MIDDLE EAST TECHNICAL UNIVERSITY

SEMESTER I EXAMINATION 2024-2025

CENG 403 – Deep Learning - RNNs, Language Modeling & Word Embeddings

January 2025 TIME ALLOWED: 3 HOURS

INSTRUCTIONS TO CANDIDATES

1. This examination paper contains **SIX (6)** questions and comprises **EIGHT (8)** 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. Show clear reasoning for your answers, especially intuitive explanations.
6. For algorithms, provide step-by-step explanations as taught in lectures.
7. Draw diagrams where requested and explain information flow clearly.
8. Connect concepts to modern applications (LLMs, etc.) where relevant.

Question 1. RNN Backpropagation and Weight Sharing (25 marks)

Based on the professor's explanation of how RNNs work as "feedforward networks with weight sharing across time."

(a) The professor emphasized that "we need to be careful about weight sharing when calculating gradients." Explain why we must sum gradients from all time steps for a single weight parameter in an RNN. Use a concrete example with 3 time steps. (8 marks)

(b) During backpropagation through time, the professor stressed that "we need to start from the end because earlier copies contribute to all following time steps." Draw and explain the gradient flow through an unfolded RNN showing why this order is essential. (10 marks)

Draw unfolded RNN with gradient flow direction

Show why we must start from the end

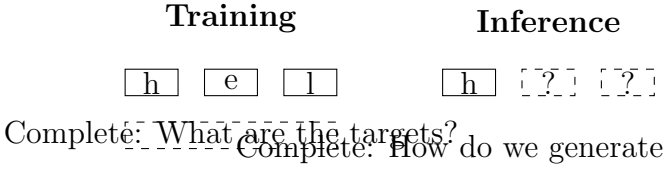
(c) The professor mentioned that RNNs suffer from "exploding gradient problem if weight norms are large, vanishing gradient problem if norms

are small." Explain the mathematical reasoning behind both problems and why LSTM addresses these issues. (7 marks)

Question 2. Autoregressive Language Modeling (20 marks)
The professor stated: "This is how large language models are trained as well - just to predict the next character."

(a) Define autoregressive modeling as explained by the professor. Write the mathematical formulation for modeling $P(x_t|x_{t-1}, x_{t-2}, \ldots, x_1)$ and explain why this is considered "self-supervised learning." (8 marks)

(b) The professor explained the difference between training and inference in language models. Complete the diagram below showing both processes for the sequence "hello": (12 marks)



Question 3. Character-Level Implementation Details (22 marks)
Based on the professor's detailed walkthrough of the "hello" example with 4-character vocabulary.

(a) The professor showed how to use one-hot encodings for characters h, e, l, o. Given the string "hello", create the complete training data with inputs and targets, including start and end tokens as explained in class. (8 marks)

(b) Follow the professor's example: given input character 'h' with one-hot encoding [1,0,0,0], weight matrix W_1 (3×4), and hyperbolic tangent activation, show the complete forward pass to predict the next character. Explain each step as the professor did. (10 marks)

(c) The professor emphasized that during inference "we use the predicted value as input for the next time step, even if it's incorrect." Explain why this creates a potential problem and how it relates to error propagation in sequence generation. (4 marks)

Question 4. Beam Search Algorithm (25 marks)

The professor explained beam search as an alternative to "greedy approach where we just take the character with highest probability."

(a) Implement the beam search algorithm as taught by the professor. Given the probability distributions below for 3 time steps with vocabulary [A, B, C], show the complete beam search process with beam size k=2: (15 marks)

Time Step	P(A)	P(B)	P(C)
t=1	0.6	0.3	0.1
t=2 (after A)	0.2	0.5	0.3
t=2 (after B)	0.4	0.1	0.5
t=3 (after AA)	0.1	0.2	0.7
t=3 (after AB)	0.3	0.3	0.4

Show the tree expansion and final top-2 sequences with their combined scores.

(b) The professor mentioned that "beam search during training is very expensive because you need to unfold this tree." Explain why beam search is typically used only during inference and what computational challenges it would create during training. (10 marks)

Question 5. Word-Level Challenges and Embeddings (28 marks)
The professor explained the transition from character-level to word-level modeling and the need for word embeddings.

(a) The professor stated that English has "170,000 different words" making one-hot encoding impractical. Calculate the number of parameters needed for: (8 marks)

- Input layer: 170,000-dimensional one-hot to 512-dimensional hidden layer
- Compare this with character-level (30 characters to 512 dimensions)
- Explain why this creates a "huge" parameter problem as the professor mentioned

(b) The professor emphasized that "in one-hot representation every word is equally distant to each other." Explain why this is problematic for semantic understanding and how word embeddings solve this issue. Use the examples the professor gave: "running and jogging should be close vs. running and swimming." (10 marks)

(c) Design the CBOW (Continuous Bag of Words) architecture as explained by the professor. For the sentence "I like running every day" with target word "running", show: (10 marks)

- Context words and their representation
- How the weight matrix W1 stores embeddings

- Why "the network is forced to learn word vectors that capture relevance"

Question 6. Deep Understanding and Modern Connections (30 marks)
Based on the professor's connections between classical methods and modern applications.

(a) The professor made a key connection: "The prompts that you provide to LLMs are actually the starting sequences that you provide to an RNN-like architecture." Explain this connection and how modern LLMs relate to the autoregressive character prediction discussed in class. (10 marks)

(b) Analyze the professor's examples of character-level RNN outputs (Shakespeare, Wikipedia, LaTeX). Explain what these results demonstrate about: (12 marks)

- Learning syntax without explicit supervision
- Generalization vs. memorization (the professor's "overfitting" discussion)
- Why the professor called these results "striking" for such simple models

(c) The professor showed word embedding arithmetic: "Paris - France + Italy = Rome." Explain: (8 marks)

- Why this works mathematically in embedding space

- How the CBOW training creates these relationships
- Give two more examples of word arithmetic that should work

END OF PAPER