Quiz 4 - Results

Attempt 1 of 1

Written Mar 31, 2024 12:31 AM - Mar 31, 2024 12:55 AM

Released Mar 31, 2024 12:00 PM

Attempt Score **21.4 / 25 - B**

Overall Grade (Highest Attempt) **21.4 / 25 - B**

Question 1 — 1.5 / 1.5 points

Which of the following is true about contrastive loss?

- ✓ ☐ It aims to minimize the inter class distance
- ✓ ☐ It aims to maximize the inter class distance
- ✓ ☐ It aims to minimize the intraclass distance.
- ✓ ☐ It aims to maximize the intraclass distance

Question 2 — 0 / 2 points

Select all that is true about multi-head mechanisms in Transformer models.

- ✗ ☐ Both self-attention and encoder-decoder attention are used in encoder
- ➡ ✗ ☐ encoder-decoder attention is used only in decoder
- ✗ ☐ Self-attention is used in only encoder
- ➡ ✗ ☐ Both self-attention and encoder-decoder attention are used in decoder

Question 3 — 1.5 / 1.5 points

Name the three vectors used to calculate self-attention in Transformers

Answer for blank # 1: Query ✓(33.33 %)

Answer for blank # 2: Key ✔(33.33 %)

Answer for blank # 3: Value ✔(33.33 %)

Question 4 **1 / 1 point**

Which of the following best describes an LSTM?

- ○ LSTM networks are an extension for recurrent neural networks, which basically involves lesser computation. Therefore it is well suited for environments with less computational resources.
- ✔ ○ LSTM networks are an extension for recurrent neural networks, which basically extends their memory. Therefore it is well suited to learn from important experiences that have very long time lags in between.

Question 5 **1.5 / 1.5 points**

What are the two gates in a GRU?

Answer for blank # 1: Update gate ✔(50 %)

Answer for blank # 2: Reset gate ✔(50 %)

Question 6 **1 / 1 point**

Suppose you download a pre-trained word embedding which has been trained on a huge corpus of text. You then use this word embedding to train an RNN for a language task of recognizing if someone is happy from a short snippet of text, using a small training set.

x (input text)	y (happy?)
I'm feeling wonderful today!	1
I'm bummed my cat is ill.	0
Really enjoying this!	1

Then even if the word "ecstatic" does not appear in your small training set, your RNN might reasonably be expected to recognize "I'm ecstatic" as deserving a label y=1.

- ✔ ○ True
- ○ False

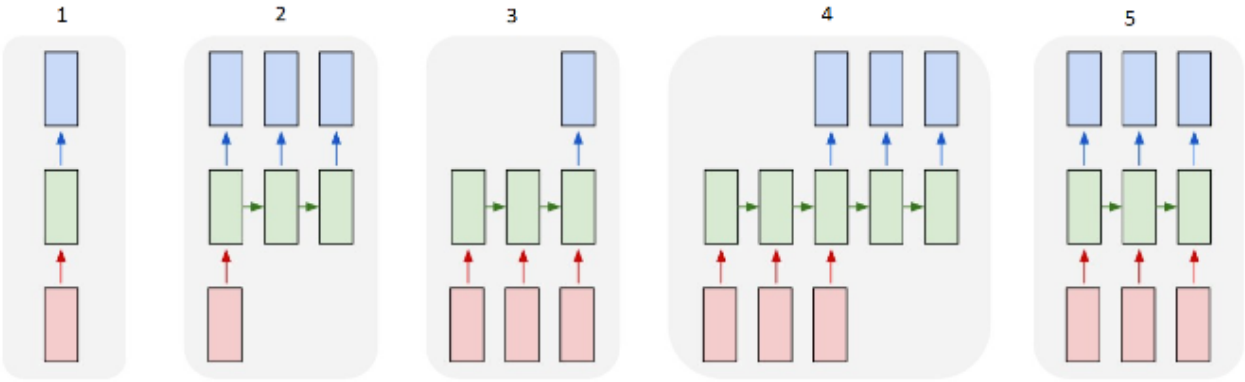
Question 7 **1.5 / 1.5 points**

Which of these equations do you think should hold for a good word embedding? (Check all that apply)

