深度学习 HW1

[redacted]

2022 年 10 月 15 日

1 BP

1.1

记 γ，$\mathbf{x}$，β 形状相同，不失一般性，不妨考虑 $\mathbf{x}$ 是一维列向量的情景。

$$\mathbf{y} = \gamma\hat{\mathbf{x}} + \beta$$

$$\begin{aligned}
\frac{\partial \mathbf{y}}{\partial \gamma} &= \mathrm{diag}(\hat{\mathbf{x}}) \\
\frac{\partial y_i}{\partial \gamma} &= x_i \mathbf{e}_i \\
\frac{\partial y_i}{\partial \beta} &= \mathbb{I} \\
\frac{\partial y_i}{\partial \beta} &= 1
\end{aligned}$$

1.2

记 Softmax 函数为，给定向量 $\mathbf{x} = [x_1, ..., x_n]^\top$，有

$$\begin{aligned}
\mathtt{softmax}(x_i) &= \frac{\exp(x_i)}{\sum_{j=1}^n \exp(x_j)} \\
\mathtt{softmax}(\mathbf{x}) &= [\mathtt{softmax}(x_1), ..., \mathtt{softmax}(x_n)]^\top
\end{aligned}$$

那么，记 $\mathbf{y}$ = softmax($\mathbf{x}$)，令 $i \neq j$

$$\begin{aligned}
\frac{\partial y_i}{\partial x_j} &= \frac{-\exp(x_i)\exp(x_j)}{\sum_{k=1}^n \exp(x_k)} = -\mathtt{softmax}(x_i)\mathtt{softmax}(x_j) \\
\frac{\partial y_i}{\partial x_i} &= \frac{\exp(x_i)[(\sum_{k=1}^n \exp(x_k)) - \exp(x_i)]}{\sum_{k=1}^n \exp(x_k)} = \mathtt{softmax}(x_i)(1 - \mathtt{softmax}(x_i))
\end{aligned}$$

1.3

$$\mathbf{x}_{1A} = \theta_{1A}\mathbf{x} + \mathbf{b}_{1A}$$

$$\begin{aligned}\mathbf{x}_{DP} &= \mathbf{M} \odot \texttt{sin}(\mathbf{x}_{1A}) \\ \mathbf{x}_{2A} &= \theta_{2A}\mathbf{x}_{DP} + \mathbf{b}_{2A} \\ \hat{\mathbf{y}}_A &= \mathbf{x}_{2A}\end{aligned}$$

$$\begin{aligned}\mathbf{x}_{1B} &= \theta_{1B}\mathbf{x} \\ \mu &= \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_{1B}^{i} \\ \mathbf{x}_{BN} &= \mathbf{x}_{1B} - \mu + \mathbf{b}_{1B} \\ \mathbf{x}_C &= \texttt{ReLU}(\mathbf{x}_{BN}) \oplus \mathbf{x}_{2A} \\ \mathbf{x}_{2B} &= \theta_{2B}(\mathbf{x}_C) + \mathbf{b}_{2B} \\ \hat{\mathbf{y}}_B &= \texttt{softmax}(\mathbf{x}_{2B})\end{aligned}$$

1.4

从简单的开始

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

$$\begin{aligned}\texttt{softmax}(\mathbf{x}) &= \frac{\texttt{exp}(\mathbf{x})}{\sum \texttt{exp}(\mathbf{x})} \\ \frac{\partial \texttt{softmax}(\mathbf{x})}{\partial \mathbf{x}} &= \frac{\mathbf{diag}(\texttt{exp}(\mathbf{x}))}{\sum \texttt{exp}(\mathbf{x})} - \frac{\texttt{exp}(\mathbf{x})\texttt{exp}(\mathbf{x})^\top}{(\sum \texttt{exp}(\mathbf{x}))^2} = \mathbf{diag}(\texttt{softmax}(\mathbf{x})) - \texttt{softmax}(\mathbf{x})\texttt{softmax}(\mathbf{x})^\top\end{aligned}$$

记 $y_B = \texttt{CrossEntropy}(\hat{\mathbf{y}}_B, \mathbf{y}_B)$，有

$$\begin{aligned}\texttt{CrossEntropy}(\hat{\mathbf{y}}_B, \mathbf{y}_B) &= \sum(\mathbf{y}_B \odot \texttt{log}(\hat{\mathbf{y}}_B)) \\ \frac{\partial \texttt{CrossEntropy}(\hat{\mathbf{y}}_B, \mathbf{y}_B)}{\partial \hat{\mathbf{y}}_B} &= \mathbf{y}_B \oslash \hat{\mathbf{y}}_B\end{aligned}$$

