FC_{1B}}}{\partial \theta_{1b}}\right)^T \\ &= \delta_{FC_{1B}} \boldsymbol{x}^T\end{aligned}$$

FC_{2A} 的残余 $\delta_{FC_{2A}}$：

$$\begin{aligned}\delta_{FC_{2A}} &= \frac{\partial L}{\partial z^{FC_{2A}}} = \frac{\partial L}{\partial \hat{y}_a} \\ &= \frac{1}{m}\sum_{i=1}^{m}\left(\hat{y}_a - y_a\right) - \frac{\partial y_{bi}^j \log\left(\widehat{y}_{bi}^j\right)}{\partial z^{FC_{2B}}} \cdot \frac{\partial z^{FC_{2B}}}{\partial \hat{y}_a} \\ &= \frac{1}{m}\sum_{i=1}^{m}\Big(\left(\hat{y}_a - y_a\right) + \left(\theta_{2b}\right)^T\left(\hat{y}_{bi} - y_{bi}\right)\Big)\end{aligned}$$

对于 θ_{2a} 的梯度：

$$\begin{aligned}\frac{\partial L}{\partial \theta_{2a}} &= \delta_{FC_{2A}} \cdot \left(\frac{\partial z^{FC_{2A}}}{\partial \theta_{2a}}\right)^T \\ &= \left(\frac{1}{m}\sum_{i=1}^{m}\left(\hat{y}_a - y_a\right) + \left(\theta_{2b}\right)^T\left(\hat{y}_{bi} - y_{bi}\right)\right)\boldsymbol{x}^T\end{aligned}$$

FC_{1A} 的残余 $\delta_{FC_{1A}}$：

$$\begin{aligned}\delta_{FC_{1A}} &= \frac{\partial L}{\partial z^{FC_{1A}}} = \left(\frac{\partial z^{FC_{2A}}}{\partial a^{FC_{1A}}}\right)^T \cdot \delta_{FC_{2A}} \cdot \frac{\partial a^{FC_{1A}}}{\partial z^{FC_{1A}}} \\ &= \left(\theta_{2a}\right)^T \delta_{FC_{2A}} \odot \mathbf{M} \odot \mathbf{ReLU}'(z^{FC_{1A}})\end{aligned}$$

其中，$\odot$ 是逐元素乘法。

对于 θ_{1a} 的梯度：

$$
\begin{aligned}
\frac{\partial L}{\partial \theta_{1a}} &= \delta_{FC_{1A}} \cdot \left(\frac{\partial z^{FC_{1A}}}{\partial \theta_{1a}}\right)^T \\
&= \left((\theta_{2a})^T \delta_{FC_{2A}} \odot \mathbf{M} \odot \mathbf{ReLU}'(z^{FC_{1A}})\right) \boldsymbol{x}^T
\end{aligned}
$$