X

Week 6 : Assignment 6

The due date for submitting this assignment has passed.

Due on 2025-03-05, 23:59 IST.

Assignment submitted on 2025-03-03, 23:20 IST

1) What is the key advantage of multi-head attention? ***1 point***

- It uses a single attention score for the entire sequence
- It allows attending to different parts of the input sequence simultaneously
- It eliminates the need for normalization
- It reduces the model size

Yes, the answer is correct.
Score: 1

Accepted Answers:
It allows attending to different parts of the input sequence simultaneously

2) What is the role of the residual connection in the Transformer architecture? ***1 point***

- Improve gradient flow during backpropagation
- Normalize input embeddings
- Reduce computational complexity
- Prevent overfitting

Yes, the answer is correct.

Score: 1
Accepted Answers:
Improve gradient flow during backpropagation

3) Which of the following elements addresses the lack of sequence information in self-attention? ***1 point***

- ○ Non-linear transformations
- ◉ Positional encoding
- ○ Masked decoding
- ○ Residual connections

Yes, the answer is correct.
Score: 1
Accepted Answers:
Positional encoding

4) For Rotary Position Embedding (RoPE), which of the following statements are true? ***1 point***

- ☐ Combines relative and absolute positional information
- ☐ Applies a multiplicative rotation matrix to encode positions
- ☐ Eliminates the need for positional encodings
- ☑ All of the above

No, the answer is incorrect.
Score: 0
Accepted Answers:
Combines relative and absolute positional information
Applies a multiplicative rotation matrix to encode positions

5) 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 w3, assuming the unmasked attention scores are: [5,2,1,3] ***2 points***

- ○ [0.6234, 0.023, 0.3424, 0.0112]
- ○ [0.2957, 0.7043, 0, 0]
- ◉ [0.9362, 0.0466, 0.0171,0]
- ○ [0.5061, 0.437, 0, 0.0569]

Yes, the answer is correct.
Score: 2
Accepted Answers:
[0.9362, 0.0466, 0.0171,0]

6) _____ maps the values of a feature in the range [0,1]. ***1 point***

- ○ Standardization
- ◉ Normalization
- ○ Transformation

Week 9 ()

Week 10 ()

Week 11 ()

Week 12 ()

Year 2025 Solutions ()

- Scaling

Yes, the answer is correct.
Score: 1

Accepted Answers:
Normalization

7) How does masked self-attention help in autoregressive models? ***1 point***

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

Yes, the answer is correct.
Score: 1

Accepted Answers:
By focusing only on past tokens to prevent information leakage.

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

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

- $sin\left(\frac{10}{10000^{1/256}}\right), cos\left(\frac{10}{10000^{1/256}}\right)$
- $cos\left(\frac{10}{10000^{1/512}}\right), sin\left(\frac{10}{10000^{1/512}}\right)$
- $cos\left(\frac{10}{10000^{4/512}}\right), sin\left(\frac{10}{10000^{7/256}}\right)$
- $sin\left(\frac{10}{10000^{2/512}}\right), cos\left(\frac{10}{10000^{3/512}}\right)$

Yes, the answer is correct.
Score: 2

Accepted Answers:

$sin\left(\frac{10}{10000^{1/256}}\right), cos\left(\frac{10}{10000^{1/256}}\right)$