MIDDLE EAST TECHNICAL UNIVERSITY

SEMESTER I EXAMINATION 2024-2025

CENG 403 – Deep Learning - RNNs, LSTM & Language Models (University Sources)

January 2025 TIME ALLOWED: 3 HOURS

INSTRUCTIONS TO CANDIDATES

1. This examination paper contains **SEVEN (7)** questions and comprises **TEN (10)** 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 all mathematical derivations clearly with proper notation.
6. For implementation questions, provide clear pseudocode or algorithms.
7. Explain computational complexity where requested.
8. Draw clear architectural diagrams with proper labels.

Question 1. Backpropagation Through Time (BPTT) (25 marks)
Based on Stanford/MIT/University of Toronto deep learning courses.

(a) Define Backpropagation Through Time (BPTT) and explain how it extends traditional backpropagation to sequential data. Include the mathematical formulation for unfolding an RNN across time steps. (8 marks)

(b) Consider an RNN with hidden state update equation: (12 marks)

$$h_t = \tanh(W_{hx}x_t + W_{hh}h_{t-1} + b_h)$$

$$y_t = W_{hy}h_t + b_y$$

For a sequence of length T=3, derive the gradient $\frac{\partial L}{\partial W_{hh}}$ where $L = \sum_{t=1}^{3} L_t$ is the total loss. Show the complete gradient computation including the chain rule applications.

(c) Explain truncated BPTT and why it's necessary for long sequences. Compare regular truncation vs. randomized truncation approaches, discussing computational complexity and numerical stability. (5 marks)

Question 2. LSTM Architecture and Gating Mechanisms (30 marks)
Based on university deep learning course materials and D2L.ai educational content.

(a) Draw the complete LSTM cell architecture showing all gates (input, forget, output) and state components (cell state, hidden state). Label all weight matrices and explain information flow. (10 marks)

Draw complete LSTM cell architecture

Include: gates, states, weight matrices, information flow

(b) Write the mathematical equations for all LSTM gates and state updates. Given input x_t, previous hidden state h_{t-1}, and previous cell state C_{t-1}, derive: (15 marks)

- Forget gate: $f_t = ?$
- Input gate: $i_t = ?$
- Candidate values: $\tilde{C}_t = ?$
- Cell state update: $C_t = ?$
- Output gate: $o_t = ?$

- Hidden state: h_t =?

(c) Compare LSTM vs. vanilla RNN in terms of gradient flow. Explain how the cell state pathway in LSTM addresses the vanishing gradient problem through mathematical analysis of gradient propagation. (5 marks)

Question 3. Word Embeddings and Distributional Semantics (22 marks)
Based on NLP course materials from Stanford CS224n and similar university programs.

(a) Explain the distributional hypothesis that underlies word embeddings. How does the CBOW (Continuous Bag of Words) model implement this principle? (6 marks)

(b) Design the Skip-gram model architecture for learning word embeddings. Given vocabulary size V=10,000 and embedding dimension d=300: (10 marks)

- Draw the network architecture
- Calculate the number of parameters
- Explain the softmax bottleneck and hierarchical softmax solution
- Derive the loss function using negative sampling

(c) Analyze word embedding arithmetic: "king - man + woman queen". Explain: (6 marks)

- Why this works mathematically in embedding space
- What linguistic relationships are captured

- Limitations of linear analogies in embeddings

Question 4. Sequence-to-Sequence Models and Attention (28 marks)
Based on modern NLP course materials covering encoder-decoder architectures.

(a) Design an encoder-decoder RNN architecture for machine translation from English to French. Show the complete architecture including: (12 marks)

- Encoder RNN processing input sequence
- Context vector computation
- Decoder RNN generating output sequence
- How teacher forcing works during training
- Inference procedure using greedy/beam search

(b) Implement beam search decoding with beam size k=3. Given the following probability distributions over vocabulary A, B, C, <EOS> for 2 time steps: (10 marks)

Context	P(A)	P(B)	P(C)	P(<EOS>)
Initial	0.5	0.3	0.15	0.05
After A	0.2	0.1	0.6	0.1
After B	0.4	0.2	0.3	0.1
After C	0.1	0.7	0.1	0.1

Show the complete beam search tree and final ranked sequences.

(c) Explain the attention mechanism as a solution to the bottleneck problem in sequence-to-sequence models. Derive the mathematical formulation for Bahdanau attention. (6 marks)

Question 5. Gradient Problems and Solutions (20 marks)
Based on theoretical analysis from university deep learning courses.

(a) Analyze the vanishing gradient problem in RNNs. For an RNN with hidden state transition $h_t = \tanh(Wh_{t-1} + Ux_t)$, show mathematically why gradients vanish for long sequences. (8 marks)

Include analysis of:

- Gradient computation through multiple time steps
- Effect of tanh derivative bounds
- Impact of weight matrix eigenvalues

(b) Compare three solutions to gradient problems in RNNs: (12 marks)

- Gradient clipping (explain algorithm and threshold selection)
- LSTM gating mechanisms (focus on gradient flow)
- Skip connections in deep RNNs

Provide mathematical justification for each approach.

Question 6. Language Modeling Evaluation and Perplexity (18 marks)
Based on NLP evaluation metrics taught in university courses.

(a) Define perplexity as a measure of language model quality. Given a test sequence $w_1, w_2, \ldots, w_N$, derive the relationship between perplexity and cross-entropy loss. (8 marks)

(b) A character-level RNN language model with vocabulary size 50 achieves the following results: (10 marks)

- Training perplexity: 1.8
- Validation perplexity: 2.3
- Test perplexity: 2.5

- Calculate the average bits per character for each dataset
- Analyze what these results indicate about model performance
- Compare with a baseline uniform model (calculate its perplexity)
- Suggest improvements to reduce the validation-test gap

Question 7. Modern RNN Variants and Applications (27 marks)
Based on advanced topics from recent university deep learning curricula.

(a) Compare GRU (Gated Recurrent Unit) with LSTM architecture. Draw both architectures and explain: (12 marks)

- Key architectural differences
- Parameter count comparison
- Computational efficiency analysis
- When to choose GRU vs LSTM

(b) Design a bidirectional RNN for named entity recognition. Explain: (8 marks)

- Forward and backward pass computations
- How to combine directional information
- Advantages for sequence labeling tasks
- Training considerations

(c) Analyze the computational complexity of different RNN variants: (7 marks)

Model	Time Complexity	Space Complexity	Parameters
Vanilla RNN	?	?	?
LSTM	?	?	?
GRU	?	?	?
Bidirectional LSTM	?	?	?

For sequence length T, hidden size H, and input size D.

END OF PAPER