Name: Bryan Jun Kit Wong
Student ID: 32882424

Question 1

CNNs and RNNs are combined in scenarios where the input data has both spatial and temporal dimensions. CNNs are adept at processing spatial information, such as images, while RNNs excel at handling sequential data, like time series or sentences. Video captioning is an example where CNN can be used to extract features from individual video frames whereas RNN can be used to generate captions to describe the video. Another example is speech recognition where CNN can be used to extract features from audio waveforms whereas RNN can be used to recognise words in the speech.

(95 words)

Question 2

One main advantage of Transformers over RNN is that it can handle long-range dependencies. This is done using a self-attention mechanism, where each token considers the entire input sentence when making predictions. On the other hand, RNN will struggle to do so due to the vanishing gradient problem. Another advantage Transformers have over RNN is its ability to process the entire input sequence in parallel. RNNs are sequential models, resulting in only being able to process input data one step at a time. Transformers can process all words in a sequence simultaneously, resulting in significantly faster training and inference times. Furthermore, Transformers have positional embeddings. As Transformers do not rely on past hidden states to capture dependencies of words, there is no risk of "forgetting" past information. In contrast, RNN uses recurrence, resulting in a dependency in past hidden states where it has a risk of "forgetting" past information.

(149 words)

Question 3

Question 3.1

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

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

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

$$h_0 = tanh\begin{pmatrix} -1 \\ 0 \\ 2 \end{pmatrix}$$

$$= \begin{pmatrix} -0.76159416762 \\ 0 \\ 0.96402758 \end{pmatrix}$$

Name: Bryan Jun Kit Wong
Student ID: 32882424

$$\begin{aligned}\bar{h}_1 &= Wh_0 + \ Ux_1 + b \\ &= \begin{pmatrix} 1 & 0 & -1 \\ 2 & 1 & 0 \\ -1 & 2 & 1 \end{pmatrix}\begin{pmatrix} -0.762 \\ 0 \\ 0.964 \end{pmatrix} + \begin{pmatrix} 1 & 2 & 3 \\ -1 & 0 & 1 \\ 2 & -1 & 0 \end{pmatrix}\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} \\ &= \begin{pmatrix} -0.726 \\ 0.477 \\ -2.274 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}h_1 &= tanh\begin{pmatrix} -0.726 \\ 0.477 \\ -2.274 \end{pmatrix} \\ &= \begin{pmatrix} -0.62037946 \\ 0.44368657 \\ -0.97906084 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}\bar{h}_2 &= Wh_1 + \ Ux_2 + b \\ &= \begin{pmatrix} 1 & 0 & -1 \\ 2 & 1 & 0 \\ -1 & 2 & 1 \end{pmatrix}\begin{pmatrix} -0.62 \\ 0.444 \\ -0.979 \end{pmatrix} + \begin{pmatrix} 1 & 2 & 3 \\ -1 & 0 & 1 \\ 2 & -1 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} \\ &= \begin{pmatrix} 3.359 \\ -0.797 \\ 0.529 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}h_2 &= tanh\begin{pmatrix} 3.359 \\ -0.797 \\ 0.529 \end{pmatrix} \\ &= \begin{pmatrix} 0.99758347 \\ -0.66239687 \\ 0.48438043 \end{pmatrix}\end{aligned}$$

Question 3.2

$$\begin{aligned}\hat{y}_0 &= softmax(Vh_0 + c) \\ &= softmax\begin{pmatrix} -0.523 \\ -0.726 \\ 1.726 \end{pmatrix} \\ &= \begin{pmatrix} 0.08854891 \\ 0.07232151 \\ 0.83912957 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}\hat{y}_1 &= softmax(Vh_1 + c) \\ &= softmax\begin{pmatrix} -0.684 \\ 1.802 \\ 0.529 \end{pmatrix}\end{aligned}$$

$$= \begin{pmatrix} 0.06102426 \\ 0.73368879 \\ 0.20528695 \end{pmatrix}$$

$$\hat{y}_2 = softmax(Vh_2 + c)$$
$$= softmax\begin{pmatrix} 3.658 \\ 0.851 \\ -1.838 \end{pmatrix}$$
$$= \begin{pmatrix} 0.93940335 \\ 0.05674045 \\ 0.00385621 \end{pmatrix}$$

