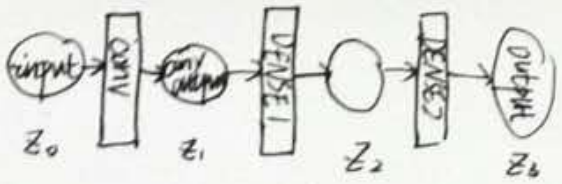
Homework 7

1.



令经过 conv/dense 运算，没经过非线性的表示为 y_i，经过 relu 非线性之后为 z_i $(i=1,\cdots 3)$

loss $L = L(z_3)$ $\quad z_i = f(y_i)$ $\quad i=1,2,3$

$y_i = g_i(z_{i-1})$ $\quad i=1,\cdots 3$ $\quad g_1 \cdots g_3$ 参数分别设为 $w_1 \cdots w_3$

其中 g_1 为 conv 操作，g_2, g_3 为 dense 操作.

$$\frac{\partial L}{\partial z_3} = L'|_{z_3} \qquad \frac{\partial L}{\partial y_3} = L'|_{z_3} f'|_{y_3} \qquad y_3 = w_3^T z_2$$

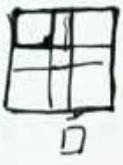
$$\frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial z_3}\frac{\partial z_3}{\partial y_3}\frac{\partial y_3}{\partial w_3} = L'|_{z_3} f'|_{y_3} z_2$$

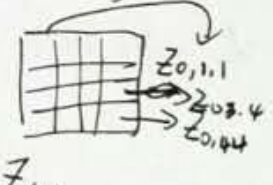
$$\frac{\partial L}{\partial z_2} = L'|_{z_3} f'|_{y_3} w_3 \qquad y_2 = w_2^T z_1$$

$$\frac{\partial L}{\partial y_2} = L'|_{z_3} f'|_{y_3} w_3 f'|_{y_2} \qquad \frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial y_2}\frac{\partial y_2}{\partial w_2} = L'|_{z_3} f'|_{y_3} w_3 f'|_{y_2} z_1$$

$$\frac{\partial L}{\partial z_1} = \frac{\partial L}{\partial y_2}\frac{\partial y_2}{\partial z_1} = L'|_{z_3} f'|_{y_3} w_3 f'|_{y_2} w_2$$

$$\frac{\partial L}{\partial y_1} = \frac{\partial L}{\partial z_1}\frac{\partial z_1}{\partial y_1} = L'|_{z_3} f'|_{y_3} w_3 f'|_{y_2} w_2 f'|_{y_1}$$

$y_1 = conv(w_1, z_0)$ $\quad w_1 =$  $\{w_{11} \cdots w_{19}, b_1\}$



$$\frac{\partial L}{\partial w_{ij}} = \sum_{k=1}^{4}\frac{\partial L}{\partial y_{1k}}\frac{\partial y_{1k}}{\partial w_{ij}}$$

对于 $j=1$ $\quad \frac{\partial L}{\partial w_{11}} = \frac{\partial L}{\partial y_{11}} z_{011} + \frac{\partial L}{\partial y_{12}} z_{012} + \frac{\partial L}{\partial y_{13}} z_{021} + \frac{\partial L}{\partial y_{14}} z_{022}$

$$\therefore \frac{\partial L}{\partial w_{ij}} = conv(\frac{\partial L}{\partial y_1}, z_0)$$

而 $\frac{\partial L}{\partial b_1} = 1 \cdot \frac{\partial L}{\partial y_1}$

$$\therefore K_{t+1} \leftarrow K_t - \eta \frac{\partial L}{\partial K}\Big|_{K_t} = K_t - \eta \frac{\partial L}{\partial w_{ij}} = K_t - \eta\, conv(L'|_{z_3} f'|_{y_3} w_3 f'|_{y_2} w_2 f'|_{y_1}, z_0)$$

$$B_{t+1} = B_t - \eta \frac{\partial L}{\partial b_1} = B_t - \eta L'|_{z_3} f'|_{y_3} w_3 f'|_{y_2} w_2 f'|_{y_1}$$

2. 序列条件分布 cross entropy 为
单个训练数据 x

$$L(P_{model}) = H(P_{true}, P_{model}) = -\sum_{y} P_{true}(y) \log P_{model}(y|x)$$

cross entropy loss 中 P_{true} 只在训练样本 y 处为 1 其它为 0. ∴ 设训练集为 $\{(x^{(i)}, y^{(i)})\}_{i=1\cdots N}$ 一共 N 组训练数据

$$L = -\sum_{i=1}^{N} \log P_{model}(y^{(i)}|x^{(i)})$$

每个 $y^{(i)}$ 长度为 $T=3$ 包含 $y_1^{(i)}$ $y_2^{(i)}$ $y_3^{(i)}$

其中 $P_{model}(y^{(i)}|x^{(i)}) = P(y_1^{(i)}|x_1^{(i)}) P(y_2^{(i)}|x_2^{(i)}, h_1^{(i)})$

$$P(y_3^{(i)}|x_3^{(i)}, h_2^{(i)})$$

$$= \gamma_1^{*}(x_0^{(i)})^T y_1^{(i)} \gamma_2(x^{(i)})^T y_2^{(i)} \gamma_3(x^{(i)})^T y_3^{(i)}$$

其中 $\gamma_1(x)$, $\gamma_2(x)$, $\gamma_3(x)$ 为将 $x^{(i)}$ 输入 RNN 得到的每一个时间 $t=1,\cdots 3$ 得到的权重向量

$y_1^{(i)}$, $y_2^{(i)}$, $y_3^{(i)}$ 讲道理应该是种型 1 下的指示向量.