其中，$\oslash$ 为 Hadamard division，即两个相同形状矩阵逐元素除法。
由课上知识，第1个全连接层（输入 a_k^{l-1} ，输出 y_k^l ）的反向传播有

$$\frac{\partial E}{\partial w_{kj}^l} = \delta_j^l a_k^{l-1}$$

这里，我的记号 w_{jk}^l 为连接（a_i^{l-1} 与 y_k^l ）的权重，排列成矩阵有

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{m\times n}} = \begin{pmatrix} \frac{\partial E}{\partial w_{11}^l} & \cdots & \frac{\partial E}{\partial w_{1n}^l} \\ \vdots & \ddots & \vdots \\ \frac{\partial E}{\partial w_{m1}^l} & \cdots & \frac{\partial E}{\partial w_{mn}^l} \end{pmatrix} = \begin{pmatrix} \delta_1^l \\ \vdots \\ \delta_m^l \end{pmatrix} \begin{pmatrix} a_1^{l-1} & \cdots & a_n^{l-1} \end{pmatrix} = \frac{\partial E}{\partial \mathbf{y}}\mathbf{a}^{(l-1)\top}$$

同时，残差在 l 层节点 j 处积累的 l+1 层残差，传播满足以下关系

$$\delta_j^l = f'(u_j^l)\sum_{k=1}^{n'}\delta_k^{l+1}w_{jk}^{l+1}$$

写成矩阵

$$\frac{\partial\mathcal{L}}{\partial\mathbf{y}^l} = f'(\mathbf{y}^l)\odot(W^{l+1\top}\frac{\partial\mathcal{L}}{\partial\mathbf{y}^{l+1}})$$

下面观察

$$\theta_{2B}\to\mathbf{x}_{2B}\to\hat{\mathbf{y}}_B\to y_B$$

先把残差从 y_B 传播到 $\mathbf{x}_{2B}$

$$\frac{\partial\mathcal{L}}{\partial\hat{\mathbf{y}}_B^i} = -\frac{1}{m}\mathbf{y}_B^i\oslash\hat{\mathbf{y}}_B^i$$

以下推导省略 batch index i，除非特殊注明（BN 推导）。

$$\frac{\partial\mathcal{L}}{\partial\mathbf{x}_{2B}^i} = -\frac{1}{m}(\mathbf{y}_B^i-\hat{\mathbf{y}}_B^i\mathbf{y}_B^i\oslash\mathtt{exp}(\mathbf{x}_{2B}^i)\mathtt{exp}(\mathbf{x}_{2B}^i))$$

注意到 $\hat{\mathbf{y}}_B^i\mathbf{y}_B^i\oslash\mathtt{exp}(\mathbf{x}_{2B}^i)=\frac{1}{\sum\mathtt{exp}(\mathbf{x}_{2B}^i)}$

$$\frac{\partial\mathcal{L}}{\partial\mathbf{x}_{2B}^i} = \frac{1}{m}(\hat{\mathbf{y}}_B^i-\mathbf{y}_B^i)$$

那么

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\theta_{2B}} &= \frac{1}{m}\sum_{i=1}^{m}(\hat{\mathbf{y}}_B^i-\mathbf{y}_B^i)\mathbf{x}_C^{i\top}\\ \frac{\partial\mathcal{L}}{\partial\mathbf{b}_{2B}} &= \frac{1}{m}\sum_{i=1}^{m}(\hat{\mathbf{y}}_B^i-\mathbf{y}_B^i)\end{aligned}$$

对于网络 2A

$$\theta_{1B}\to\mathbf{x}_{1B}\to\mathbf{x}_{BN}\to\mathbf{x}_C\to\mathbf{x}_{2B}\to\hat{\mathbf{y}}_B\to y_B$$

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\mathbf{x}_C^i} &= \theta_{2B}^\top\frac{\partial\mathcal{L}}{\partial\mathbf{x}_{2B}^i} = \frac{1}{m}\theta_{2B}^\top(\hat{\mathbf{y}}_B^i-\mathbf{y}_B^i)\\ \frac{\partial\mathcal{L}}{\partial\mathbf{x}_{BN}^i} &= \mathtt{ReLU}'(\mathbf{x}_{BN}^i)\frac{\partial\mathcal{L}}{\partial\mathbf{x}_C^i}\end{aligned}$$

研究 BN 的误差传递，由链式法则，记一个 batch BN 的输出为 $\mathbf{y^1}, \mathbf{y^2}, ..., \mathbf{y^m}$ ，输入为 $\mathbf{x^1}, \mathbf{x^2}, ..., \mathbf{x^m}$ ，满足

$$
\begin{aligned}
\vec{\mu} &= \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}^i \\
\mathbf{y^i} &= \mathbf{x}^i - \vec{\mu} \\
\frac{\partial\mathcal{L}}{\partial x_q^i} &= \sum_{k=1}^{m}\frac{\partial\mathcal{L}}{\partial y_k^i}\frac{\partial y_k^i}{\partial x_q^i} = \frac{\partial\mathcal{L}}{\partial y_q^i}\frac{\partial y_q^i}{\partial x_q^i} = (1-\frac{1}{m})\frac{\partial\mathcal{L}}{\partial y_q^i} \\
\frac{\partial\mathcal{L}}{\partial \mathbf{x}^i} &= (1-\frac{1}{m})\frac{\partial\mathcal{L}}{\partial \mathbf{y}^i}
\end{aligned}
$$

