CS 224N **Name: Xin Li**
Homework #2 **SUNet ID: 06121602**

Answer to Question 1: Tensorflow Softmax

(1.a) q1_softmax.py is submitted.

(1.b) q1_softmax.py is submitted.

(1.c) q1_classifier.py is submitted.
Placeholder variables are nodes to which we will later assign inputs.
Feed dictionaries are used to map from placeholders to actual values.

(1.d) q1_classifier.py is submitted.

(1.e) q1_classifier.py is submitted.
Tensorflow will use back propagation to generate the gradients automatically.

Answer to Question 2: Deep Networks for Named Entity Recognition

(2.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 → I	LEFT-ARC
[ROOT, parsed, this]	[sentence, correctly]		SHIFT
[ROOT, parsed, this, sentence]	[correctly]		SHIFT
[ROOT, parsed, sentence]	[correctly]	sentence → this	LEFT-ARC
[ROOT, parsed]	[correctly]	parsed → sentense	RIGHT-ARC
[ROOT, parsed, correctly]	$\emptyset$		SHIFT
[ROOT, parsed]	$\emptyset$	parsed → correctly	RIGHT-ARC
[ROOT]	$\emptyset$	root → parsed	RIGHT-ARC

(2.b)
A sentence containing n words will be parsed in 2*n steps.
Because it takes n steps to shift each word into the stack if we don't consider the sequence.
Also, it takes another n steps to remove each word from the stack.
So after 2*n steps, the stack remains its initial state, which means that parsing is finished.

(2.c) q2_parser_transitions.py is submitted.

(2.d) q2_parse_transitions.py is submitted.

(2.e) q2_initialization.py is submitted.

(2.f)

$$\begin{aligned}
\mathbb{E}_{p_{drop}}[\boldsymbol{h}_{drop}]_i &= \mathbb{E}_{p_{drop}}[\gamma \boldsymbol{d} \circ \boldsymbol{h}]_i \\
&= \gamma \mathbb{E}_{p_{drop}}[\boldsymbol{d} \circ \boldsymbol{h}]_i \\
&= \gamma \cdot prob(\boldsymbol{h}_i = 1) \cdot \boldsymbol{h}_i \\
&= \gamma (1 - p_{drop}) \boldsymbol{h}_i \\
\mathbb{E}_{p_{drop}}[\boldsymbol{h}_{drop}]_i &= \boldsymbol{h}_i \\
\therefore \gamma &= \frac{1}{1 - p_{drop}}
\end{aligned}$$

(2.g.i)

β_1 controls the rate that $\boldsymbol{m}$ changes from the previous $\boldsymbol{m}$. So if $\nabla_\theta J$ vary much and β_1 is set to be small enough, the updates won't change much from its previous value, which means the updates are more stable.

The learning rate depends on α and β_1. It can help preventing θ from oscillating.

(2.g.ii)

When $\nabla_\theta J$ is smaller, the model parameters will get larger updates.

So that updates won't vary much even if $\Delta_\theta J$ is very large. Consequently θ won't oscillate much.

(2.h)

Best UAS the model achieves on dev set: 88.58

UAS the model achieves on test set: 89.05

Answer to Question 3: Recurrent Neural Networks: Language Modeling

(3.a)

$$
\begin{aligned}
CE(\boldsymbol{y}^{(t)}, \hat{\boldsymbol{y}}^{(t)}) &= -\sum_{i=1}^{|V|} \boldsymbol{y}_{\boldsymbol{i}}^{(t)} log(\hat{\boldsymbol{y}}_i{}^{(t)}) \\
&= -\boldsymbol{y}_j^{(t)} log(\hat{\boldsymbol{y}}_j{}^{(t)})|_{\boldsymbol{y}_j^{(t)}=1} \\
&= log(\frac{1}{\hat{\boldsymbol{y}}_j{}^{(t)}}) \\
&= log(\frac{1}{\boldsymbol{y}_j^{(t)} \hat{\boldsymbol{y}}_j{}^{(t)}})|_{\boldsymbol{y}_j^{(t)}=1} \\
&= log(\frac{1}{\sum_{i=1}^{|V|} \boldsymbol{y}_{\boldsymbol{i}}^{(t)} \hat{\boldsymbol{y}}_i{}^{(t)}}) \\
&= log(\boldsymbol{PP}^{(t)}(\boldsymbol{y}_{\boldsymbol{i}}^{(t)}, \hat{\boldsymbol{y}}_i{}^{(t)}))
\end{aligned}
$$

So that minimizing the mean cross-entropy loss will also minimize the mean perplexity.

