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

Feb 16, 2024

Full Name: _____

UW Net ID: _____

Question		Score
True/False	(5 pts)	
Multiple Choice	(8 pts)	
Short Answer	(9 pts)	
Total	(22 pts)	

Welcome to the CSE 493 G1 Quiz 3!

- The exam is 20 min and is **double-sided**.
- No electronic devices are allowed.

I understand and agree to uphold the University of Washington Honor 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 (5 points) - Recommended 5 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 Skip-gram algorithm makes use of global co-occurrence statistics.

◯ True

◯ False

1.2 Word embeddings generated by Skip-gram model are sparse representations of words in a high-dimensional space.

◯ True

◯ False

1.3 RNNs with attention mechanisms enable the model to focus on specific parts of the input sequence while making predictions, improving its ability to capture long-range dependencies.

◯ True

◯ False

1.4 Transformers are inherently more memory-efficient than RNNs when processing long sequences due to their ability to capture long-range dependencies in parallel.

◯ True

◯ False

1.5 Transformers are primarily used for sequence-to-sequence tasks and are not a good choice for tasks involving non-sequential data.

◯ True

◯ False

2 Multiple Choices (8 points) - Recommended 6 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 Select all statements that are true about recurrent neural networks.

- ◯ A: Training recurrent neural networks can be affected by the exploding gradient problem.
- ◯ B: Gradient clipping might help if your RNN is troubled by vanishing gradients.
- ◯ C: Unlike standard feedforward networks, recurrent neural networks can learn from sequences of variable length.
- ◯ D: Unlike traditional RNNs, LSTMs do not suffer from the exploding gradient problem.
- ◯ E: None of the above.

2.2 Which of the following statements are true about batch normalization?

- ◯ A: Batch Normalization is helpful in regularizing neural networks and reducing overfitting.
- ◯ B: BatchNorm parameters are recalculated based on batch statistics during inference.
- ◯ C: BatchNorm helps mitigate the vanishing gradient problem during training by maintaining more stable activations throughout the network.
- ◯ D: Batch normalization mitigates the effects of poor weight initialization and allows the network to initialize our weights to smaller values close to zero.
- ◯ E: BatchNorm requires the computation of mean and standard deviation across the entire dataset before training can begin.

3 Short Answers (9 points) - Recommended 9 Minutes

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

Consider the task of sentiment classification in movie reviews using two distinct neural network architectures depicted in Figure 1. "Architecture 1" relies on LSTM, while "Architecture 2" leverages self-attention mechanism **(without using positional encoding or masking)**. Both models are trained to classify movie reviews as either **positive (1)** or **negative (0)**. Both models have similar embedding layers, resulting in identical word representations. These representations undergo further processing, with the resulting tokens passed through sigmoid layers to predict the likelihood of a review being "positive". Subsequently, these probabilities are rounded to binary decisions (0 or 1) to make one prediction after each new word of the sentence. The final sentiment classification is determined by aggregating all predictions through majority voting. For instance, the LSTM-based model has 5 predictions consisting of four "0"s and one "1". Therefore the final label is negative (0).

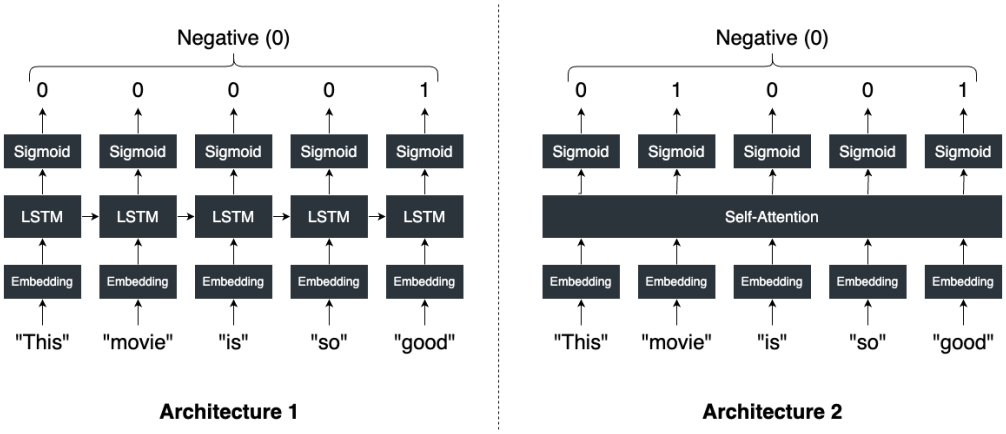

Figure 1: The 2 architectures used for sentiment classification.

Given this setup, consider the following scenarios:

1. The LSTM-based architecture labels the review *"This movie is so good"* as negative due to the output (00001) containing four "0"s and one "1" as shown in Figure 1 (left). Predict how this model will classify the review *"This movie is so bad"*, justifying your answer with appropriate reasoning.

   (a) Positive (1)

   (b) Negative (0)

   (c) Insufficient information

2. Architecture 2, based on self-attention, also categorizes the same review, *"This movie is so good"* as negative because the output (01001) consists of three "0"s and two "1"s as shown in Figure 1 (right). Predict the sentiment classification for the review *"This movie is so bad"* using this model, justifying your answer with appropriate reasoning.

   (a) Positive (1)

   (b) Negative (0)

   (c) Insufficient information

3. The self-attention architecture classifies the following reviews as both positive: *"The movie was not good; it was bad."* and *"The movie was not bad; it was good."* Analyze the underlying reason for this mis-classification and propose a potential solution.