NLP Assignment

Assignment 1

Part 1

Problem 1:

a)

we have $y_w = \begin{cases} 1 & \text{if } w = o \\ 0 & \text{if } w \neq o \end{cases} \Rightarrow -\sum_{w \in vocab} y_w \cdot \log(\hat{y}_w) = -y_o \cdot \log(\hat{y}_o)$

$= -\log(\hat{y}_o)$

b) $J_{naive\text{-}softmax}(v_c, o, U) = -\log p(O = o, C = c)$

$$= -\log \frac{\exp(u_o^T \cdot v_c)}{\sum_{w \in vocab} \exp(u_w^T \cdot v_c)}$$

$$\frac{\partial J}{\partial v_c} = \frac{\partial}{\partial v_c}\left(-\log \frac{\exp(u_o^T \cdot v_c)}{\sum_{w \in vocab} \exp(u_w^T \cdot v_c)}\right)$$

$$= -\frac{\partial}{\partial v_c} \log(\exp(u_o^T \cdot v_c)) + \frac{\partial}{\partial v_c} \cdot \log\left(\sum_w \exp(u_w^T \cdot v_c)\right)$$

$$= -\frac{\partial}{\partial v_c} u_o^T \cdot v_c + \frac{\partial}{\partial v_c} \log\left(\sum_w \exp(u_w^T \cdot v_c)\right)$$

$$= -u_o + \frac{\partial}{\partial v_c} \log\left(\sum_w \exp(u_w^T \cdot v_c)\right)$$

we have:

$$\frac{\partial}{\partial v_c} \log\left(\sum_w \exp(u_w^T \cdot v_c)\right) = \frac{1}{\sum_w \exp(u_w^T \cdot v_c)} \cdot \frac{\partial}{\partial v_c} \sum_{x \in vocab} \exp(u_x^T \cdot v_c)$$

$$= \frac{1}{\sum_w \exp(u_w^T \cdot v_c)} \cdot \sum_x \frac{\partial}{\partial v_c} \exp(u_x^T \cdot v_c)$$

$$= \frac{1}{\sum_w \exp(u_w^T \cdot v_c)} \cdot \sum_x \exp(u_x^T \cdot v_c) \cdot \frac{\partial}{\partial v_c} (u_x^T \cdot v_c)$$

$$= \frac{1}{\sum_w \exp(u_w^T \cdot v_c)} \cdot \sum_x \exp(u_x^T \cdot v_c) \cdot u_x$$

$$= \sum_x \frac{\exp(u_x^T \cdot v_c)}{\sum_w \exp(u_w^T \cdot v_c)} \cdot u_x$$

$$= \sum_x p(O=x|C=c) \cdot u_x = \sum_x \hat{y}_x \cdot u_x$$

So: $\frac{\partial J}{\partial v_c} = -u_o + \sum_{x \in vocab} \hat{y}_x \cdot u_x$

c) $$\frac{\partial J}{\partial u_w} = \frac{\partial}{\partial u_w} \left(-\log \frac{\exp(u_o^T \cdot v_c)}{\sum_{w \in vocab} \exp(u_w^T \cdot v_c)} \right)$$

$$= -\frac{\partial}{\partial u_w} \log(\exp(u_o^T \cdot v_c)) + \frac{\partial}{\partial u_w} \log \sum_w \exp(u_w^T \cdot v_c))$$

$$= -\frac{\partial}{\partial u_w} u_o^T \cdot v_c + \frac{\partial}{\partial u_w} \log\left(\sum_w \exp(u_w^T \cdot v_c)\right)$$

In case $w = O$:

$$\frac{\partial J}{\partial u_o} = -\frac{\partial}{\partial u_o} u_o^T \cdot v_c + \frac{\partial}{\partial u_o} \log\left(\sum_w \exp(u_w^T \cdot v_c)\right)$$

$$= -v_c + \frac{\partial}{\partial u_o} \log\left(\sum_w \exp(u_w^T \cdot v_c)\right)$$

we have:

$$\frac{\partial}{\partial u_o} \log\left(\sum_w \exp(u_w^T \cdot v_c)\right) = \frac{1}{\sum_w \exp(u_w^T \cdot v_c)} \cdot \frac{\partial}{\partial u_o} \sum_{x \in vocab} \exp(u_x^T \cdot v_c)$$

$$= \frac{1}{\sum_w \exp(u_w^T \cdot v_c)} \cdot \sum_x \frac{\partial}{\partial u_o} \exp(u_x^T \cdot v_c)$$

