FIT 3181 Assignment2

Xiaoyu sun
32950683

Question1:
CNN is good for analyzing visual patterns in fixed-size input, such as images.
RNN is good for sequential data such as text and video as it can maintain memory of previous inputs.

Example:
video classification is good for CNN combined with RNN as videos consists of sequential images. CNN can extract features from each images and RNN can process the sequence of the extracted features.

image captioning. CNN can be used to analyze the image to extract features and RNN can generate a sequence of words based on those features.

Question2:
The main advantages of transformer:
Transformer enables parallel computation so that it can make most use of GPU by removing the dependency between words. The parallel characteristic is achieved by each attention performing attention independently
Also, transformer is good at capturing long-range dependency

Question 3:
As shown in the below two handwritten screenshot

Question 3.1

$$\bar{h}_0 = Ux_0 + b$$

$$= \begin{bmatrix} 1 & 2 & 3 \\ -1 & 0 & 1 \\ 2 & -1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}$$

$$= \begin{bmatrix} -1 \\ -1 \\ 3 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}$$

$$= \begin{bmatrix} -1 \\ 0 \\ 2 \end{bmatrix}$$

$$h_0 = \tanh(\bar{h}_0)$$

$$= \begin{bmatrix} -0.762 \\ 0 \\ 0.964 \end{bmatrix}$$

$$h_1 = \tanh(Wh_0 + Ux_1 + b)$$

$$= \tanh\left(\begin{bmatrix} 1 & 0 & -1 \\ 2 & 1 & 0 \\ -1 & 2 & 1 \end{bmatrix} \begin{bmatrix} -0.762 \\ 0 \\ 0.964 \end{bmatrix} + \begin{bmatrix} 1 & 2 & 3 \\ -1 & 0 & 1 \\ 2 & -1 & 0 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} \right)$$

$$= \tanh\left(\begin{bmatrix} -1.726 \\ -1.524 \\ 1.726 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ -3 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} \right)$$

$$= \tanh\left(\begin{bmatrix} -0.726 \\ 0.476 \\ -2.274 \end{bmatrix} \right) = \begin{bmatrix} -0.621 \\ 0.443 \\ -0.979 \end{bmatrix}$$

$$h_2 = \tanh(Wh_1 + Ux_2 + b)$$

$$= \tanh\left(\begin{bmatrix} 1 & 0 & -1 \\ 2 & 1 & 0 \\ -1 & 2 & 1 \end{bmatrix} \begin{bmatrix} -0.621 \\ 0.443 \\ -0.979 \end{bmatrix} + \begin{bmatrix} 1 & 2 & 3 \\ -1 & 0 & 1 \\ 2 & -1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} \right)$$

$$= \tanh\left(\begin{bmatrix} 0.358 \\ -0.799 \\ 0.528 \end{bmatrix} + \begin{bmatrix} 3 \\ -1 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} \right)$$

$$= \tanh\left(\begin{bmatrix} 3.358 \\ -0.799 \\ 0.528 \end{bmatrix} \right) = \begin{bmatrix} 0.998 \\ -0.663 \\ 0.484 \end{bmatrix}$$

Question 3.2

$$\hat{y}_0 = \text{softmax}(Vh_0 + c)$$

$$= \text{softmax}\left(\begin{bmatrix} 2 & -1 & 0 \\ 1 & 1 & -1 \\ -1 & 2 & 1 \end{bmatrix} \begin{bmatrix} -0.762 \\ 0 \\ 0.964 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \right)$$

$$= \text{softmax}\left(\begin{bmatrix} -1.524 \\ -1.726 \\ 1.726 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \right)$$

$$= \text{softmax}\left(\begin{bmatrix} -0.524 \\ -0.726 \\ 1.726 \end{bmatrix} \right) = \begin{bmatrix} 0.0885 \\ 0.0723 \\ 0.839 \end{bmatrix}$$

$$\hat{y}_1 = \text{softmax}(Vh_1 + c)$$

$$= \text{softmax}\left(\begin{bmatrix} 2 & -1 & 0 \\ 1 & 1 & -1 \\ -1 & 2 & 1 \end{bmatrix} \begin{bmatrix} -0.621 \\ 0.443 \\ -0.979 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \right)$$

$$= \text{softmax}\left(\begin{bmatrix} -1.685 \\ 0.801 \\ 0.528 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \right)$$

$$= \text{softmax}\left(\begin{bmatrix} -0.685 \\ 1.801 \\ 0.528 \end{bmatrix} \right) = \begin{bmatrix} 0.0611 \\ 0.734 \\ 0.205 \end{bmatrix}$$

$$\hat{y}_2 = \text{softmax}(Vh_2 + c)$$

$$= \text{softmax}\left(\begin{bmatrix} 2 & -1 & 0 \\ 1 & 1 & -1 \\ -1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 0.998 \\ -0.663 \\ 0.484 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \right)$$

$$= \text{softmax}\left(\begin{bmatrix} 2.659 \\ -0.149 \\ -1.84 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \right)$$

$$= \text{softmax}\begin{pmatrix} 3.659 \\ 0.851 \\ -1.84 \end{pmatrix} = \begin{bmatrix} 0.939 \\ 0.0567 \\ 0.00384 \end{bmatrix}$$

