AI2651 智能语音识别:循环神经网络

521030910387 薛翔元

假定输入序列为 x, 输出序列为 \hat{r} , 标签序列为 r, 序列长度为 T_r , 类别数为 C, 使用 Sigmoid 激活函数和交叉熵损失函数。前向传播过程可以描述为

$$egin{aligned} oldsymbol{a}_t^{ ext{(in)}} &= \sigma \left(oldsymbol{W}^{ ext{(in)}} oldsymbol{x}_t + oldsymbol{b}^{ ext{(in)}}
ight) \ oldsymbol{h}_t &= \sigma \left(oldsymbol{U} oldsymbol{a}_t^{ ext{(in)}} + oldsymbol{V} oldsymbol{h}_{t-1} + oldsymbol{b}_h
ight) \ oldsymbol{h}_t^{ ext{(out)}} &= oldsymbol{W}^{ ext{(out)}} oldsymbol{o}_t + oldsymbol{b}^{ ext{(out)}} \ oldsymbol{\hat{r}}_t &= \operatorname{Softmax} \left(oldsymbol{h}_t^{ ext{(out)}}
ight) \ oldsymbol{\mathcal{L}} &= \sum_{t=1}^{T_r} \mathcal{L}_t = \sum_{t=1}^{T_r} -oldsymbol{r}_t^T \log oldsymbol{\hat{r}}_t \end{aligned}$$

Sigmoid 激活函数的导数为

$$\frac{\partial \sigma(x)}{\partial x} = \sigma(x) \left(1 - \sigma(x) \right)$$

将 Softmax 函数记为 s, 即

$$s_k(\mathbf{x}) = \frac{e^{x_k}}{\sum_{c=1}^{C} e^{x_c}}$$

则 Softmax 函数的导数为

$$\frac{\partial s_i(\boldsymbol{x})}{\partial x_j} = \begin{cases} s_i(\boldsymbol{x}) \left(1 - s_i(\boldsymbol{x})\right), & i = j \\ -s_i(\boldsymbol{x}) s_j(\boldsymbol{x}), & i \neq j \end{cases}$$

对于输出层有

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{h}_{t,i}^{(\text{out})}} = \sum_{t=1}^{T_r} \sum_{c=1}^{C} \frac{\partial \mathcal{L}_t}{\partial \hat{\boldsymbol{r}}_{t,c}} \frac{\partial \hat{\boldsymbol{r}}_{t,c}}{\partial \boldsymbol{h}_{t,i}^{(\text{out})}} = -\sum_{t=1}^{T_r} \left[\frac{\boldsymbol{r}_{t,i}}{\hat{\boldsymbol{r}}_{t,i}} \hat{\boldsymbol{r}}_{t,i} \left(1 - \hat{\boldsymbol{r}}_{t,i} \right) - \sum_{j \neq i} \frac{\boldsymbol{r}_{t,j}}{\hat{\boldsymbol{r}}_{t,j}} \hat{\boldsymbol{r}}_{t,i} \hat{\boldsymbol{r}}_{t,j} \right]$$

$$= -\sum_{t=1}^{T_r} \left(\boldsymbol{r}_{t,i} - \hat{\boldsymbol{r}}_{t,i} \sum_{j=1}^{C} \boldsymbol{r}_{t,j} \right) = \sum_{t=1}^{T_r} \left(\hat{\boldsymbol{r}}_{t,i} - \boldsymbol{r}_{t,i} \right)$$

于是

$$egin{aligned} rac{\partial \mathcal{L}}{\partial oldsymbol{W}_{i,j}^{(ext{out})}} &= rac{\partial \mathcal{L}}{\partial oldsymbol{h}_{t,i}^{(ext{out})}} rac{\partial oldsymbol{h}_{t,i}^{(ext{out})}}{\partial oldsymbol{W}_{i,j}^{(ext{out})}} = \sum_{t=1}^{T_r} \left(\hat{oldsymbol{r}}_{t,i} - oldsymbol{r}_{t,i}
ight) oldsymbol{o}_{t,j} \ rac{\partial \mathcal{L}}{\partial oldsymbol{b}_i^{(ext{out})}} &= rac{\partial \mathcal{L}}{\partial oldsymbol{h}_{t,i}^{(ext{out})}} rac{\partial oldsymbol{h}_{t,i}^{(ext{out})}}{\partial oldsymbol{b}_i^{(ext{out})}} = \sum_{t=1}^{T_r} \left(\hat{oldsymbol{r}}_{t,i} - oldsymbol{r}_{t,i}
ight) \end{aligned}$$