$$
\begin{aligned}
E[\boldsymbol{PP}^{(t)}(\boldsymbol{y}_{\boldsymbol{i}}^{(t)}, \hat{\boldsymbol{y}}_i{}^{(t)})] &= E[\frac{1}{\bar{p}(\boldsymbol{x}_{pred}^{(t+1)} = \boldsymbol{x}_{(t+1)}|\boldsymbol{x}_{(t)}, \cdots, \boldsymbol{x}_{(1)})}] \\
&= E[\frac{1}{\frac{1}{|V|}}] \\
&= E[|V|] \\
&= 10000 \\
E[CE(\boldsymbol{y}^{(t)}, \hat{\boldsymbol{y}}^{(t)})] &= E[log(\boldsymbol{PP}^{(t)}(\boldsymbol{y}_{\boldsymbol{i}}^{(t)}, \hat{\boldsymbol{y}}_i{}^{(t)}))] \\
&= log(E[\boldsymbol{PP}^{(t)}(\boldsymbol{y}_{\boldsymbol{i}}^{(t)}, \hat{\boldsymbol{y}}_i{}^{(t)})]) \\
&= log(10000) \\
&= 9.2103
\end{aligned}
$$

(3.b)

$$J^{(t)} = -\sum_{j=1}^{|V|} y_j^{(t)} \log \hat{y}_j^{(t)}$$

$$= -\log \hat{y}_i^{(t)}$$

$$= -\log \text{softmax}(h^{(t)}U + b_2)_i$$

$$= \log \frac{\sum_{j=1}^{|V|} \exp(h^{(t)}U+b_2)_j}{\exp(h^{(t)}U+b_2)_i}$$

$$= \log \sum_{j=1}^{|V|} \exp(h^{(t)}U+b_2)_j - (h^{(t)}U+b_2)_i$$

$$\frac{\partial J^{(t)}}{\partial (b_2)_i} = \frac{\exp(h^{(t)}U+b_2)_i}{\sum_{j=1}^{|V|} \exp(h^{(t)}U+b_2)_j} - 1 = \hat{y}_i^{(t)} - 1 = \hat{y}_i^{(t)} - y_i^{(t)}$$

$$\left.\frac{\partial J^{(t)}}{\partial (b_2)_j}\right|_{j \neq i} = \frac{\exp(h^{(t)}U+b_2)_j}{\sum_{j=1}^{|V|} \exp(h^{(t)}U+b_2)_j} = \hat{y}_i^{(t)} = \hat{y}_i^{(t)} - 0 = \hat{y}_i^{(t)} - y_i^{(t)}$$

$$\boxed{\therefore \frac{\partial J^{(t)}}{\partial b_2} = \hat{y}^{(t)} - y^{(t)}}$$

$$\frac{\partial J^{(t)}}{\partial L_{x^{(t)}}} = \frac{\partial J^{(t)}}{\partial e^{(t)}}$$

$$= \frac{\partial J^{(t)}}{\partial h^{(t)}} \odot \frac{\partial h^{(t)}}{\partial e^{(t)}}$$

$$= (\hat{y} - y)U^T \odot \text{sigmoid}'(h^{(t)}) \cdot \frac{\partial (h^{(t-1)}H + e^{(t)}I + b_1)}{\partial e^{(t)}}$$

$$= (\hat{y} - y)U^T \odot \left(h^{(t)}(1 - h^{(t)})\right) \cdot I^T$$

$$\boxed{\therefore \frac{\partial J^{(t)}}{\partial L_{x^{(t)}}} = (\hat{y} - y)U^T \odot (h^{(t)}(1 - h^{(t)})) \cdot I^T}$$

$$\left.\frac{\partial J^{(t)}}{\partial I}\right|_{(t)} = \frac{\partial J^{(t)}}{\partial h^{(t)}} \odot \left.\frac{\partial h^{(t)}}{\partial I}\right|_{(t)}$$

$$= (\hat{y} - y)U^T \odot \text{sigmoid}'(h^{(t)}) \cdot \left.\frac{\partial h^{(t)}}{\partial I}\right|_{(t)}$$

$$= (\hat{y}-y)U^T \odot (h^{(t)}(1-h^{(t)})) \cdot \frac{\partial (h^{(t-1)}H + e^{(t)}I + b_1)}{\partial I}\Big|_{(t)}$$

$$= (\hat{y}-y)U^T \odot \left(e^{(t)T} h^{(t)}(1-h^{(t)})\right)$$

$$\therefore \boxed{\frac{\partial J^{(t)}}{\partial I} = (\hat{y}-y)U^T \odot (e^{(t)T} h^{(t)}(1-h^{(t)}))}$$

$$\frac{\partial J^{(t)}}{\partial H}\Big|_{(t)} = \frac{\partial J^{(t)}}{\partial h^{(t)}} \odot \text{sigmoid}'(h^{(t)}) \cdot \frac{\partial h^{(t)}}{\partial H}\Big|_{(t)}$$

$$= (\hat{y}-y)U^T \odot (h^{(t)}(1-h^{(t)})) \cdot \frac{\partial (h^{(t-1)}H + e^{(t)}I + b_1)}{\partial H}\Big|_{(t)}$$

$$= (\hat{y}-y)U^T \odot \left(h^{(t-1)T} h^{(t)}(1-h^{(t)})\right)$$

$$\therefore \boxed{\frac{\partial J^{(t)}}{\partial H}\Big|_{(t)} = (\hat{y}-y)U^T \odot (h^{(t-1)T} h^{(t)}(1-h^{(t)}))}$$