- ✔ ☐ $e_{boy} - e_{brother} \approx e_{girl} - e_{sister}$
- ✔ ☐ $e_{boy} - e_{girl} \approx e_{brother} - e_{sister}$
- ✔ ☐ $e_{boy} - e_{girl} \approx e_{sister} - e_{brother}$
- ✔ ☐ $e_{boy} - e_{brother} \approx e_{sister} - e_{girl}$

Question 8

2 / 2 points



Fill in the blanks based on which type of RNN is specified in the above image. You need to write the RNN number (1, 2,...). If there are more than two RNNs for the answer you need to put comma separared values (like 1, 5).

Many-to-many

___4, 5___ ✔(25 %)

One-to-one

___1___ ✔(25 %)

One-to-many

___2___ ✔(25 %)

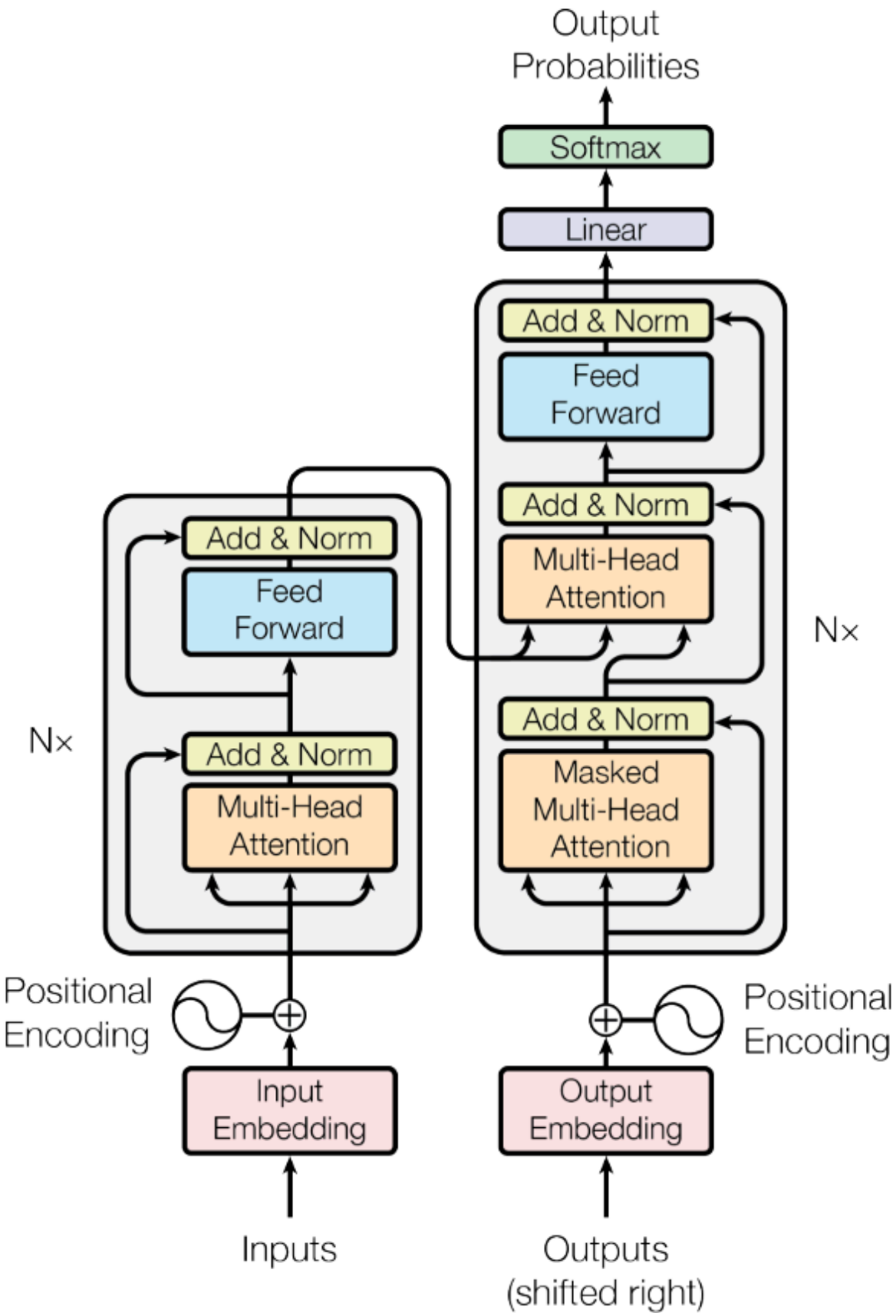
Many-to-one

___3___ ✔(25 %)

Question 9

2 / 2 points

Following is the architecture of a transformer network



What information does the decoder take from the encoder for its second block of multi-head mechanism?

✔ ☐ query

- ✓ key
- ✓ input sequence
- ✓ value

Question 10 **0.9 / 1.5 points**

Which of the following are challenges that can't be tackled efficiently using Multi-layered Perceptron?

- ✓ Classification prediction problems
- ➡ ✓ Text data
- ➡ ✗ Image data
- ➡ ✓ Time series data
- ✓ Regression prediction problems

Question 11 **1.5 / 1.5 points**

Which of the following are true about Transformers?

- ✓ Residuals are applied to attention layers to combat the vanishing gradients problem.
- ✓ Weight sharing occurs between different encoders in a transformer network.
- ✓ Transformers consist of a single encoder and decoder block.
- ✓ Unlike RNNs, transformers process the entire input at once.

Question 12 **1 / 1 point**

In self-attention mechanism (select all that apply)

- ✓ Given a word, its neighboring words are used to compute its context by taking a weighted sum of the word values to map the Attention related to that given word.
- Given a word, its neighboring words are used to compute its context by selecting the lowest of those word values to map the Attention related to that given word.
- Given a word, its neighboring words are used to compute its context by selecting the highest of the word values to map the Attention related to that given word.

