Indian Institute of Technology Kharagpur
Computer Sciene & Engineering Department
CS60010 Spring 2023
Quiz 2
Date: 30 March 2023

Full Marks = 30

Answer the questions in the spaces provided on the question sheets.

Roll Number: ______________________________

Name: ______________________________

1. (1 mark) Which of the following statements is INCORRECT?

   A. Recurrent neural networks can handle a sequence of arbitrary length, while feedforward neural networks can not.
   B. Training recurrent neural networks is hard because of vanishing and exploding gradient problems.
   C. Gradient clipping is an effective way of solving the vanishing gradient problem.
   D. Gated recurrent units (GRUs) have fewer parameters than LSTMs.

   1. ____C____

2. (1 mark) Dense word vectors learned through word2vec, or GloVe, has many advantages over sparse one-hot word vectors. Which of the following is NOT an advantage dense vectors have over sparse vectors?

   A. Models using dense word vectors generalize better to unseen words than those using sparse vectors.
   B. Dense word vectors encode similarity between words, while sparse vectors do not.
   C. Dense word vectors are easier to include as features in machine learning systems than sparse vectors.

   2. ____A____

3. (1 mark) A popular model used for sentiment classification is an LSTM model. It takes word vectors at each time step and uses the last hidden state vector to predict the sentiment label (y). Suppose we use a simple "bag-of-vectors" model for sentiment classification: we used the average of all the word vectors in a sentence to predict the sentiment label. Name one benefit of the LSTM model over the bag-of-vectors model.

   Solution: The LSTM model is able to integrate information from word ordering, e.g. this was not an amazing fantastic movie" while the bag-of- vectors model can not.

4. (1 mark) What is the primary use of the CLS token in BERT for sentence classification tasks?

A. To provide a classification decision
B. To identify the start of a sentence
C. To mask words for pertaining
D. To separate sentences in a pair

4. A

5. (1 mark) Which of the following tasks is an example of a tagging task that BERT can be used for?

A. Sentiment analysis
B. Name-entity recognition
C. Textual entailment
D. Document classification

5. B

6. (1 mark) What is one limitation of using pretrained encoders like BERT for sequence generation tasks?

A. They don't perform well on NLP tasks
B. They don't naturally lead to autoregressive generation methods
C. They are computationally expensive
D. They require manual feature engineering

6. B

7. (1 mark) What is a key improvement made by RoBERTa over the original BERT model?

A. Adding more layers to the Transformer architecture
B. Training BERT for a longer duration and removing the next sentence prediction
C. Using a different pretraining task
D. Focusing on span-based question-answering tasks only

7. B

8. (1 mark) What is the main difference between full finetuning and lightweight finetuning?

A. Full finetuning adapts all parameters, while lightweight finetuning trains only a few existing or new parameters
B. Full finetuning uses smaller batch sizes, while lightweight finetuning uses larger batch sizes
C. Full finetuning is used for pretrained encoders, while lightweight finetuning is used for pretrained decoders
D. Full finetuning focuses on improving representations, while lightweight finetuning focuses on text generation

8. A

9. (1 mark) What is the main goal of parameter-efficient finetuning techniques?

A. To improve the quality of text generation
B. To adapt pretrained models in a constrained way, leading to less overfitting or more efficient finetuning and inference
C. To increase the size of the model for better performance
D. To introduce new pretraining tasks for better representations

