B.Sc. Engg. SWE 8th Semester 14 May 2024 30

ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)

ORGANISATION OF ISLAMIC COOPERATION (OIC)

Department of Computer Science and Engineering (CSE)

SEMESTER FINAL EXAMINATION
DURATION: 3 HOURS

SUMMER SEMESTER, 2022-2023
FULL MARKS: 150

SWE 4841: Natural Language Processing

Programmable calculators are not allowed. Do not write anything on the question paper.

Answer **all 6 (six)** questions. Figures in the right margin indicate full marks of questions with corresponding COs and POs in parentheses.

1. a) Answer the following questions based on Figure 1: $5 + 5 + 8$ (CO3) (PO1)

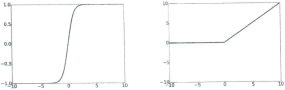

Figure 1: Activation functions for Question 1.a

i. What limitations of the sigmoid activation function do the above activation functions solve?

ii. What is the vanishing gradient problem and which activation function partly solves this problem?

iii. Why is it necessary to use these non-linear activation functions for each layer in a neural network? Explain mathematically.

b) Differentiate between multinomial logistic regression and a neural network. Comment on the utility of hidden layers of a neural network in cases of text inputs. 7 (CO3) (PO1)

2. A simple feedforward neural network is depicted in Figure 2. Answer the following questions based on this figure:

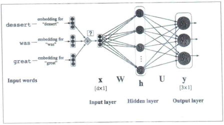

Figure 2: A simple feedforward network for Question 2

a) Which downstream task of NLP is presented in the figure? 3 (CO1) (PO1)

b) What operation is being performed on the question marked component (?) in Figure 2? Mention a few methods to perform this operation. 6 (CO2) (PO1)

c) Suppose the hidden layer has the dimension of $d_h \times 1$. Determine the dimension of the weight matrices mentioned in the figure. 6 (CO3) (PO1)

d) Modify Figure 2 to illustrate how you would train a neural language model using self-supervision. 10 (CO3) (PO1)

3. a) Discuss the differences between the NLP tasks listed below in terms of dataset annotation, training strategy, etc. 4 + 4 (CO1) (PO1)

   i. Sequence Labeling and Sequence Classification

   ii. Language Modeling and Causal LM Generation

   b) Explain why stacking recurrent networks generally outperforms single-layer networks. Similarly, how have bidirectional RNNs proven to be quite effective at sequence classification? 4 + 5 (CO3) (PO1)

   c) Figure 3 is a diagram of an LSTM unit that learns to forget textual information that is no longer needed and to remember information required for future decisions. 8 (CO3) (PO1)

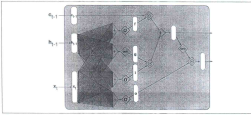

Figure 3: A single LSTM unit for Question 3.c

Describe the gates in the above diagram that allow the network to forget and remember distant information. Provide mathematical deductions for the responsible gates.

4. Four architectures for NLP tasks are depicted in Figure 4. Answer the following questions based on this diagram: 5 × 5 (CO3) (PO1)

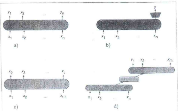

Figure 4: Four architectures for NLP tasks for Question 4

a) Identify the NLP tasks and architectures depicted in Figure 4.

b) How is the loss calculated in each of these NLP tasks?

c) In what NLP downstream tasks is architecture d) used? Justify the need for this architecture in those tasks.

d) How would you turn a language model with autoregressive generation into architecture d) for machine translation tasks?

e) What limitation of architecture d) can be solved with attention mechanism?

5. a) Explain how the notion of distributional hypothesis is utilized to build powerful language models. 5 (CO1) (PO1)

b) Differentiate between the following concepts 5 + 5 (CO3) (PO3)

   i. attention and self-attention mechanism

   ii. pre-norm and post-norm transformer

c) Label the components of the incomplete transformer block provided in Figure 5. Draw the incomplete parts of the block as well. 10 (CO3) (PO3)

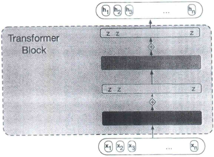

Figure 5: An incomplete transformer block for Question 5.c

6. a) Consider the following sentence that passes through a single causal self-attention unit: 20 (CO3) (PO1)

Sally sells seashells by the seashore

The query, key, and values for each of the tokens are depicted in Table 1 :

Table 1: Query, Key, and Value Vectors for question Question 6.a

Token	Query	Key	Value
Sally	[0.72, 0.45, 0.31, 0.81]	[0.13, 0.65, 0.29, 0.97]	[0.55, 0.24, 0.83, 0.06]
sells	[0.63, 0.29, 0.91, 0.44]	[0.52, 0.07, 0.72, 0.31]	[0.89, 0.46, 0.71, 0.15]
seashells	[0.19, 0.98, 0.57, 0.34]	[0.36, 0.81, 0.94, 0.02]	[0.74, 0.23, 0.09, 0.61]
by	[0.65, 0.52, 0.83, 0.18]	[0.79, 0.03, 0.49, 0.68]	[0.21, 0.97, 0.33, 0.12]
the	[0.48, 0.91, 0.11, 0.72]	[0.28, 0.39, 0.55, 0.66]	[0.93, 0.14, 0.27, 0.75]
seashore	[0.81, 0.36, 0.77, 0.09]	[0.42, 0.65, 0.79, 0.54]	[0.68, 0.59, 0.03, 0.83]

Calculate the output of the self-attention unit for the token *seashells*.

b) What is hallucination in language models? What strategies can be employed to address this issue? 5 (CO1) (PO1)