转化为矩阵形式有

$$egin{aligned} rac{\partial \mathcal{L}}{\partial oldsymbol{h}_t^{(ext{out})}} &= \sum_{t=1}^{T_r} \left(\hat{oldsymbol{r}}_t - oldsymbol{r}_t
ight) \ rac{\partial \mathcal{L}}{\partial oldsymbol{W}^{(ext{out})}} &= \sum_{t=1}^{T_r} \left(\hat{oldsymbol{r}}_t - oldsymbol{r}_t
ight) oldsymbol{o}_t^T \ rac{\partial \mathcal{L}}{\partial oldsymbol{b}^{(ext{out})}} &= \sum_{t=1}^{T_r} \left(\hat{oldsymbol{r}}_t - oldsymbol{r}_t
ight) \ rac{\partial \mathcal{L}}{\partial oldsymbol{o}_t} &= \sum_{t=1}^{T_r} \left[oldsymbol{W}^{(ext{out})}
ight]^T \left(\hat{oldsymbol{r}}_t - oldsymbol{r}_t
ight) \end{aligned}$$

对于隐藏层有

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}} = \sum_{t=1}^{T_r} \left[\boldsymbol{W}^{(\text{out})} \right]^T \left[(\hat{\boldsymbol{r}}_t - \boldsymbol{r}_t) \circ \boldsymbol{o}_t \circ (1 - \boldsymbol{o}_t) \right] \boldsymbol{h}_t^T$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}_o} = \sum_{t=1}^{T_r} \left[\boldsymbol{W}^{(\text{out})} \right]^T \left[(\hat{\boldsymbol{r}}_t - \boldsymbol{r}_t) \circ \boldsymbol{o}_t \circ (1 - \boldsymbol{o}_t) \right]$$

令 $\boldsymbol{\delta}_t = \frac{\partial \mathcal{L}_t}{\partial \boldsymbol{h}_t}$. 当 $t < T_r$ 时,我们有

$$\begin{aligned} \boldsymbol{\delta}_{t} &= \frac{\partial \mathcal{L}_{t}}{\partial \boldsymbol{h}_{t}} = \left(\frac{\partial \boldsymbol{o}_{t}}{\partial \boldsymbol{h}_{t}}\right)^{T} \frac{\partial \mathcal{L}_{t}}{\partial \boldsymbol{o}_{t}} + \left(\frac{\partial \boldsymbol{h}_{t+1}}{\partial \boldsymbol{h}_{t}}\right)^{T} \frac{\partial \mathcal{L}_{t}}{\partial \boldsymbol{h}_{t+1}} \\ &= \boldsymbol{W}^{T} \left[\boldsymbol{W}^{(\text{out})}\right]^{T} \left[(\hat{\boldsymbol{r}}_{t} - \boldsymbol{r}_{t}) \circ \boldsymbol{o}_{t} \circ (1 - \boldsymbol{o}_{t}) \right] + \boldsymbol{V}^{T} \left[\boldsymbol{\delta}_{t+1} \circ \boldsymbol{h}_{t+1} \circ (1 - \boldsymbol{h}_{t+1}) \right] \end{aligned}$$

当 $t = T_r$ 时,我们有

$$\boldsymbol{\delta}_t = \frac{\partial \mathcal{L}_t}{\partial \boldsymbol{h}_t} = \left(\frac{\partial \boldsymbol{o}_t}{\partial \boldsymbol{h}_t}\right)^T \frac{\partial \mathcal{L}_t}{\partial \boldsymbol{o}_t} = \boldsymbol{W}^T \left[\boldsymbol{W}^{(\text{out})} \right]^T \left[(\hat{\boldsymbol{r}}_t - \boldsymbol{r}_t) \circ \boldsymbol{o}_t \circ (1 - \boldsymbol{o}_t) \right]$$

因此

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{U}} = \sum_{t=1}^{T_r} \left(\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{U}} \right)^T \frac{\partial \mathcal{L}_t}{\partial \boldsymbol{h}_t} = \sum_{t=1}^{T_r} \left[\boldsymbol{h}_t \circ (1 - \boldsymbol{h}_t) \circ \boldsymbol{\delta}_t \right] \left[\boldsymbol{a}_t^{(\text{in})} \right]^T
\frac{\partial \mathcal{L}}{\partial \boldsymbol{V}} = \sum_{t=1}^{T_r} \left(\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{V}} \right)^T \frac{\partial \mathcal{L}_t}{\partial \boldsymbol{h}_t} = \sum_{t=1}^{T_r} \left[\boldsymbol{h}_t \circ (1 - \boldsymbol{h}_t) \circ \boldsymbol{\delta}_t \right] \boldsymbol{h}_{t-1}^T
\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}_h} = \sum_{t=1}^{T_r} \left(\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{b}_h} \right)^T \frac{\partial \mathcal{L}_t}{\partial \boldsymbol{h}_t} = \sum_{t=1}^{T_r} \left[\boldsymbol{h}_t \circ (1 - \boldsymbol{h}_t) \circ \boldsymbol{\delta}_t \right]$$