$$= \frac{1}{\sum_w \exp(u_w^T \cdot v_c)} \cdot \sum_x \exp(u_x^T \cdot v_c) \cdot \frac{\partial}{\partial u_o} u_x^T \cdot v_c$$

$$= \frac{1}{\sum_w \exp(u_w^T \cdot v_c)} \cdot \left(\exp(u_o^T \cdot v_c) \cdot \frac{\partial}{\partial u_o} \cdot u_o^T \cdot v_c + \sum_{x \neq o} \exp(u_x^T v_c) \cdot \frac{\partial}{\partial u_o} u_x^T \cdot v_c \right)$$

$$= \frac{1}{\sum_w \exp(u_w^T \cdot v_c)} \cdot (\exp(u_o^T \cdot v_c) \cdot v_c + 0)$$

$$= \frac{\exp(u_o^T \cdot v_c) \cdot v_c}{\sum_w \exp(u_w^T \cdot v_c)} = p(O=o \mid C=c) \cdot v_c = \hat{y}_o \cdot v_c$$

So: $$\frac{\partial J}{\partial u_o} = -v_c + \hat{y}_o \cdot v_c = v_c (1 - \hat{y}_o)$$

In case $w \neq O$

$$\frac{\partial J}{\partial u_w} = -\frac{\partial}{\partial u_w} u_o^T \cdot v_c + \frac{\partial}{\partial u_w} \log \left(\sum_w \exp(u_w^T \cdot v_c) \right)$$

$$= 0 + \frac{1}{\sum_w \exp(u_w^T \cdot v_c)} \cdot \frac{\partial}{\partial u_w} \sum_{x \in vocab} \exp(u_x^T \cdot v_c)$$

$$= \frac{1}{\sum_w \exp(u_w^T \cdot v_c)} \cdot \sum_x \frac{\partial}{\partial u_w} \exp(u_x^T \cdot v_c)$$

$$= \frac{1}{\sum_w \exp(u_w^T \cdot v_c)} \cdot \sum_x \exp(u_x^T \cdot v_c) \cdot \frac{\partial}{\partial u_w} (u_x^T \cdot v_c)$$

~~with each $w \neq O$:~~ $\cancel{\frac{\partial J}{\partial u_w} = \frac{1}{\sum_w \exp(u_w^T \cdot v_c)} \cdot \frac{\partial}{\partial u_w} (u_w^T v}$

$$= \frac{1}{\sum_w \exp(u_w^T \cdot v_c)} \cdot \Big(\exp(u_w^T \cdot v_c) \cdot \frac{\partial}{\partial u_w} (u_w^T \cdot v_c) + \sum_{x \neq w} \exp(u_x^T \cdot v_c) \cdot \frac{\partial}{\partial u_w} (u_x^T \cdot v_c) \Big)$$

$$= \frac{1}{\sum_w \exp(u_w^T \cdot v_c)} \cdot (\exp(u_w^T \cdot v_c) \cdot v_c + 0)$$

$$= \frac{\exp(u_w^T \cdot v_c)}{\sum_w \exp(u_w^T \cdot v_c)} \cdot v_c = p(O=w \mid C=c) \cdot v_c$$

$$= \hat{y}_w \cdot v_c$$

with $w \neq o$

d)

$$\frac{d}{dx} \sigma(x) = \frac{d}{dx} \frac{1}{1+e^{-x}} = \frac{d}{dx} (1+e^{-x})^{-1}$$

$$= -(1+e^{-x})^{-2} \cdot \frac{d}{dx} (1+e^{-x})$$

$$= -(1+e^{-x})^{-2} \cdot \frac{d}{dx} e^{-x} = (1+e^{-x})^{-2} \cdot e^{-x} \cdot \frac{d}{dx}(-x)$$

$$= (1+e^{-x})^{-2} \cdot e^{-x}$$

$$= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}}$$

$$= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}+1-1}{1+e^{-x}}$$

$$= \frac{1}{1+e^{-x}} \cdot \left(1 - \frac{1}{1+e^{-x}}\right)$$

$$= \sigma(x)\,(1 - \sigma(x))$$

c) $J_{neg\text{-}sample}(v_c, o, U) = -\log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^T \cdot v_c))$

$$\frac{\partial J}{\partial v_c} = \underbrace{\frac{-\partial}{\partial v_c} \log(\sigma(u_o^T \cdot v_c))}_{(1)} - \underbrace{\frac{\partial}{\partial v_c} \sum_k \log(\sigma(-u_k^T \cdot v_c))}_{(2)}$$

