Introduction to Large Language Models

Week-5 Assignment

Number of questions: 9 **Total mark: 8 X 1 + 1 X 2 = 10**

Question 1: [1 mark]

Which of the following best explains the vanishing gradient problem in RNNs?

a. RNNs lack memory mechanisms for long-term dependencies.
b. Gradients grow too large during backpropagation.
c. Gradients shrink exponentially over long sequences.
d. RNNs cannot process variable-length sequences.

Correct Answer: c
Solution: Please refer to slides.

Question 2: [1 mark]

In an attention mechanism, what does the softmax function ensure?

a. Normalization of decoder outputs
b. Stability of gradients during backpropagation
c. Values lie between -1 and 1
d. Attention weights sum to 1

Correct Answer: d
Solution:
The softmax is applied to attention scores to produce a probability distribution over encoder hidden states. This ensures the weights sum to 1.

Question 3: [1 mark]

Which of the following is true about the difference between a standard RNN and an LSTM?

a. LSTM does not use any non-linear activation.
b. LSTM has a gating mechanism to control information flow.
c. RNNs have fewer parameters than LSTMs because they use convolution.
d. LSTMs cannot learn long-term dependencies.

Correct Answer: b
Solution: Please refer to slides.

Question 4: [1 mark]

Which gate in an LSTM is responsible for deciding how much of the cell state to keep?

a. Forget gate
b. Input gate
c. Output gate
d. Cell candidate gate

Correct Answer: a

Solution:
The forget gate determines what fraction of the previous cell state should be retained in the current timestep.

Question 5: [1 mark]

What improvement does attention bring to the basic Seq2Seq model?

a. Reduces training time
b. Removes the need for an encoder
c. Allows access to all encoder states during decoding
d. Reduces the number of model parameters

Correct Answer: c
Solution:
Attention allows the decoder to consider all encoder hidden states dynamically.

Question 6: [1 mark]

Which of the following is a correct statement about the encoder-decoder architecture?

a. The encoder generates tokens one at a time.
b. The decoder summarizes the input sequence.
c. The decoder generates outputs based on encoder representations and its own prior outputs.
d. The encoder stores only the first token of the sequence.

Correct Answer: c
Solution:
The decoder uses both the encoder's output and its own previously generated tokens to produce the next output.

Question 7: [1 mark]

What is self-attention in Transformers used for?

a. To enable sequential computation
b. To attend to the previous layer's output
c. To relate different positions in the same sequence
d. To enforce fixed-length output

Correct Answer: c
Solution:
Self-attention allows each token to focus on all other tokens in the same sequence.

Question 8: [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 9: [2 marks]
Given the following encoder and decoder hidden states, compute the attention scores. (Use dot product as the scoring function)

Encoder hidden states: h1=[7,3], h2=[0,2], h3=[1,4]
Decoder hidden state: s=[0.2,1.5]

a. 0.42, 0.02, 0.56
b. 0.15, 0.53, 0.32
c. 0.64, 0.18, 0.18
d. 0.08, 0.91, 0.01

Correct Answer: a
Solution:
e1 = 7*0.2+3*1.5 = 5.9
e2 = 0*0.2+2*1.5 = 3
e3 = 1*0.2+4*1.5 = 6.2
$\alpha 1 = e^{5.9}/(e^{5.9} + e^{3} + e^{6.2}) = 0.42$

$\alpha 2 = e^{3}/(e^{5.9} + e^{3} + e^{6.2}) = 0.02$
$\alpha 3 = e^{6.2}/(e^{5.9} + e^{3} + e^{6.2}) = 0.56$