CS 224N:Assignment #2
WeiYang

2. Neural Transition-Based Dependency Parsing

a)

stack	buffer	new dependency	transition
$ROOT$	$I, parsed, this, sentence, correctly$		$Initial\ Configuration$
$ROOT, I$	$parsed, this, sentence, correctly$		$SHIFT$
$ROOT, I, parsed$	$this, sentence, correctly$		$SHIFT$
$ROOT, parsed$	$this, sentence, correctly$	$parsed \to I$	$LEFT-ARC$
$ROOT, parsed, this$	$sentence, correctly$		$SHIFT$
$ROOT, parsed, this, sentence$	$correctly$		$SHIFT$
$ROOT, parsed, sentence$	$correctly$	$sentence \to this$	$LEFT-ARC$
$ROOT, parsed$	$correctly$	$parsed \to sentence$	$RIGHT-ARC$
$ROOT, parsed, correctly$			$SHIFT$
$ROOT, parsed$		$parsed \to correctly$	$RIGHT-ARC$
$ROOT$		$ROOT \to parsed$	$RIGHT-ARC$

b)

$2n$ 步，因为每个单词移进栈需要 n 步，移出栈需要 n 步。

f)

$$\mathrm{E}_{p_{drop}}[h_{drop}]_i = \mathrm{E}_{p_{drop}}[\gamma d_i h_i] = p_{drop} \cdot 0 + (1-p_{drop})\gamma h_i = (1-p_{drop})\gamma h_i = h_i$$

$$\Rightarrow \gamma = \frac{1}{1-p_{drop}}$$

g)

i. 因为 β_1 接近 1，所以每次更新量与上一次基本相同，不会导致梯度振荡过大的情况。

ii. 那些梯度较小的参数也会得到较大的更新。

3. Recurrent Neural Networks: Language Modeling

a)

$$\mathrm{CE}(y^{(t)}, \hat{y}^{(t)}) = -\log \hat{y}_i^{(t)} = \log \frac{1}{\hat{y}_i^{(t)}}$$

$$\mathrm{PP}^{(t)}(y^{(t)}, \hat{y}^{(t)}) = \frac{1}{\hat{y}_i^{(t)}} = e^{\mathrm{CE}(y^{(t)}, \hat{y}^{(t)})}$$

$$E(\hat{y}^{(t)}) = \frac{1}{|V|}$$
$$E(\mathrm{PP}^{(t)}(y^{(t)}, \hat{y}^{(t)})) = |V|$$
$$E(\mathrm{CE}(\mathrm{y}^{(\mathrm{t})}, \hat{\mathrm{y}}^{(\mathrm{t})})) = \log|V| = \log 10000 \approx 9.21$$

b)

令

$$v^{(t)} = h^{(t-1)}H + e^{(t)}I + b_1$$
$$\theta^{(t)} = h^{(t)}U + b_2$$

所以

$$\delta_1^{(t)} = \frac{\partial J}{\partial \theta^{(t)}} = \hat{y}^{(t)} - y^{(t)}$$
$$\delta_2^{(t)} = \frac{\partial J}{\partial v^{(t)}} = \delta_1^{(t)} U^T h^{(t)} (1 - h^{(t)})$$

所以

$$\frac{\partial J}{\partial b_2} = \frac{\partial J}{\partial \theta^{(t)}} \frac{\partial \theta^{(t)}}{\partial b_2} = \delta_1^{(t)}$$
$$\frac{\partial J}{\partial \mathrm{L}_{x^{(t)}}} = \frac{\partial J}{\partial v^{(t)}} \frac{\partial v^{(t)}}{\partial e^{(t)}} \frac{\partial e^{(t)}}{\partial \mathrm{L}_{x^{(t)}}} = \delta_2^{(t)} I^T$$
$$\frac{\partial J}{\partial I} = \frac{\partial J}{\partial v^{(t)}} \frac{\partial v^{(t)}}{\partial I} = (e^{(t)})^T \delta_2^{(t)}$$
$$\frac{\partial J}{\partial H} = \frac{\partial J}{\partial v^{(t)}} \frac{\partial v^{(t)}}{\partial H} = (h^{(t-1)})^T \delta_2^{(t)}$$
$$\frac{\partial J}{\partial h^{(t-1)}} = \frac{\partial J}{\partial v^{(t)}} \frac{\partial v^{(t)}}{\partial h^{(t-1)}} = \delta_2^{(t)} H^T$$

c)

令

$$\sigma'(v^{(t-1)}) = \frac{\partial h^{(t-1)}}{\partial v^{(t-1)}} = diag(h^{(t-1)}(1 - h^{(t-1)}))$$

所以

$$\frac{\partial J}{\partial L_{x^{(t-1)}}} = \frac{\partial J}{\partial h^{(t-1)}} \frac{\partial h^{(t-1)}}{\partial v^{(t-1)}} \frac{\partial v^{(t-1)}}{\partial L_{x^{(t-1)}}} = \delta^{(t-1)} \sigma'(v^{(t-1)}) I^T$$

$$\frac{\partial J}{\partial I} = \frac{\partial J}{\partial h^{(t-1)}} \frac{\partial h^{(t-1)}}{\partial v^{(t-1)}} \frac{\partial v^{(t-1)}}{\partial I} = (e^{(t-1)})^T \delta^{(t-1)} \sigma'(v^{(t-1)})$$

$$\frac{\partial J}{\partial H} = \frac{\partial J}{\partial h^{(t-1)}} \frac{\partial h^{(t-1)}}{\partial v^{(t-1)}} \frac{\partial v^{(t-1)}}{\partial H} = (h^{(t-2)})^T \delta^{(t-1)} \sigma'(v^{(t-1)})$$

d)

前向传播：

$$O(dD_h + D_h^2 + |V| D_h)$$

反向传播：

$$O(dD_h + D_h^2 + \tau |V| D_h)$$

计算 softmax 速度最慢，可以用 NCE 代替。