Question 3.3

$l_0 = CE(y_0, \hat{y}_0)$

$= -1 \cdot \log(0.839)$

$= 0.0762$ (base 10)

$l_1 = CE(y_1, \hat{y}_1)$

$= -\log(0.734)$

$= 0.134$ (base 10)

$l_2 = CE(y_2, \hat{y}_2)$

$= -\log(0.00384)$

$= 2.416$ (base 10)

$L = \frac{1}{3}(l_1 + l_2 + l_3)$

$= 0.875$

Question 3.4

$$\frac{\partial L}{\partial h_1} = \frac{1}{3} \Sigma_t \frac{\partial l_t}{\partial h_1} = \frac{1}{3}\Big($$

$$\underbrace{\frac{\partial l_2}{\partial \hat{y}_2} \times \frac{\partial \hat{y}_2}{\partial h_2}}_{(\hat{y}_2 - y_2)V} \times \underbrace{\frac{\partial h_2}{\partial \bar{h}_2}}_{diag(1-h_2^2)} \times \underbrace{\frac{\partial \bar{h}_2}{\partial h_1}}_{W}$$

$$+ \underbrace{\frac{\partial l_1}{\partial \hat{y}_1} \times \frac{\partial \hat{y}_1}{\partial h_1}}_{(\hat{y}_1 - y_1)V}$$

$$+ 0 \Big)$$

$$= \frac{1}{3}\left(\left(\begin{bmatrix} 0.939 \\ 0.0567 \\ 0.00384 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right) \begin{bmatrix} 2 & -1 & 0 \\ 1 & 1 & -1 \\ -1 & 2 & 1 \end{bmatrix} diag\left(\begin{bmatrix} 1-0.998^2 \\ 1-(0.663)^2 \\ 1-(0.484)^2 \end{bmatrix} \right) \begin{bmatrix} 1 & 0 & -1 \\ 2 & 1 & 0 \\ -1 & 2 & 1 \end{bmatrix} + \left(\begin{bmatrix} 0.0885 \\ 0.0723 \\ 0.839 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right) \begin{bmatrix} 2 & -1 & 0 \\ 1 & 1 & -1 \\ -1 & 2 & 1 \end{bmatrix} \right)$$

Question 3.5

$$\frac{\partial L}{\partial V} = \frac{1}{3} \Sigma_t \frac{\partial l_t}{V} = \frac{1}{3}\Big($$

$$\frac{\partial l_2}{\partial \hat{y}_2} \times \frac{\partial \hat{y}_2}{\partial V}$$

$$+$$

$$\frac{\partial l_1}{\partial \hat{y}_1} \times \frac{\partial \hat{y}_1}{\partial V}$$

$$+$$

$$\frac{\partial l_0}{\partial \hat{y}_0} \times \frac{\partial \hat{y}_0}{\partial V} \Big)$$

$$= \frac{1}{3}\left[(\hat{y}_2 - y_2) h_2 + (\hat{y}_1 - y_1) h_1 + (\hat{y}_0 - y_0) h_0 \right]$$

plug in number and calculate

Question 3.6

$$\frac{\partial L}{\partial U} = \frac{1}{3} \Sigma_t \frac{\partial l_t}{U} = \frac{1}{3}\Big($$

$$\frac{\partial l_2}{\partial \hat{y}_2} \times \frac{\partial \hat{y}_2}{\partial h_2} \times \frac{\partial h_2}{\partial \bar{h}_2} \times \frac{\partial \bar{h}_2}{\partial U}$$

$$+$$

$$\frac{\partial l_1}{\partial \hat{y}_1} \times \frac{\partial \hat{y}_1}{\partial h_1} \times \frac{\partial h_1}{\partial \bar{h}_1} \times \frac{\partial \bar{h}_1}{\partial U}$$

$$+$$

$$\frac{\partial l_0}{\partial \hat{y}_0} \times \frac{\partial \hat{y}_0}{\partial h_0} \times \frac{\partial h_0}{\partial \bar{h}_0} \times \frac{\partial \bar{h}_0}{\partial U} \Big)$$

$$= \frac{1}{3}\Big((\hat{y}_2 - y_2) V diag(1-h_2^2) X_2$$

$$+$$

$$(\hat{y}_1 - y_1) V diag(1-h_1^2) X_1$$

$$+$$

$$(\hat{y}_0 - y_0) V diag(1-h_0^2) X_0 \Big)$$