CS 224n : Assignment 2

3. Recurrent Neural Networks Language Modelling

a) i) Perplexity

$$PP^{(t)}\left(y^{(t)}, \hat{y}^{(t)}\right)$$

$$= \frac{1}{\bar{P}\left(x_{pred}^{(t+1)} \mid x^{(t)} \dots x^{(1)}\right)}$$

$$= \frac{1}{\sum_{j=1}^{V} y_j^{(t)} \cdot \hat{y}_j^{(t)}}$$

We know that $y^{(t)}$ is one-hot :

$$y^{(t)} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow k = (\delta_{ik})_i$$

$$PP^{(t)}\left(y^{(t)}, \hat{y}^{(t)}\right) = \frac{1}{1 \cdot \hat{y}_k^{(t)}}$$

$$= \frac{1}{\hat{y}_k^{(t)}}$$

Cross-entropy loss

$$= J^{(t)}(\theta) = -\sum_{j=1}^{|V|} \underbrace{y_j^{(t)}}_{\substack{=0 \ \forall j \neq k \\ =1 \text{ si } j=k}} \log \hat{y}_j^{(t)}$$

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

$$= -\log \frac{1}{PP^{(t)}}$$

$$J^{(t)}(\theta) = \log PP^{(t)}$$

$$\boxed{PP^{(t)}\left(y^{(t)}, \hat{y}^{(t)}\right) = \exp\left(J^{(t)}(\theta)\right)}$$

a) ii) We want to:

$$\text{minimize } \left[\prod_{t=1}^{T} PP\left(y^{(t)}, \hat{y}^{(t)}\right)\right]^{1/T}$$

$$\Big\downarrow \log.$$

$$\text{minimize } \frac{1}{T} \log\left(\prod_{t=1}^{T} PP\left(y^{(t)}, \hat{y}^{(t)}\right)\right)$$

$$\Big\downarrow =$$

$$\text{minimize } \frac{1}{T} \sum_{t=1}^{T} \log PP\left(y^{(t)}, \hat{y}^{(t)}\right)$$

$$\Big\downarrow =$$

$$\text{minimize } \frac{1}{T} \sum_{t=1}^{T} CE\left(y^{(t)}, \hat{y}^{(t)}\right)$$

a) iii)

Let's pick the next word uniformly.

$$PP^{(t)}\left(y^{(t)}, \hat{y}^{(t)}\right) = \frac{1}{\bar{P}\left(x_{pred}^{(t+1)} = x^{(t+1)} \mid x^{(t)} \ldots x^{(1)}\right)}$$

$$= \frac{1}{\frac{1}{|V|}} = |V|.$$

Then, the corresponding cross-entropy loss would be:

$$J^{(t)}(\theta) = \log PP^{(t)}$$
$$= \log |V|$$
$$= \log(10^4)$$

$$J^{(t)}(\theta) = 4.$$

b).

Let's define the following intermediary terms:

* $z^{(t)} = W_h h^{(t-1)} + W_e e^{(t)} + b_1$. (1)

So that: * $h^{(t)} = \sigma(z^{(t)})$ (2)

* $\theta^{(t)} = U \cdot h^{(t)} + b_2$ (3)

* $\hat{y}^{(t)} = \text{softmax}(\theta^{(t)})$. (4)

$J^{(t)}(\theta) = CE(y^{(t)}, \hat{y}^{(t)})$

$$\boxed{\frac{\partial J^{(t)}}{\partial U}}$$

$$\underbrace{\frac{\partial J^{(t)}}{\partial U}}_{\in \mathbb{R}^{|V| \times D_h}} = \frac{\partial J^{(t)}}{\partial \theta^{(t)}} \frac{\partial \theta^{(t)}}{\partial U}$$

$\partial J^{(t)} \in \mathbb{R}$

$\frac{\partial J^{(t)}}{\partial \theta^{(t)}} = \underbrace{\hat{y}^{(t)} - y^{(t)}}_{\in (|V|, 1)}$ (see assignment 1)

$\frac{\partial \theta^{(t)}}{\partial U} = \underbrace{h^{(t)}}_{(D_h, 1)}$.

$$\boxed{\frac{\partial J^{(t)}}{\partial U} = \delta_1^{(t)} \cdot h^{(t)T}. \quad \text{where } \delta_1^{(t)} = \hat{y}^{(t)} - y^{(t)}}$$

$$\underbrace{\frac{\partial J^{(t)}}{\partial e^{(t)}}}_{(d,1)} = \frac{\partial J^{(t)}}{\partial \theta^{(t)}} \frac{\partial \theta^{(t)}}{\partial h^{(t)}} \frac{\partial h^{(t)}}{\partial z^{(t)}} \frac{\partial z^{(t)}}{\partial e^{(t)}}$$

element-wise multiplication, as

$$= \underbrace{(\hat{y} - y)^T}_{(|V|,1)} \cdot \underbrace{U}_{(|V|,D_h)} \cdot \underbrace{\mathrm{Diag}(\sigma'(z^{(t)}))}_{(D_h,D_h)} ; \underbrace{W_e}_{(D_h, d)}$$

