CSE 493 G1/ 599 G1
Deep Learning
Autumn 2024 Quiz 3

November 8, 2024

Full Name: _____

UW Net ID: _____

Question		Score
True/False	(4 pts)	
Multiple Choice	(8 pts)	
Short Answer	(8 pts)	
Total	(20 pts)	

Welcome to the CSE 493 G1 Quiz 3!

- The exam is 30 min and is **double-sided**.
- No electronic devices are allowed.

I agree to uphold the University of Washington Student Conduct Code during this exam.

Signature: _____ Date: _____

Good luck!

This page is left blank for scratch work only. DO NOT write your answers here.

1 True / False (4 points) - Recommended 4 Minutes

Fill in the circle next to True or False, or fill in neither. Fill it in completely like this: ●. No explanations are required.

Scoring: Correct answer is worth 1 points.

1.1 One-hot vector encodings of words must be the length of the entire vocabulary while learned embeddings are generally much smaller.

- ◯ True
- ◯ False

1.2 One advantage that RNNs have over transformers is that they can process any sequence length while transformers can only take in sequences of a fixed length.

- ◯ True
- ◯ False

1.3 In self-attention, the attention weights for a given position are computed using only the tokens that come before it in the sequence

- ◯ True
- ◯ False

1.4 You are doing general attention. The sequence length of the queries is N and the sequence length of the keys and values is M. The sum of all of the values in the attention map is M.

- ◯ True
- ◯ False

2 Multiple Choices (8 points) - Recommended 8 Minutes

Fill in the circle next to the letter(s) of your choice (like this: ●). No explanations are required. **Choose ALL options that apply.**

Each question is worth 4 points and the answer may contain one or more options. Selecting all of the correct options and none of the incorrect options will get full credits. For questions with multiple correct options, each incorrect or missing selection gets a 2-point deduction (up to 4 points).

2.1 Which of the following statements about Recurrent Neural Networks (RNNs) are true?

◯ A: The vanishing gradient problem in RNNs occurs because the same weights are repeatedly multiplied during backpropagation through time

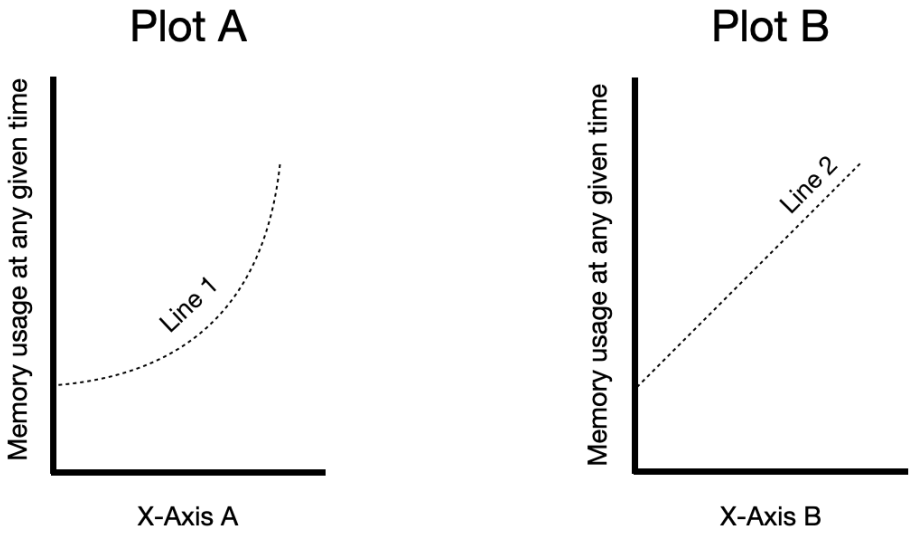
◯ B: LSTMs maintain two hidden states (cell state and hidden state) while basic RNNs have only one hidden state

◯ C: When processing a sequence in an RNN, the computation at each timestep must be done sequentially, making RNNs difficult to parallelize

◯ D: LSTMs help mitigate the vanishing gradient problem by creating direct pathways for gradients to flow through their cell state

2.2 We are shown two plots that compare memory usage patterns, but we're missing some key information about what they represent. Here's what we know:

- We have two plots (A and B) of memory usage
- One plot shows memory usage vs sequence length
- One plot shows memory usage vs size of hidden state
- We don't know which plot represents which relationship
- On each plot, we show either a transformer model and/or an RNN model. Each plot could show the same model or each show a different model.