9. B

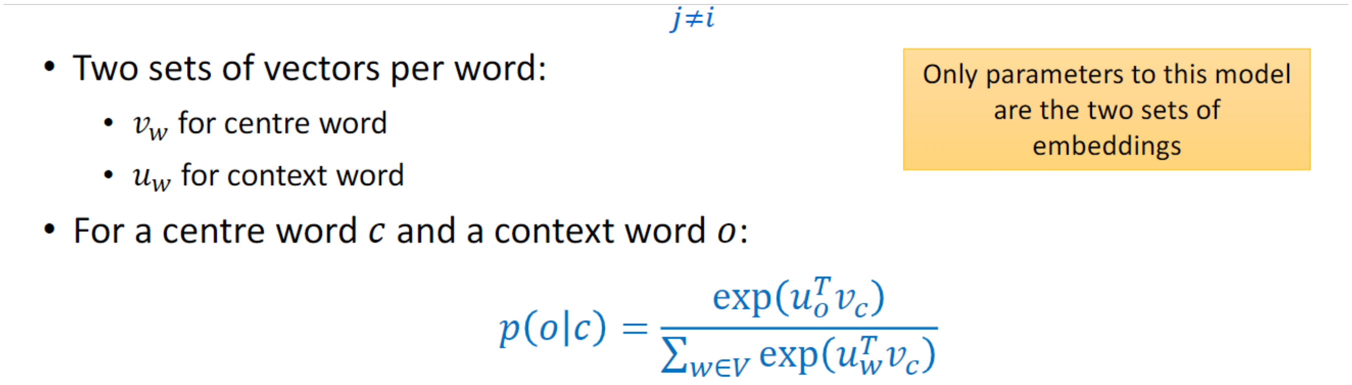
10. (1 mark) What is the main advantage of one-directional forward transformer models like GPT over bidirectional transformer models like BERT?

A. Better at generating text
B. Better representations
C. Faster training
D. More versatile for different tasks

10. A

11. (1 mark) In Word2Vec architecture define the probability for the outside (context) word o given the center word c i.e $P(o|c)$. Describe the variable names you used.

Solution



$j \neq i$

- Two sets of vectors per word:
  - v_w for centre word
  - u_w for context word
- For a centre word c and a context word o:

$$p(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$$

Only parameters to this model are the two sets of embeddings

Figure 1: Solution for Question 11

12. (1 mark) In Word2Vec given the T context windows, define the objective function $J(\theta)$ used to train the network. Describe the variable names you used.

Solution

13. (2 marks) Suppose you are training a Word2Vec network with Skip-Gram architecture. The vocabulary V contains 5 words and the hidden layer dimension is 2. After the last step of training, you have the following word vectors:

$$L_\theta = \frac{1}{T} \times \sum_{i=1}^{T} \sum_{\substack{j\in\{i-m \ldots i+m\} \\ j\neq i}} \log\left(p_\theta(w_j|w_i)\right)$$

Figure 2: Solution for question 12

$$\begin{aligned} v_1 &= [0.3, 0.4]^T & u_1 &= [0.3, 0.6]^T \\ v_2 &= [0.5, 0.8]^T & u_2 &= [0.6, 0.7]^T \\ v_3 &= [0.3, 0.7]^T & u_3 &= [0.3, 0.7]^T \\ v_4 &= [0.7, 0.8]^T & u_4 &= [1.0, 2.0]^T \\ v_5 &= [0.3, 5]^T & u_5 &= [5.0, 0.3]^T \end{aligned}$$

Consider the input, hidden and output vectors to be column vectors. Write down the values of the matrix W_1 (2 $\times$5) and W_2 (5$\times$2).

Solution:

$$W_1 = [v_1, v_2, v_3, v_4, v_5] = \begin{bmatrix} 0.3 & 0.5 & 0.3 & 0.7 & 0.3 \\ 0.4 & 0.8 & 0.7 & 0.8 & 5 \end{bmatrix}$$

$$W_2 = [u_1, u_2, u_3, u_4, u_5]^T = \begin{bmatrix} 0.3 & 0.6 \\ 0.6 & 0.7 \\ 0.3 & 0.7 \\ 1.0 & 2.0 \\ 5.0 & 3.0 \end{bmatrix}$$

14. (2 marks) Suppose you have the word embeddings for the following words:

India: $[1, 2, 3]$ Delhi: $[7, 6, 3]$ and Great Britain: $[4, 2, 1]$

Which of the following word embeddings is most likely for the word – London?