$= \cancel{W_e^T}$

$$= \underbrace{(\hat{y} - y)^T U \circ \sigma'(z^{(t)}) \quad W_e}_{(1,d)}.$$

Should be equal to its transpose

$$\boxed{\frac{\partial J^{(t)}}{\partial e^{(t)}} = (\underbrace{\delta_2^{(t)} W_e}_{(d,1)})^T}$$

Where $\delta_2^{(t)} = (\hat{y} - y)^T U \circ \underbrace{\sigma'(z^{(t)})}$

$= \sigma(z^{(t)}) \circ (1 - \sigma(z^{(t)}))$

$= h^{(t)} \circ (1 - h^{(t)})$

$$\boxed{\delta_2^{(t)} = \underbrace{\underbrace{(\hat{y} - y)^T}_{(1,|V|)} \underbrace{U}_{(|V|, D_h)}}_{(1, D_h)} \circ \underbrace{h^{(t)} \circ (1 - h^{(t)})}_{(|V|, D_h)}}$$

$$\underbrace{\left.\frac{\partial J^{(t)}}{\partial W_e}\right|_t}_{(D_h,\, d)} = \frac{\partial J}{\partial z^{(t)}} \cdot \frac{\partial z^{(t)}}{\partial W_e}$$

$$= \underbrace{\delta_2^{(t)}}_{(1,\, D_h)} \cdot \underbrace{e^{(t)}}_{(d,\, 1)}$$

$$\boxed{\left.\frac{\partial J^{(t)}}{\partial W_e}\right|_t = \delta_2^{(t)\,T}\; e^{(t)\,T}}$$

$$\underbrace{\left.\frac{\partial J^{(t)}}{\partial W_h}\right|_t}_{(D_h,\, D_h)} = \frac{\partial J}{\partial z^{(t)}} \; \frac{\partial z^{(t)}}{\partial W_h}$$

$$= \underbrace{\delta_2^{(t)}}_{(1,\, D_h)} \; \underbrace{h^{(t-1)}}_{(D_h,\, 1)}$$

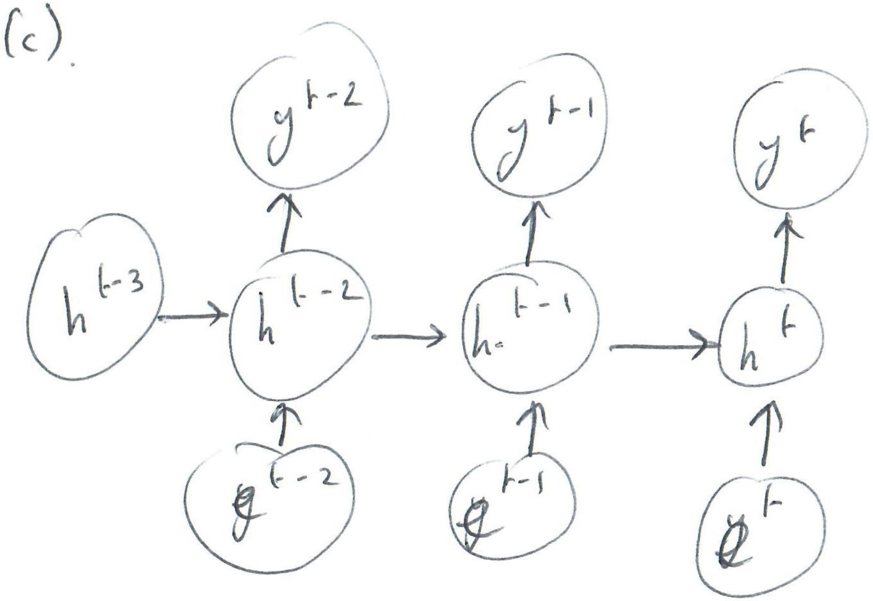
$$\boxed{\left.\frac{\partial J^{(t)}}{\partial W_h}\right|_t = \delta_2^{(t)\,T}\; h^{(t-1)\,T}}$$

$$\underbrace{\frac{\partial J}{\partial h^{(t-1)}}}_{(D_h,\, 1)} = \frac{\partial J}{\partial z^{(t)}} \; \frac{\partial z^{(t)}}{\partial h^{(t-1)}}$$

$$= \underbrace{\delta_2^{(t)}}_{(1,\, D_h)} \; \underbrace{W_h}_{(D_h,\, D_h)}$$

$$\boxed{\frac{\partial J}{\partial h^{(t-1)}} = \left(\delta_2^{(t)}\; W_h\right)^T}$$