Question 3.3

$$l_0 = CE\left(y_0, \hat{y}_0\right)$$
$$= -[log(0.83912957)]$$
$$= 0.17539015058878907$$
$$\approx 0.175$$

$$l_1 = CE\left(y_1, \hat{y}_1\right)$$
$$= -[log(0.73368879)]$$
$$= 0.3096703321031836$$
$$\approx 0.310$$

$$l_2 = CE\left(y_2, \hat{y}_2\right)$$
$$= -[log(0.00385621)]$$
$$= 5.558070443135849$$
$$\approx 5.56$$

$$L = \frac{1}{3}\ (l_0 +\ l_1 +\ l_2)$$
$$= \frac{1}{3}\ (6.043130925827821)$$
$$= 2.0143769752759404$$
$$\approx 2.01$$

Question 3.4

$$\frac{\partial L}{\partial h_1} = \sum \frac{\partial l_t}{\partial h_1}$$

$$
\begin{aligned}
&= \frac{\partial l_1}{\partial h_1} + \frac{\partial l_2}{\partial h_1} \\
&= \left(\frac{\partial L}{\partial \hat{y}_1} * \frac{\partial \hat{y}_1}{\partial h_1}\right) + \left(\frac{\partial L}{\partial \hat{y}_2} * \frac{\partial \hat{y}_2}{\partial h_2} * \frac{\partial h_2}{\partial \bar{h}_2} * \frac{\partial \bar{h}_2}{\partial h_1}\right) \\
&= \begin{pmatrix} -2.7565646 \\ -3.1419743 \\ -0.34840563 \end{pmatrix} \\
&\approx \begin{pmatrix} -2.76 \\ -3.14 \\ -0.348 \end{pmatrix}
\end{aligned}
$$

Question 3.5

$$
\begin{aligned}
\frac{\partial L}{\partial V} &= \left(\frac{\partial L}{\partial \hat{y}_0} * \frac{\partial \hat{y}_0}{\partial V}\right) + \left(\frac{\partial L}{\partial \hat{y}_1} * \frac{\partial \hat{y}_1}{\partial V}\right) + \left(\frac{\partial L}{\partial \hat{y}_2} * \frac{\partial \hat{y}_2}{\partial V}\right) \\
&= \begin{pmatrix} 0.83183672 & -0.5951822 & 0.48064573 \\ 0.16673769 & -0.1557434 & 0.35793877 \\ -0.99857442 & 0.7509256 & -0.8385845 \end{pmatrix} \\
&\approx \begin{pmatrix} 0.832 & -0.595 & 0.481 \\ 0.167 & -0.156 & 0.358 \\ -0.999 & 0.751 & -0.839 \end{pmatrix}
\end{aligned}
$$

Question 3.6

$$
\begin{aligned}
\frac{\partial L}{\partial V} &= \left(\frac{\partial L}{\partial \hat{y}_0} * \frac{\partial \hat{y}_0}{\partial h_0} * \frac{\partial h_0}{\partial \bar{h}_0} * \frac{\partial \bar{h}_0}{\partial U}\right) + \left(\frac{\partial L}{\partial \hat{y}_1} * \frac{\partial \hat{y}_1}{\partial h_1} * \frac{\partial h_1}{\partial \bar{h}_1} * \frac{\partial \bar{h}_1}{\partial U}\right) + \left(\frac{\partial L}{\partial \hat{y}_2} * \frac{\partial \hat{y}_2}{\partial h_2} * \frac{\partial h_2}{\partial \bar{h}_2} * \frac{\partial \bar{h}_2}{\partial U}\right) \\
&= \begin{pmatrix} 0.69636995 & -0.67760952 & 0 \\ -1.83694602 & -1.62310579 & 0 \\ -0.43115442 & -0.37476464 & 0 \end{pmatrix} \\
&\approx \begin{pmatrix} 0.696 & -0.678 & 0 \\ -1.84 & -1.62 & 0 \\ -0.431 & -0.375 & 0 \end{pmatrix}
\end{aligned}
$$