CS 162 Natural Language Processing
Sample Midterm Exam

February 12 2024 at 8am

Instructions

1. The exam is 1 hour 30 minutes long and contains 7 pages, including this one. We will begin promptly at 8:00 am and end at 9:30 am. You may not begin until instructed to do so and you must stop once instructed to do so.

2. You may use as reference one double-sided paper with notes. All of the following are considered academically dishonest and are not allowed: communicating with fellow students during an exam, copying or attempting to copy material from another student's exam; allowing another student to copy from an exam or assignment; possession or use of unauthorized notes (i.e. any other than the double-sided paper you bring in), graphic calculator, or other materials during exams, unauthorized removal of exam materials, and submission of work altered after grading, including but not limited to changing answers after an exam has been returned or submitting another's exam as your own to gain credit. Violation of this policy is subject to immediate confiscation of the exam and other penalties as outlined by the University Student Conduct Code.

3. All numeric answers may be expressed with sums, fractions, exponents, and decimal values.

4. Please follow the instructions closely.

5. If you have a question raise your hand and an instructor will come to you. Speak quietly. If necessary, clarifications will be written on the board.

6. If you finish the exam early raise your hand and an instructor will take it from you.

Name: _____

Signature: _____

UID: _____

Question 1: Multiple Choice (3 points each for total 24)

For each question circle all answers that apply. One or more than one may apply in each case. You will be given partial credit if some of your answers are correct.

1a If $P(A) = \frac{1}{2}, P(B|A) = \frac{1}{4}$, and $P(A|B) = \frac{3}{10}$, what is $P(B) =$?

a. $\frac{1}{8}$

b. 1

c. $\frac{3}{20}$

d. $\frac{5}{12}$

1b Consider GloVe embeddings with 300 dimensions trained on some large-scale language corpora for the following word vectors: v_{France}, v_{Paris}, $v_{Germany}$, and v_{Berlin}; which of following answers is closest to $v_{Germany}$

a. $v_{France} + v_{Berlin} - v_{Paris}$

b. $v_{France} + v_{Paris} - v_{Berlin}$

c. $v_{France} - v_{Berlin} - v_{Paris}$ *(Please note there was a typo in this option in the sample midterm)*

d. $v_{France} + v_{Paris} + v_{Berlin}$

1c Which of the following statement(s) about n-gram language models are True?

a. The model will contain more syntactic information if the n is larger
b. We could use perplexity to conduct the intrinsic evaluation for language models
c. The model will need to save more parameters if n is larger
d. Language model's perplexity is the higher the better

1d Which level of language cares about the literal meaning of a sentence?

a. Pragmatics
b. Morphology
c. Syntax
d. Semantics

1e Which of the following statement(s) are True for Word2Vec?

a. Word2Vec model utilizes co-occurrences of words as a training objective.
b. The continuous bag of words (CBOW) module is implemented with a Recurrent Neural Network (RNN) module.
c. Word2Vec model can be utilized to compute the cosine similarity of two sentences and/or documents.
d. Word2Vec model contains an input embedding matrix and an output embedding matrix as its parameters.

1f The word "engine" is a _____ of "submarine"

a. Synonym
b. Meronym
c. Holonym
d. Homonym

Question 2: Short-Answer Questions (12 points)

Please show your work as partial credit can be given.

2a The structural difference between LSTMs and RNNs is that LSTM has three gates to control the memory in the cells. (3 pts)

2b If we have $p(x) = 1/2, p(y) = 1/2, p(x, y) = 1/4$, then $PMI(x, y) = 0$. (3 points)

2c The cosine similarity between $u = [20, 1]$ and $v = [1, 10]$ is 30. (3 pts)

2d Assume you use a bi-gram to model English sentences, then for sentence $P(I, like, a, cup, of, tea, .)$, the probability of like and cup is independent or dependent?

independent

Question 3: Transformers (10 points)

3a Fill in the blanks. (5 points)

a. A Transformer encoding layer consists of a self-attention block and a feed forward block
b. The single-head self-attention mechanism obtains a weighted sum of all the value vectors.
c. What used to compute a compatibility score of a target element with all elements in the same input sequence - what the element is "offering" when computed against the target one? key
d. What is the matrix which we aim to assign a weight to each of its element and compute the final output? value
e. What is used to compute a compatibility score of a target element with all elements in the same input sequence - what we are focusing on in the input sequence, i.e., the target? query

3b In the Transformer model, what is one advantage of multi-head attention over single-head attention? Write your answers in one sentence. (3 pts)