A. $[1, 4, 2]$ B. $[4, 2, 6]$ C. $[7, 5, 3]$ D. $[3, 4, 6]$

14. ____________

Solution:

$$\begin{aligned}
w_{India} - w_{Delhi} &\approx w_{GreatBritain} - w_{London} \\
w_{London} &\approx w_{GreatBritain} - w_{India} + w_{Delhi} \\
&= [4,2,1] - [1,2,3] + [7,6,3] \\
&= [10,6,1]
\end{aligned}$$

Find cosine similarity of $[10, 6, 1]$ with all options and select the option with highest value.

$$\begin{aligned}
[1,4,2]^T \times [10,6,1] &= 0.67 \\
[4,2,6]^T \times [10,6,1] &= 0.66 \\
[7,5,3]^T \times [10,6,1] &= 0.96 \\
[3,4,6]^T \times [10,6,1] &= 0.67
\end{aligned}$$

Answer: Option c

15. (3 marks) Suppose you are training a Skip-Gram Word2Vec network. Your vocabulary contains the words {brown, dog, fox, jumps, lazy, over, quick, the} indexed in lexicographic order. You are given the example sentence – **the quick brown fox jumps over the lazy dog**.

Consider that you have the following word vectors (in order listed in Vocabulary):

$$[1,-2], [-4,3], [1,0], [6,-4], [4,-5], [2,-1], [3,-4], [4,-6]$$

Consider that both word vectors for a word are the same.

The network is given the example: the quick brown fox jumps over the lazy dog. Considering "fox" to be your center word and a context window of 2 words, the loss for the word "jumps" is of the form $-\log x$. What is the value of x ? Round your answer to 2 decimal places.

Solution:

Input One hot vector $(i_c) = [0, 0, 1, 0, 0, 0, 0, 0]^T$

Center word vector (v_c)

$$= W_1 \times i_c$$

$$= \begin{bmatrix} 1 & -4 & 1 & 6 & 4 & 2 & 3 & 4 \\ -2 & 3 & 0 & -4 & -5 & -1 & -4 & -6 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Model output vector (o)

$$= W_2 \times v_c$$

$$= \begin{bmatrix} 1 & -2 \\ -4 & 3 \\ 1 & 0 \\ 6 & -4 \\ 4 & -5 \\ 2 & -1 \\ 3 & -4 \\ 4 & -6 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ -4 \\ 1 \\ 6 \\ 4 \\ 2 \\ 3 \\ 4 \end{bmatrix}$$

Softmax Output vector (q) =

$$\begin{bmatrix} 0.0049826 \\ 3.35725 * 10^{-5} \\ 0.0049826 \\ 0.739484 \\ 0.100078 \\ 0.0135441 \\ 0.0368167 \\ 0.100078 \end{bmatrix}$$

Loss = Cross Entropy Loss $([0,0,0,1,0,0,0,0], q) = -\log 0.739$.

Therefore $x = 0.74$ (rounded)

16. (1 mark) Which is not a form of attention used in the classical transformer architecture?

A. Self attention
B. Cross attention
C. Masked attention
D. Bidirectional attention

16. D

17. (1 mark) Which of the following is not a problem of RNN that Transformer solves?

A. Difficulty in capturing long range dependencies
B. Problem in processing the sequences parallelly
C. Problem of Vanishing gradients
D. Difficulty in learning hierarchical structures

17. D

18. (2 marks) What is the time complexity of self-attention in Transformers in big-O notation, given a sequence length of N and a dimensionality of d?

18. (O(n$^2 d$)

19. (2 marks) What is the purpose of the attention mechanism in a transformer?

A. To calculate the importance of each input token
B. To randomly select input tokens for encoding
C. To add noise to the input tokens
D. To remove redundant input tokens

19. ___A___

20. (2 marks) Let $e_{t,l}$=$k_t.q_l$ denote the attention score for encoder step t to decoder step l where k_t is key vector and q_l is query vector. Why is h_t (hidden state of encoder step t) with highest $e_{t,l}$ not chosen for decoder step l and instead we go for a softmax approach? Give answer in 1 line.

Solution: Because argmax is not differentiable

21. (3 marks) Consider a Transformer model with an input sequence length of 3, a key vector dimension of 3, and a query vector dimension of 3. If the dot-product attention mechanism is used, and the query vectors at each time step are $[1, 0, 0]$, $[0, 1, 0]$, and $[0, 0, 1]$, and the key vectors are $[2, 3, 4]$, $[5, 6, 7]$, and $[8, 9, 10]$, respectively, what is the value of the attention weight $\alpha_{1,3}$ from the first time step to the third time step (i.e. query from 1st time step to all 3 time steps is considered)? Round off your answer to 2 decimal places.

Answer: 0.87

Rough Work