MIDDLE EAST TECHNICAL UNIVERSITY

SEMESTER I EXAMINATION 2024-2025

CENG 403 – Deep Learning - Self-Attention and Transformers

January 2025 TIME ALLOWED: 3 HOURS

INSTRUCTIONS TO CANDIDATES

1. This examination paper contains **FIVE (5)** questions and comprises **SIX (6)** printed pages.
2. Answer all questions. The marks for each question are indicated at the beginning of each question.
3. Answer each question beginning on a **FRESH** page of the answer book.
4. This **IS NOT an OPEN BOOK** exam.
5. Calculators are allowed for numerical computations.
6. Show all mathematical derivations and computational steps clearly.
7. For matrix operations, clearly indicate dimensions and show intermediate steps.
8. Draw clear diagrams where requested and label all components.

Question 1. Text Encoding Methods and Challenges (20 marks)
Deep learning models can process text using different encoding approaches, each with distinct advantages and limitations.

(a) Compare and contrast three text encoding methods: character-level encoding, word-level encoding (word embeddings), and subword-level encoding (Byte Pair Encoding). For each method, discuss: (12 marks)

- How new/unseen words are handled
- Vocabulary size considerations
- Suitability for morphologically rich languages (e.g., Turkish)

(b) Explain why Byte Pair Encoding might be particularly advantageous for machine translation between languages that share some common subword structures. Provide a concrete example. (8 marks)

Question 2. RNN-based Sequence-to-Sequence Models (25 marks) Consider the encoder-decoder architecture used for neural machine translation in 2014, which achieved state-of-the-art results but had significant limitations.

(a) Draw a detailed diagram of the encoder-decoder architecture for machine translation. Label the encoder RNN, decoder RNN, hidden states, and show how information flows from input sequence to output sequence. (8 marks)

(b) Explain the long-term dependency problem in this architecture. Why does this problem become severe when translating long sentences (e.g., 100+ words)? Discuss both forward pass and backward pass challenges. (10 marks)

(c) Calculate the effective depth of the unfolded network when translating a 50-word source sentence to a 50-word target sentence. Explain why this creates optimization difficulties. (7 marks)

Question 3. Attention Mechanism and Bahdanau Attention (30 marks)
The attention mechanism was introduced to address the limitations of basic encoder-decoder models.

(a) Explain the core intuition behind attention mechanism. How does it solve the long-term dependency problem in encoder-decoder models? (8 marks)

(b) Given decoder hidden state h_i^{dec} and encoder hidden states $h_1^{enc}, h_2^{enc}, \ldots, h_T^{enc}$, write the mathematical formulation for Bahdanau attention mechanism. Include: (12 marks)

- Alignment score calculation using a 2-layer MLP
- Attention weight normalization
- Context vector computation

(c) Compare Bahdanau attention with three alternative similarity measures: cosine similarity, scaled dot product, and simple dot product. Write the mathematical expressions and discuss the computational complexity of each. (10 marks)

Question 4. Self-Attention Mechanism and Scaled Dot-Product Attention (35 marks)

Self-attention allows parallel processing of sequence elements and forms the foundation of transformer architectures.

(a) Starting from the basic self-attention concept, derive the complete mathematical formulation for scaled dot-product attention. Given input embeddings $E_0, E_1, \ldots, E_{T-1}$, show how to compute: (15 marks)

- Query, Key, and Value matrices using learnable parameters W_Q, W_K, W_V
- Attention scores and their normalization
- Final output computation

Include the scaling factor and explain its necessity.

(b) Consider two word embeddings $x_1 = [1, 2, 0]$ and $x_2 = [0, 1, 2]$ with dimension $d = 3$. Given: (12 marks)

$$W_Q = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}, \quad W_K = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \quad W_V = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

Calculate the updated representation for x_1 using scaled dot-product attention. Show all intermediate steps including query/key/value computation, attention scores, and final weighted combination.

(c) Explain why self-attention is position-invariant and why this is both an advantage and a limitation. How does this property affect the computational complexity compared to RNNs? (8 marks)

Question 5. Transformer Architecture and Multi-Head Attention (40 marks)
The transformer architecture revolutionized sequence modeling by replacing recurrence with attention mechanisms.

(a) Draw a complete transformer encoder block. Include and label: multi-head self-attention, skip connections, layer normalization, feed-forward network, and show the flow of information. Explain the purpose of each component. (12 marks)

(b) Explain multi-head attention mechanism. Why do we use multiple attention heads instead of a single large attention mechanism? Include the mathematical formulation showing how heads are computed in parallel and combined. (10 marks)

(c) Design and explain a positional encoding scheme. Compare learnable positional embeddings vs. trigonometric positional encoding. What are the trade-offs, especially regarding generalization to longer sequences than seen during training? (10 marks)

(d) A transformer decoder differs from the encoder in two key ways. Explain: (8 marks)

- Masked self-attention and why it's necessary
- Cross-attention mechanism and how it connects to the encoder

END OF PAPER