$$(1) = -\frac{1}{\sigma(u_o^T \cdot v_c)} \cdot \frac{\partial}{\partial v_c} \sigma(u_o^T \cdot v_c)$$

$$= -\frac{1}{\sigma(u_o^T \cdot v_c)} \cdot \sigma(u_o^T \cdot v_c) \cdot (1 - \sigma(u_o^T \cdot v_c)) \cdot \frac{\partial}{\partial v_c} u_o^T \cdot v_c$$

$$= (\sigma(u_o^T \cdot v_c) - 1) \cdot u_o$$

$$(2) = \sum_k \frac{\partial}{\partial v_c} \log(\sigma(-u_k^T \cdot v_c))$$

$$= \sum_k \frac{1}{\sigma(-u_k^T \cdot v_c)} \cdot \frac{\partial}{\partial v_c} \sigma(-u_k^T \cdot v_c)$$

$$= \sum_k \frac{1}{\sigma(-u_k^T \cdot v_c)} \cdot \sigma(-u_k^T \cdot v_c)(1 - \sigma(-u_k^T \cdot v_c)) \frac{\partial}{\partial v_c}(-u_k^T \cdot v_c)$$

$$= \sum_k (\sigma(-u_k^T \cdot v_c) - 1) \cdot u_k$$

So: $$\frac{\partial J}{\partial v_c} = (\sigma(u_o^T \cdot v_c) - 1) \cdot u_o - \sum_{k=1}^{K} (\sigma(-u_k^T \cdot v_c) - 1) \cdot u_k$$

$$\frac{\partial J}{\partial u_w} = \frac{-\partial}{\partial u_w} \log(\sigma(u_o^T . v_c)) - \frac{\partial}{\partial u_w} \sum_k \log(\sigma(-u_k^T . v_c))$$

$$= -(1-\sigma(u_o^T.v_c)) . \frac{\partial}{\partial u_w} u_o^T.v_c - \sum_k (1-\sigma(-u_k^T.v_c)) . \frac{\partial (-u_k^T.v_c)}{\partial u_w}$$

In case $w=o$, we have:

$$\frac{\partial J}{\partial u_o} = (\sigma(u_o^T.v_c)-1).v_c - 0$$ // this is because $o \notin \{w_1 ... w_k\}$ so the derivative $=0$

$$= (\sigma(u_o^T.v_c)-1).v_c$$

In case $w \neq o$, we have:

$$\frac{\partial J}{\partial u_w} = 0 - \sum_k (1-\sigma(-u_k^T.v_c)) . \frac{\partial}{\partial u_w}(-u_k^T.v_c)$$

$$= \sum_k (1-\sigma(-u_k^T.v_c)).v_c$$

// this because there is only one ~~w~~ $k=w$ so the ~~de~~ derivative of other $=0$

Conclusion:

- This loss function is much better ~~than~~ to compute more than the naive-softmax loss ~~cause~~ because it doesn't go through all the word in the vocabulary ~~so that~~ therefore, the computation is less expensive.

Problem 2:

a)

- Neural window-based models can be parallelized, but RNN models cannot

b)

- Use a network with fewer layers
- Increase L2 regularization weight

c) $-\log(0,2)$

d)

- True

e)

- True

Problem 4:

i) 1. Number of outputs: n

// because each output will correspond to an input at each time step.

2. $\hat{y}^{(t)}$ is the probability distribution over 4 categories including: person, organization, location, none

3. Each input will be a word in the sentence and will produce an output ~~at each~~ correspond to predicted category at each time step

ii) 1. Number of outputs: arbitrary

// because we don't know how many words the model will generate

2. $\hat{y}^{(t)}$ is the probability distribution over all words in vocabulary

3. Each input will be ~~the output of~~ the previous output and will produce a new output which is the next predicted word for the sentence at each time step

Được quét bằng CamScanner

Problem 3.

a) Yes, graph-based dependency can produce non-projective ~~[illegible]~~ dependency trees while transition-based can't

b) For example, in case $i=5$, the score will be

$$s(i,i) = h_i^T \cdot h_i$$

The result of this product will always be positive and higher than the scores of other edges

=> This can cause a problem that words can get linked to themselves

=> Incorrectly predict dependencies that involve a word and itself.

c) The drawback is that graph-based is slower and require more computational resources since it has to compute all the scores for all possible dependencies between all pairs of words in the sentence while transition-based ~~only~~ just simply build a tree from left to right or right to left incrementally