对于输入层有

$$\begin{split} \frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^{(\text{in})}} &= \sum_{t=1}^{T_r} \left(\frac{\partial \boldsymbol{a}_t^{(\text{in})}}{\partial \boldsymbol{W}^{(\text{in})}} \right)^T \frac{\partial \mathcal{L}_t}{\partial \boldsymbol{a}_t^{(\text{in})}} = \sum_{t=1}^{T_r} \boldsymbol{U}^T \left[\boldsymbol{h}_t \circ (1 - \boldsymbol{h}_t) \circ \boldsymbol{\delta}_t \right] \left[\boldsymbol{x}_t \right]^T \\ \frac{\partial \mathcal{L}}{\partial \boldsymbol{b}^{(\text{in})}} &= \sum_{t=1}^{T_r} \left(\frac{\partial \boldsymbol{a}_t^{(\text{in})}}{\partial \boldsymbol{b}^{(\text{in})}} \right)^T \frac{\partial \mathcal{L}_t}{\partial \boldsymbol{a}_t^{(\text{in})}} = \sum_{t=1}^{T_r} \boldsymbol{U}^T \left[\boldsymbol{h}_t \circ (1 - \boldsymbol{h}_t) \circ \boldsymbol{\delta}_t \right] \end{split}$$

综上所述, 反向传播的更新公式如下

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^{(\text{out})}} = \sum_{t=1}^{T_r} (\hat{\boldsymbol{r}}_t - \boldsymbol{r}_t) \, \boldsymbol{o}_t^T \\
\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}^{(\text{out})}} = \sum_{t=1}^{T_r} (\hat{\boldsymbol{r}}_t - \boldsymbol{r}_t) \\
\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}} = \sum_{t=1}^{T_r} \left[\boldsymbol{W}^{(\text{out})} \right]^T \left[(\hat{\boldsymbol{r}}_t - \boldsymbol{r}_t) \circ \boldsymbol{o}_t \circ (1 - \boldsymbol{o}_t) \right] \boldsymbol{h}_t^T \\
\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}_o} = \sum_{t=1}^{T_r} \left[\boldsymbol{W}^{(\text{out})} \right]^T \left[(\hat{\boldsymbol{r}}_t - \boldsymbol{r}_t) \circ \boldsymbol{o}_t \circ (1 - \boldsymbol{o}_t) \right] \\
\frac{\partial \mathcal{L}}{\partial \boldsymbol{U}} = \sum_{t=1}^{T_r} \left[\boldsymbol{h}_t \circ (1 - \boldsymbol{h}_t) \circ \boldsymbol{\delta}_t \right] \left[\boldsymbol{a}_t^{(\text{in})} \right]^T \\
\frac{\partial \mathcal{L}}{\partial \boldsymbol{V}} = \sum_{t=1}^{T_r} \left[\boldsymbol{h}_t \circ (1 - \boldsymbol{h}_t) \circ \boldsymbol{\delta}_t \right] \boldsymbol{h}_{t-1}^T \\
\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}_h} = \sum_{t=1}^{T_r} \left[\boldsymbol{h}_t \circ (1 - \boldsymbol{h}_t) \circ \boldsymbol{\delta}_t \right] \\
\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^{(\text{in})}} = \sum_{t=1}^{T_r} \boldsymbol{U}^T \left[\boldsymbol{h}_t \circ (1 - \boldsymbol{h}_t) \circ \boldsymbol{\delta}_t \right] \left[\boldsymbol{x}_t \right]^T \\
\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}^{(\text{in})}} = \sum_{t=1}^{T_r} \boldsymbol{U}^T \left[\boldsymbol{h}_t \circ (1 - \boldsymbol{h}_t) \circ \boldsymbol{\delta}_t \right] \\
\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}^{(\text{in})}} = \sum_{t=1}^{T_r} \boldsymbol{U}^T \left[\boldsymbol{h}_t \circ (1 - \boldsymbol{h}_t) \circ \boldsymbol{\delta}_t \right]$$