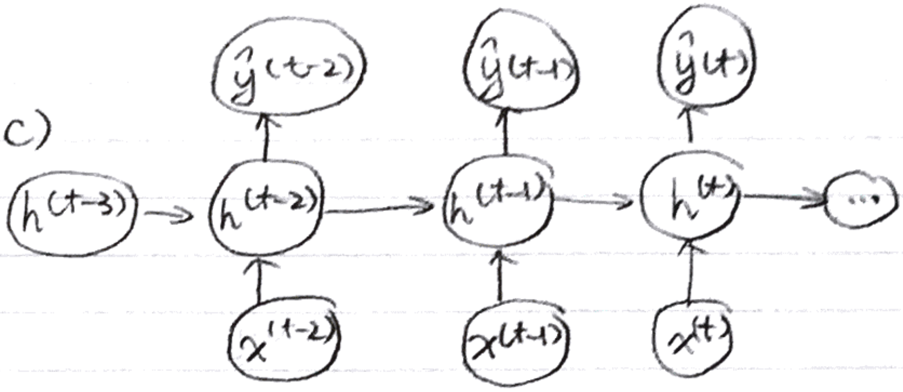
$$\frac{\partial J^{(t)}}{\partial h^{(t-1)}} = \frac{\partial J^{(t)}}{\partial h^{(t)}} \odot \text{sigmoid}'(h^{(t)}) \cdot \frac{\partial h^{(t)}}{\partial h^{(t-1)}}$$

$$= (\hat{y}-y)U^T \odot (h^{(t)}(1-h^{(t)})) \cdot \frac{\partial (h^{(t-1)}H + e^{(t)}I + b_1)}{\partial h^{(t-1)}}$$

$$= (\hat{y}-y)U^T \odot (h^{(t)}(1-h^{(t)})) \cdot H^T$$

$$\therefore \boxed{\frac{\partial J^{(t)}}{\partial h^{(t-1)}} = (\hat{y}-y)U^T \odot (h^{(t)}(1-h^{(t)})) \cdot H^T}$$

(3.c)



$$\frac{\partial J^{(t)}}{\partial L_{x^{(t-1)}}} = \frac{\partial J^{(t)}}{\partial h^{(t-1)}} \odot \frac{\partial h^{(t-1)}}{\partial J^{(t-1)}} \odot \frac{\partial J^{(t-1)}}{\partial L_{x^{(t-1)}}}$$

$$= \delta^{(t-1)} \odot \frac{1}{\cancel{(\hat{y}-y)U^T}} \odot \cancel{(\hat{y}-y)U^T} \odot (1-h^{(t-1)})\, h^{(t-1)} I^T$$

$$= \delta^{(t-1)} \odot (1-h^{(t-1)})\, h^{(t-1)} I^T$$

$$\left.\frac{\partial J^{(t)}}{\partial I}\right|_{(t-1)} = \left.\frac{\partial J^{(t)}}{\partial h^{(t-1)}} \odot \frac{\partial h^{(t-1)}}{\partial J^{(t-1)}} \odot \frac{\partial J^{(t-1)}}{\partial I}\right|_{(t-1)}$$

$$= \delta^{(t-1)} \odot \frac{1}{\cancel{(\hat{y}-y)U^T}} \odot e^{(t-1)T} \cancel{(\hat{y}-y)U^T} (1-h^{(t-1)})\, h^{(t-1)}$$

$$= \delta^{(t-1)} \odot e^{(t-1)T} (1-h^{(t-1)})\, h^{(t-1)}$$

$$\left.\frac{\partial J^{(t)}}{\partial H}\right|_{(t-1)} = \left.\frac{\partial J^{(t)}}{\partial h^{(t-1)}} \odot \frac{\partial h^{(t-1)}}{\partial J^{(t-1)}} \odot \frac{\partial J^{(t-1)}}{\partial H}\right|_{(t-1)}$$

$$= \delta^{(t-1)} \odot \frac{1}{\cancel{(\hat{y}-y)U^T}} \odot \cancel{(\hat{y}-y)U^T} \odot (h^{(t-2)T}(1-h^{(t-1)})\, h^{(t-1)})$$

$$= \delta^{(t-1)} \odot h^{(t-2)T} (1-h^{(t-1)})\, h^{(t-1)}$$

(3.d)

Forward : $O(D_h^2 + |V|D_h) = O(|V|D_h)$ $(\because |V| >> D_h)$

Backward : same as forward propagation $O(|V|D_h)$

τ steps : $O(\tau |V| D_h)$

Slow step : computation of $\hat{y}^{(t)}$