Introduction to Large Language Models
Assignment- 5

Number of questions: 8 **Total mark: 6 X 1 + 2 X 2 = 10**

QUESTION 1: [1 mark]

Which of the following is a disadvantage of Recurrent Neural Networks (RNNs)?

a. Can only process fixed-length inputs.
b. Symmetry in how inputs are processed.
c. Difficulty accessing information from many steps back.
d. Weights are not reused across timesteps.

Correct Answer: c

Solution: Please refer to the lecture slides.

QUESTION 2: [1 mark]

Why are RNNs preferred over fixed-window neural models?
a. They have a smaller parameter size.
b. They can process sequences of arbitrary length.
c. They eliminate the need for embedding layers.
d. None of the above.

Correct Answer: b
Solution: Please refer to lecture slides.

QUESTION 3: [1 mark]

What is the primary purpose of the cell state in an LSTM?
a. Store short-term information.
b. Control the gradient flow across timesteps.
c. Store long-term information.
d. Perform the activation function.

Correct Answer: c
Solution: The cell stores long-term information in LSTM.

QUESTION 4: [1 mark]

In training an RNN, what technique is used to calculate gradients over multiple timesteps?
a. Backpropagation through Time (BPTT)

b. Stochastic Gradient Descent (SGD)
c. Dropout Regularization
d. Layer Normalization

Correct Answer: a
Solution: Please refer to lecture slides.

QUESTION 5: [2 mark]

Consider a simple RNN:

- Input vector size: 3
- Hidden state size: 4
- Output vector size: 2
- Number of timesteps: 5

How many parameters are there in total?

a. 210
b. 190
c. 90
d. 42

Correct Answer: d
Solution:
Input to hidden weights: 3×4=12
Hidden to hidden weights: 4×4=16
Hidden to output weights: 4×2=8
Bias terms: 4(hidden) + 2(output) = 6
Total: 12+16+8+6=42

QUESTION 6: [1 mark]

What is the time complexity for processing a sequence of length 'N' by an RNN, if the input embedding dimension, hidden state dimension, and output vector dimension are all 'd'?

a. $O(N)$
b. $O(N^2d)$
c. $O(Nd)$
d. $O(Nd^2)$

Correct answer: d
Solution: The time complexity of processing a sequence of length N by an RNN depends on the computational cost of updating the hidden state at each time step.

At each time step, the RNN updates its hidden state h_t using the previous hidden state h_{t-1} and the current input x_t. This update typically involves matrix multiplications:

I. Input-to-hidden transformation: $W_x * x_t$, where W_x is a d × d matrix, leading to a complexity of $O(d^2)$.
II. Hidden-to-hidden transformation: $W_h * h_{t-1}$, where W_h is also a d × d matrix, leading to a complexity of $O(d^2)$.
III. Activation function application: This is typically O(d) and negligible compared to matrix multiplications.

Since these computations occur at every time step, the total complexity for a sequence of length N is: $O(N * d^2)$

QUESTION 7: [1 mark]

Which of the following is true about Seq2Seq models?
(i) Seq2Seq models are always conditioned on the source sentence.
(ii) The encoder compresses the input sequence into a fixed-size vector representation.
(iii) Seq2Seq models cannot handle variable-length sequences.
a. (i) and (ii)
b. (ii) only
c. (iii) only
d. (i), (ii), and (iii)

Correct Answer: a
Solution: Seq2Seq models are designed to encode variable-length sequences but compress them into fixed-size vector representations.

QUESTION 8: [2 marks]

Given the following encoder and decoder hidden states, compute the attention scores. (Use dot product as the scoring function)

Encoder hidden states: h_1=[1,2], h_2=[3,4], h_3=[5,6]

Decoder hidden state: s=[0.5,1]

a. 0.00235,0.04731,0.9503
b. 0.0737,0.287,0.6393
c. 0.9503,0.0137,0.036
d. 0.6393,0.0737,0.287

Correct Answer: a
Solution:
e_1 = 1*0.5+2*1 =0.5+2 = 2.5

$e_2 = 3*0.5+4*1 = 1.5+4 = 5.5$
$e_3 = 5*0.5+6*1 = 2.5+6 = 8.5$

$\alpha 1 = e^{2.5}/(e^{2.5} + e^{5.5} + e^{8.5}) = 0.00235$

$\alpha 2 = e^{5.5}/(e^{2.5} + e^{5.5} + e^{8.5}) = 0.04731$

$\alpha 3 = e^{8.5}/(e^{2.5} + e^{5.5} + e^{8.5}) = 0.9503$
