Introduction to Large Language Models
Assignment- 6

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

QUESTION 1: [1 mark]

What is the key advantage of multi-head attention?
a. It uses a single attention score for the entire sequence
b. It allows attending to different parts of the input sequence simultaneously
c. It eliminates the need for normalization
d. It reduces the model size

Correct Answer: b

Solution: Please refer to the lecture slides.

QUESTION 2: [1 mark]

What is the role of the residual connection in the Transformer architecture?
a. Improve gradient flow during backpropagation
b. Normalize input embeddings
c. Reduce computational complexity
d. Prevent overfitting

Correct Answer: a

Solution: Please refer to lecture slides.

QUESTION 3: [1 mark]

Which of the following elements addresses the lack of sequence information in self-attention?

a. Non-linear transformations
b. Positional encoding
c. Masked decoding
d. Residual connections

Correct Answer: b

Solution: Please refer to lecture slides.

QUESTION 4: [1 mark]

For Rotary Position Embedding (RoPE), which of the following statements are true?

a. Combines relative and absolute positional information

b. Applies a multiplicative rotation matrix to encode positions

c. Eliminates the need for positional encodings

d. All of the above

Correct Answer: a, b

Solution: Please refer to the slides.

QUESTION 5: [2 mark]
Consider a sequence of tokens of length 4: [w_1,w_2,w_3,w_4]. Using masked self-attention, compute the attention weights for token w_3, assuming the unmasked attention scores are: [5,2,1,3]

a. [0.6234, 0.023, 0.3424, 0.0112]
b. [0.2957, 0.7043, 0, 0]
c. [0.9362, 0.0466, 0.0171,0]
d. [0.5061, 0.437, 0, 0.0569]

Correct Answer: c

Solution:
Note that for masked self-attention, token w_3 cannot attend to future tokens. Set scores for future tokens w_4 to $-\infty$.

$$[5,2,1,-\infty]$$

Applying softmax, ($e^{-\infty} = 0$)

$$\left[\frac{e^5}{e^5 + e^2 + e^1}, \frac{e^2}{e^5 + e^2 + e^1}, \frac{e^1}{e^5 + e^2 + e^1}, 0\right]$$

= [0.9362, 0.0466, 0.0171,0]

QUESTION 6: [1 mark]

_____ maps the values of a feature in the range [0,1].
a. Standardization
b. Normalization
c. Transformation
d. Scaling

Correct Answer: b

Solution: Normalization works by mapping all values of a feature to be in the range [0,1] using the transformation.

QUESTION 7: [1 mark]

How does masked self-attention help in autoregressive models?

a. By attending to all tokens, including future ones.
b. By focusing only on past tokens to prevent information leakage.
c. By ignoring positional information in the sequence.
d. By disabling the attention mechanism entirely.

Correct Answer: b

Solution: Masked self-attention ensures that the model only attends to past tokens in the sequence, preventing information leakage during training.

QUESTION 8: [2 marks]
For a transformer with d_{model} = 512, calculate the positional encoding for position p=10 and dimensions 2 and 3 using the sinusoidal formula:

$$PE(p,2i) = sin\left(\frac{p}{10000^{2i/d_{model}}}\right) \quad PE(p,2i+1) = cos\left(\frac{p}{10000^{2i/d_{model}}}\right)$$

a. $sin\left(\frac{10}{10000^{1/256}}\right), cos\left(\frac{10}{10000^{1/256}}\right)$
b. $cos\left(\frac{10}{10000^{1/512}}\right), sin\left(\frac{10}{10000^{1/512}}\right)$
c. $cos\left(\frac{10}{10000^{4/512}}\right), sin\left(\frac{10}{10000^{7/256}}\right)$
d. $sin\left(\frac{10}{10000^{2/512}}\right), cos\left(\frac{10}{10000^{3/512}}\right)$

Correct Answer: a

Solution:

For dimension 2, $PE(10,2) = sin\left(\frac{10}{10000^{2/512}}\right) = sin\left(\frac{10}{10000^{1/256}}\right)$

For dimension 3, $PE(10,3) = cos\left(\frac{10}{10000^{2/512}}\right) = cos\left(\frac{10}{10000^{1/256}}\right)$