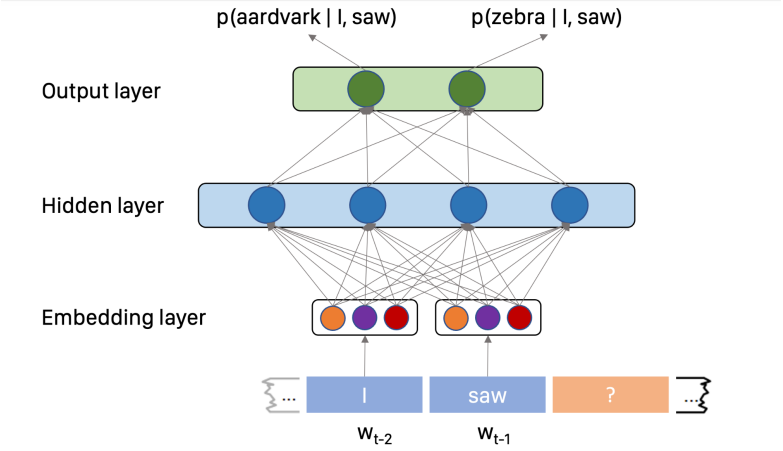
MHA helps encode different aspects of the query token with respect to all other tokens, like different heads can represent syntax semantics, etc. Whereas SHA, will represent only one aspect of query token with respect to all other tokens.

3c For an eight-head transformer, suppose the input token length is 15 and the total hidden dimension size is 512, what is the size of Q/K/V? (2 pts)

15×64 as total dimension is 512 and there are 8 heads. As hidden dimension per head will be the same, we have dimension for one head 64. Dimension of $Q/K/V$ = *input token length* $\times$ *hidden dimension per head*

Question 4: Forward and Backward Pass in Neural Language Model (12 points)

Assume we are training a feedforward neural language model. The model structure is shown as the figure below. The model takes in two words ($w_{t\text{-}2}$ and $w_{t\text{-}1}$) and output probability for the next word $P\ (w_t \mid w_{t\text{-}2}\ , w_{t\text{-}1})$. The embedding layer provide word vectors $e_{t\text{-}2}$ and $e_{t\text{-}1}$ for the



two words $w_{t\text{-}2}$ and $w_{t\text{-}1}$, these two vectors are concatenated together to produce e as the input to the hidden layer. We multiply e by a weight matrix W and add bias b and pass through the ReLU activation function to get h. Then we multiply h by another weight matrix U. After the softmax, each node in the output layer estimates the final probability distribution over words. Let's assume that word vector dimension for each word is 3, there are 4 hidden units in the hidden layer, and the vocabulary size (all possible unique words that can be the next word) is 2.

$$
\begin{aligned}
e &= [e_{t\text{-}2}; e_{t\text{-}1}] \\
h &= \sigma(We + b) \\
z &= Uh \\
\hat{y} &= \text{softmax}(z)
\end{aligned}
$$

4a Write down sizes of the parameters or intermediate variables (in the form such as 1000 × 10): (4 points)

a. Size of e : $A \times 1$ A = 6

b. Size of W : $A \times B$ A = 4, B = 6

c. Size of b : $A \times B$ A = 4, B = 1

d. Size of U : $A \times B$ A = 2, B = 4

e. Size of $\hat{y}$: $A \times 1$ A = 2

4b At a certain iteration, h is $[0.2, 0.3, 0.9, 0.4]$ and U is $[[0.5, 0.6, 0.1, 0.2]$, $[0.3, 2.4, 0.1, \text{-}0.2]]$. What is the value of z (Please show the calculation process and the final value of z)?(2 points)

The first element is $z_0 = 0.20.5 + 0.3 \times 0.6 + 0.9 \times 0.1 + 0.4 \times 0.2 = 0.45$ The secondelement is $z_1 = 0.2 \times 0.3 + 0.3 \times 2.4 + 0.9 \times 0.1 + 0.4 \times (0.2) = 0.79$ $z = [0.45, 0.79]$ (if we flatten it to 1-D vector) or $[[0.45], [0.79]]$ (if we do not flatten it to 1- D vector)

4c Assume the vocabulary is indexed as $[\text{"}aardvark\text{"}, \text{"}zebra\text{"}]$. According to the calculated result of part 4b, which word should be the predicted next word?(2 points)

zebra, as $z[1]$ has higher value than $z[0]$, indicating the predicted probability will be higher.

4d Assume we are performing backward propagation at a certain iteration, and we've computed that gradient of flattened (converted it to 1D vector) z (a.k.a L) is $[3, 1]$. h is $[0.2, 0.3, 0.9, 0.4]$ and current value of U is $[[0.5, 0.6, 0.1, 0.2]$, $[0.3, 2.4, 0.1, 0.2]]$. What is the gradient of U (a.k.a $\frac{\partial L}{\partial U}$?). (4 points)

Note: You are not required to show the calculation process, you can just give the gradient matrix.

$$\frac{\partial L}{\partial U} = \frac{\partial L}{\partial z}\frac{\partial z}{\partial U} = \frac{\partial L}{\partial z}h = \begin{bmatrix} -0.6 & -0.9 & -2.7 & -1.2 \\ -0.2 & -0.3 & -0.9 & -0.4 \end{bmatrix}$$