- ○ Given a word, its neighboring words are used to compute its context by taking a simple average of the word values to map the Attention related to that given word.

Question 13 **1 / 1 point**

In RNNs, the weights are shared among all the nodes in a layer, but the bias for each node is different.

- ○ True
- ✓ ○ False

Question 14 **1.5 / 1.5 points**

What are some of the advantages of Parameter sharing and Unrolling in recurrent neural networks?

- ✓ ☐ Allows modeling arbitrary sequence lengths
- ✓ ☐ Keeps number of parameter in check
- ✓ ☐ Provides better prediction for every type of problem
- ✓ ☐ Flexible in terms of inputs and outputs
- ✓ ☐ Prevents exploding and vanishing gradients

Question 15 **1.5 / 1.5 points**

What are the three gates in an LSTM unit?

Answer for blank # 1: Input gate ✓(33.33 %)

Answer for blank # 2: Forget gate ✓(33.33 %)

Answer for blank # 3: Output gate ✓(33.33 %)

Question 16 **1 / 1 point**

Which of these tasks can be seen as a sequence-to-sequence problem?

- ○ Translating a text from Chinese into English
- ○ Answering questions about a document
- ✓ ○ All of them
- ○ Writing short reviews of long documents

Question 17 **1 / 1 point**

Which of the following best describes the basic concept of Recurrent Neural Network?

- ✔ ◯ Use previous inputs to find the next output according to the training set.
- ◯ Use recurrent features from dataset to find the best answers
- ◯ Use loops between the most important features to predict next output
- ◯ Use a loop between inputs and outputs in order to achieve the better prediction

Question 18 **0 / 1 point**

In the Teacher-Student model for Pseudo-labeling, we select top-K examples for each target label. For each image, we retain only the classes associated with the P highest scores.
Which of the following is True?

- ➡ ◯ The reason for choosing P > 1 is that it is difficult to identify accurately under-represented concepts, or some may occulted by more prominent co-occurring concepts
- ✖ ◯ All of these
- ◯ Probability of introducing False Negatives increases as we increase the value of K.
- ◯ Increasing the value of K will always lead to better performance because we are selecting more examples for each target label.

Done