◯ A: X-axis A could only be sequence length

◯ B: There is not enough information to know if X-axis A is sequence length or size of hidden state

◯ C: Line 1 could only be the transformer based model.

◯ D: There is not enough information to know if line 1 is the RNN or Transformer

◯ E: Line 2 could only be the RNN based model.

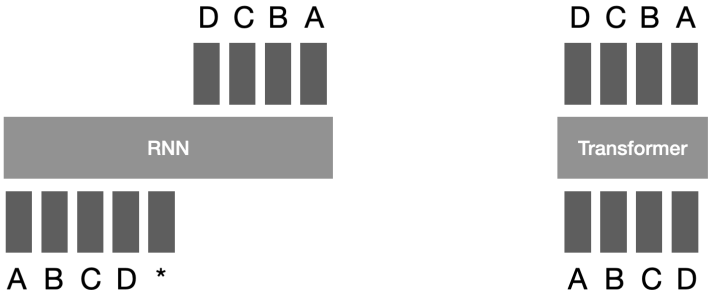
◯ F: There is not enough information to know if line 2 is the RNN or Transformer

3 Short Answers (8 points) - Recommended 8 Minutes

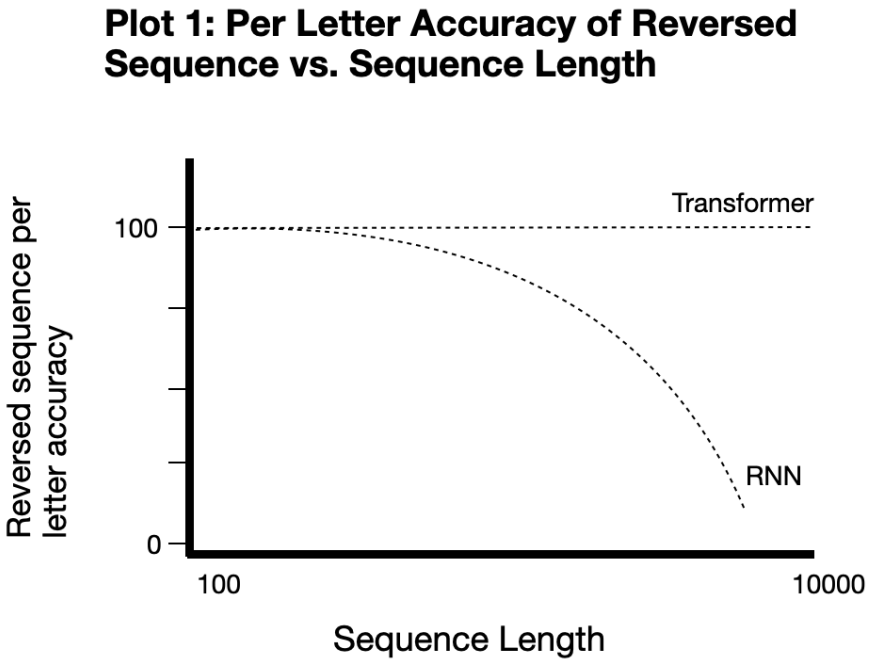
Please make sure to write your answer only in the provided space.

3.1 RNN vs Transformer

You are attempting to build a model to complete the task of reversing a sequence of letters. You are considering two options, a transformer based model and an RNN. For the RNN, you input the sequence in the first x steps, then a token indicating the end of the sequence, and then on next x steps the model outputs the reversed sequence. For the transformer model, you input the sequence of length x with positional embeddings. Then it outputs a sequence of length x, corresponding to the reversed input.



You train both models on a variety of sequences. You monitor the gradients and find there are no vanishing or exploding gradients. You evaluate the models using per letter accuracy (so what percent of the letters the model get correct, rather than what percent of sequences the model gets entirely correct). When you evaluate the models on sequences of different length, you find the following results:



3.1.1 Accuracy (3 points)

Explain the trends you see in each line in Plot 1.

3.1.2 Smaller RNN (2 points)

You create another RNN with a hidden dimension that is half the size of your first RNN. Plot the accuracy of reversed sequence vs sequence length of this smaller RNN on Plot 1. Label this line "RNN 2".

3.1.3 Masked Attention Transformer (3 points)

You create another transformer. You now use the below masked self-attention, identical to what we discussed in lecture. Plot the accuracy of reversed sequence vs sequence length of this masked transformer on Plot 1. Label this line "masked".