代回，得到

$$
\begin{aligned}
\frac{\partial\mathcal{L}}{\partial \mathbf{x}_{1B}^i} &= (1-\frac{1}{m})\frac{\partial\mathcal{L}}{\partial \mathbf{x}_{BN}^i} \\
\frac{\partial\mathcal{L}}{\partial \theta_{1B}} &= \sum_{i=1}^{m}\frac{\partial\mathcal{L}}{\partial \mathbf{x}_{1B}^i}\mathbf{x}^{i\top}
\end{aligned}
$$

回头算 2A，观察 θ_{2A} 到 $\mathcal{L}$ 的传播链条，共有两个梯度来源，一个是 $\frac{\partial\mathcal{L}}{\partial \mathbf{y}_A}$ ，一个是 $\frac{\partial\mathcal{L}}{\partial \mathbf{x}_C}$。

$$
\begin{aligned}
\frac{\partial\mathcal{L}}{\partial \mathbf{y}_A^i} &= \frac{1}{m}2(\hat{\mathbf{y}}_A^i - \mathbf{y}_A^i) \\
\frac{\partial\mathcal{L}}{\partial \mathbf{x}_C^i} &= \frac{1}{m}\theta_{2B}^\top(\mathbf{y}_B^i - \hat{\mathbf{y}}_B^i)
\end{aligned}
$$

如果把 $\oplus$ 看做一层网络，那么 loss 对输出节点的梯度等于输入节点的梯度，两者相等。研究 $\frac{\partial\mathcal{L}}{\partial \mathbf{x}_{2A}}$ ，由链式法则，应该去两者之和，便有：

$$
\frac{\partial\mathcal{L}}{\partial \mathbf{x}_{2A}^i} = \frac{1}{m}2(\hat{\mathbf{y}}_A^i - \mathbf{y}_A^i) + \frac{1}{m}\theta_{2B}^\top(\mathbf{y}_B^i - \hat{\mathbf{y}}_B^i)
$$

再运用上文结论：

$$
\begin{aligned}
\frac{\partial\mathcal{L}}{\partial \theta_{2A}} &= \sum_{i=1}^{m}\frac{\partial\mathcal{L}}{\partial \mathbf{x}_{2A}^i}\mathbf{x}_{DP}^{i\ \top} \\
\frac{\partial\mathcal{L}}{\partial \mathbf{b}_{2A}} &= \sum_{i=1}^{m}\frac{\partial\mathcal{L}}{\partial \mathbf{x}_{2A}^i}\mathbf{x}_{DP}^i
\end{aligned}
$$

最后算 1A，有了 $\frac{\partial\mathcal{L}}{\partial \mathbf{x}_{2A}^i}$，剩下的就好算了很多：

$$
\frac{\partial\mathcal{L}}{\partial \mathbf{x}_{DP}^i} = \theta_{2A}^\top\frac{\partial\mathcal{L}}{\partial \mathbf{x}_{2A}^i}
$$

记 $\mathbf{x}_s = \mathtt{sin}(\mathbf{x}_{1A})$，考虑 DP 层：

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial x_{si}} &= \sum_j \frac{\partial \mathcal{L}}{\partial x_{DPj}} \frac{\partial x_{DPj}}{\partial x_{si}} = \frac{\partial \mathcal{L}}{\partial x_{DPi}} \frac{\partial x_{DPi}}{\partial x_{si}} \\
\frac{\partial \mathcal{L}}{\partial \mathbf{x}_s^i} &= \mathbf{M} \odot \frac{\partial \mathcal{L}}{\partial \mathbf{x}_{DP}^i}
\end{aligned}
$$

因此，有结果

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{x}_{1A}^i} &= \cos(\mathbf{x}_{1A}^i) \odot \mathbf{M} \odot \frac{\partial \mathcal{L}}{\partial \mathbf{x}_{DP}^i} \\
\frac{\partial \mathcal{L}}{\partial \theta_{1A}} &= \sum_{i=1}^{m} \frac{\partial \mathcal{L}}{\partial \mathbf{x}_{1A}^i} \mathbf{x}^{i\top} \\
\frac{\partial \mathcal{L}}{\partial \mathbf{b}_{1A}} &= \sum_{i=1}^{m} \frac{\partial \mathcal{L}}{\partial \mathbf{x}_{1A}^i}
\end{aligned}
$$