(c)



$$\underbrace{\frac{\partial J^{(t)}}{\partial e^{(t-1)}}}_{d,1} = \frac{\partial J^{(t)}}{\partial h^{(t-1)}} \frac{\partial h^{(t-1)}}{\partial z^{(t-1)}} \frac{\partial z^{(t-1)}}{\partial e^{(t-1)}}$$

$$= \underbrace{\gamma^{(t-1)}}_{D_h,1} \cdot \underbrace{\circ h^{(t-1)} \circ (1-h^{(t-1)})}_{D_h,D_h} \underbrace{W_e}_{D_h,d}$$

$$\boxed{\frac{\partial J^{(t)}}{\partial e^{(t-1)}} = \left(\gamma^{(t-1)} \circ h^{(t-1)} \circ (1-h^{(t-1)}) \, W_e\right)^T}$$

$$\underbrace{\left.\frac{\partial J^{(t)}}{\partial W_e}\right|_{t-1}}_{(D_h,d)} = \frac{\partial J^{(t)}}{\partial h^{(t-1)}} \frac{\partial h^{(t-1)}}{\partial z^{(t-1)}} \left.\frac{\partial z^{(t-1)}}{\partial W_e}\right|_{t-1}$$

$$\underbrace{\gamma^{(t-1)}}_{(D_h,1)} \quad \underbrace{\circ h^{(t-1)} \circ (1-h^{(t-1)})}_{(D_h,D_h)} \quad \underbrace{e^{(t-1)}}_{(d,1)}$$

$$\boxed{\left.\frac{\partial J^{(t)}}{\partial W_e}\right|_{t-1} = \gamma^{(t-1)T} \circ h^{(t-1)} \circ (1-h^{(t-1)}) e^{(t-1)T}}$$

$$\underbrace{\left.\frac{\partial J^{(t)}}{\partial W_h}\right|_{t-1}}_{(D_h,D_h)} = \frac{\partial J^{(t)}}{\partial h^{(t-1)}} \frac{\partial h^{(t-1)}}{\partial z^{(t-1)}} \left.\frac{\partial z^{(t-1)}}{\partial W_h}\right|_{t-1}$$

$$= \underbrace{\gamma^{(t-1)}}_{(D_h,1)} \quad \underbrace{\circ h^{(t-1)} \circ (1-h^{(t-1)})}_{(D_h,D_h)} \quad = \underbrace{h^{(t-2)}}_{(D_h,1)}$$

$$\boxed{\left.\frac{\partial J^{(t)}}{\partial W_h}\right|_{t-1} = \gamma^{(t-1)T} \circ h^{(t-1)} (1-h^{(t-1)}) \, h^{(t-2)T}}$$

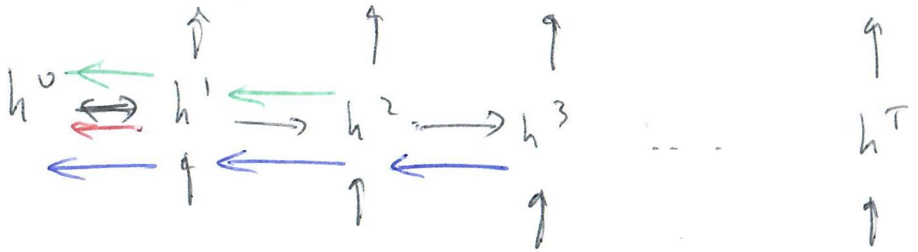
(d) Calculating complexity

- $\frac{\partial J^{(t)}}{\partial U}$: $D_h \cdot |V|$ operands
- $\frac{\partial J^{(t)}}{\partial e^{(t)}}$: $|V| \cdot D_h + D_h \cdot d$ operations
- $\left.\frac{\partial J^{(t)}}{\partial W_e}\right|_t$: $D_h \cdot d$ operations
- $\left.\frac{\partial J^{(t)}}{\partial W_h}\right|_t$: $D_h \cdot D_h$ operations
- $\frac{\partial J^{(t)}}{\partial h^{(t-1)}}$: $D_h \cdot D_h$ operations.

Which makes a total of:

$$\boxed{\mathcal{O}\left(D_h \cdot |V| + D_h \cdot d + D_h^2\right) \text{ operations}}$$

e) We could backpropagate all our losses at the same time, which would make us do:



1 pass for h^1

2 for h^2

$\vdots$

T for h^T

$$\Sigma = \frac{T(T+1)}{2} = \mathcal{O}(T^2)$$

But we could also wait for the passes coming from the ~~left~~ right to all be computed, before "packaging" them together and sending them on the left.

$\Rightarrow$ aggregating $\left.\begin{array}{l}\Leftarrow\\ \Leftarrow\\ \Leftarrow\end{array}\right\}$ flows into 1.

Hence, we'd need just T passes.

$$\boxed{\text{Complexity for } T \text{ words} = \mathcal{O}\left(T D_h(|V| + d + D_h)\right)}.$$

f). $|V|$ = vocabulary : $\sim 10^{4-6}$

d = dimension of our embeddings : $\sim 10^2$

D_h : dimension of hidden layer : $\sim 10^2$.

$$\Rightarrow \boxed{\mathcal{O}(T \cdot D_h \cdot |V|)}$$