2) 为表示方便. 样本仍由 $x^{(i)}$, $y^{(i)}$ 表示. 模型得到的用大写 Γ, γ 表示.

① 对于任意参数 $w \in$ W, U, V 分别有

$$\Delta w = -lr\frac{\partial L}{\partial w} = lr \frac{\partial \log P_{model}(y^{(i)}|x^{(i)})}{\partial w} = lr \frac{\frac{\partial P_{model}(y^{(i)}|x^{(i)})}{\partial w}}{P_{model}(y^{(i)}|x^{(i)})}$$

$$= lr \frac{1}{\prod_{j=1} \Gamma_j(x^{(i)})^T y_j^{(i)}} \frac{\partial P_{model}(y^{(i)}|x^{(i)})}{\partial w}$$

① 对于 V：$\frac{\partial P_{model}}{\partial V} = \sum_{t=1}^{T} \frac{\partial P_{model}}{\partial Y_t} \frac{\partial Y_t}{\partial V}$

$\frac{\partial Y_t}{\partial V} = \text{softmax}'\big|_{res=Y_t} H_j^T$　　$\text{softmax}'\big|_{res=Y_t}$ 第 i,j 项为 $\begin{cases} i=j & Y_{ti}(1-Y_{ti}) \\ j\neq i & -Y_{ti}Y_{tj} \end{cases}$ (*)

$\frac{\partial P_{model}}{\partial Y_t} = \prod_{j=1}^{T} Y_j^T y_j^{(i)} \left(\frac{y_t^{(i)}}{Y_t^T y_t^{(i)}} \right)$

$\therefore \frac{\partial P_{model}}{\partial V} = \sum_{t=1}^{T} \prod_{j=1}^{T} Y_j^T y_i^{(i)} \left(\frac{y_t^{(i)}}{Y_t^T y_{(t)}^{(i)}} \right) \text{softmax}'\big|_{res=Y_t} H_j^T$

$\Delta V = lr \cdot \frac{\frac{\partial P_{model}}{\partial V}}{\prod_{j=1}^{T} Y_j^T y_i^{(i)}} = \sum_{t=1}^{T} \frac{y_t^{(i)}}{Y_t^T y_{(t)}^{(i)}} \text{softmax}'\big|_{res=Y_t} H_j^T$

其中 softmax′ 在上面 (*) 给出

② 对于 U：考虑 $\frac{\partial P_{model}}{\partial H_t}$

对于 $t=3$：$\frac{\partial P}{\partial H_3} = \frac{\partial P}{\partial Y_3} \cdot \frac{\partial Y_3}{\partial H_3} = \frac{\partial P}{\partial Y_3} \odot \text{softmax}'\big|_{Y_3} V$

其中 $\text{softmax}'\big|_{Y_3}$ 与 $\frac{\partial P}{\partial Y_3}$ 都已在上面给出

$t=2, t=1$：$\frac{\partial P}{\partial H_t} = \frac{\partial P}{\partial Y_t} \frac{\partial Y_t}{\partial H_t} + \frac{\partial P}{\partial H_{t+1}} \frac{\partial H_{t+1}}{\partial H_t}$

其中 $\frac{\partial H_{t+1}}{\partial H_t} = \text{sigmoid}'\big|_{res=H_{t+1}} W$　其中 $\text{sigmoid}'\big|_{res=H_{t+1}} = diag[H_{t+1} \odot (1 - H_{t+1})]$

"$\odot$" 为 element wise 乘.

其中 $\frac{\partial P}{\partial Y_t}$, $\frac{\partial Y_t}{\partial H_t}$ 以及 $\frac{\partial P}{\partial H_{t+1}}$ 在前已计算得到.

$\frac{\partial P}{\partial U} = \sum_{t=1}^{T} \frac{\partial P}{\partial H_t} \frac{\partial H_t}{\partial U}$　其中 $\frac{\partial H_t}{\partial U} = \text{sigmoid}'\big|_{res=H_t} \odot X_t^T$

$\frac{\partial P}{\partial H_t}$ 在上面给出　　$\Delta U = lr \frac{\sum_{t=1}^{T} \frac{\partial P}{\partial H_t} \frac{\partial H_t}{\partial U}}{\prod_{j=1}^{T} Y_j^T y_i^{(i)}}$

③ 对于 W：$\frac{\partial P}{\partial W} = \sum_{t=1}^{T} \frac{\partial P}{\partial H_t} \frac{\partial H_t}{\partial W} = \sum_{t=2}^{T} \frac{\partial P}{\partial H_t} \text{sigmoid}'\big|_{res=H_t} H_{t-1}^T$　(设 $H_0 = 0$)

其中 $\text{sigmoid}'\big|_{res=H_t}$, $\frac{\partial L}{\partial H_t}$ 都在上节给出.

有 $\Delta W = lr \frac{1}{\prod_{j=1}^{T} Y_j^T y_i^{(i)}} \cdot \sum_{t=2}^{T} \frac{\partial P}{\partial H_t} \odot \frac{\partial H_t}{\partial W}$