Started on	Monday, 20 January 2025, 10:04 PM
State	Finished
Completed on	Monday, 20 January 2025, 10:21 PM
Time taken	16 mins 21 secs
Grade	**4.00** out of 5.00 (**80**%)

Question **1**

Correct

Mark 1.00 out of 1.00

Which of the following tasks is used for pre-training of the latest BERT class of models?

- Question answering
- Neural Machine translation
- Mask language modeling ✔
- Next sentence prediction

Your answer is correct.

MLM is the pre-training strategy used for encoder only model. First BERT model was trained on Next sentence prediction tasks as well, in addition to MLM. The pretrained models are then used for transfer learning and tuned for specific tasks.

The correct answer is:
Mask language modeling

Question **2**

Correct

Mark 1.00 out of 1.00

What distinguishes the masking strategy used in the decoder from that in the encoder within the transformer architecture?

- Masking is identical in both encoder and decoder
- In the decoders, masking ensures exclusion of attention to pad tokens
- In the decoders, masking enforces attention weights to only consider preceding input tokens ✔

In the decoders, masking enforces attention weights to only consider succeeding input tokens

Your answer is correct.

The Transformer's decoder and encoder components implement distinct masking techniques. In the decoder, the masking approach compels attention weights to focus exclusively on preceding input tokens, enabling the prediction of subsequent tokens in the sequence. This mechanism, often referred to as causal masking, ensures the decoder doesn't cheat by attending to tokens that should only come after the predicted token.

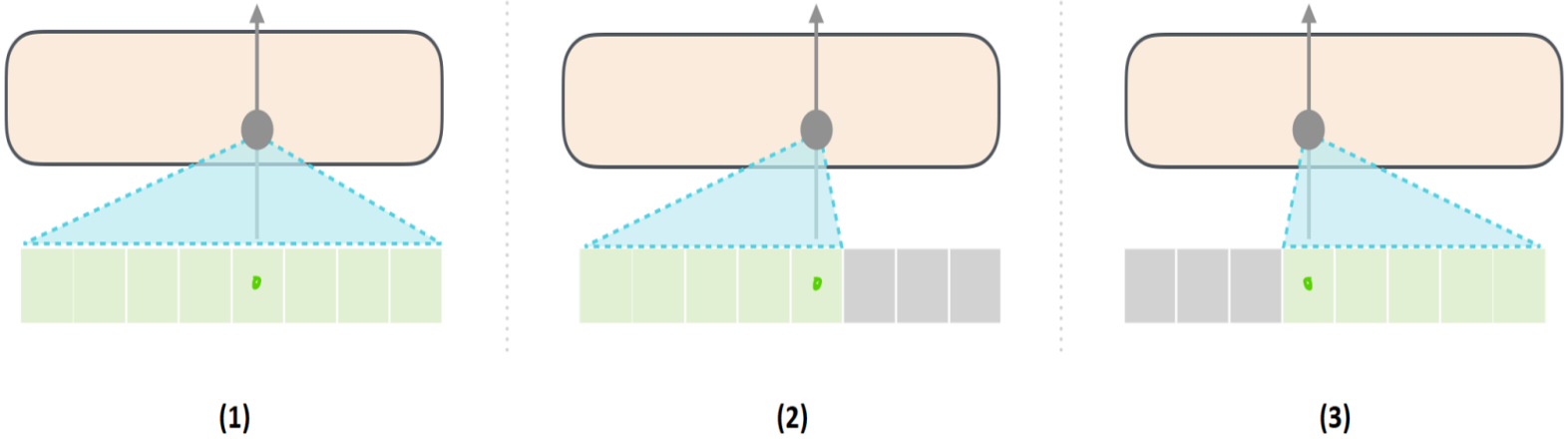
The correct answer is:
In the decoders, masking enforces attention weights to only consider preceding input tokens

Question **3**

Incorrect

Mark 0.00 out of 1.00

Select the correct option for 'Masked Self-Attention' that can be used inside Transformer Decoder architecture for performing Neural Machine Translation:



(1) (2) (3)

- Only 1
- Only 2
- Only 3 ✖
- Both 2 and 3

Your answer is incorrect.

The correct answer is:
Only 2

Question **4**

Correct

Mark 1.00 out of 1.00

How does the Transformer Decoder utilize positional encodings?

- To encode token types
- To add information about the position of tokens in the sequence ✔
- To mask future tokens
- To integrate with the attention scores

Your answer is correct.

To add information about the position of tokens in the sequence is correct because positional encodings are specifically designed to provide information about the position of tokens within the sequence. They are added to the token embeddings before they are fed into the attention layers. This addition allows the model to distinguish between tokens based on their position in the sequence, which is crucial for understanding the order and relationships between tokens.

Other options are incorrect because:

● Positional encodings do not encode token types. They are solely concerned with the positions of tokens in the sequence, not their types or categories.

not their types or categories.

● Masking future tokens is related to the masked self-attention mechanism in the Transformer Decoder, not to positional encodings.

● While positional encodings are added to token embeddings before the attention mechanism is applied, they are not directly integrated with the attention scores. Their purpose is to add positional information to the embeddings, which then become part of the input to the attention mechanism.

The correct answer is:
To add information about the position of tokens in the sequence

Question **5**

Correct

Mark 1.00 out of 1.00

What is the purpose of the cross-attention mechanism in the T5 (Text-to-Text Translation Transformer) type of model?

- To capture relationships between tokens in the output sequence
- To integrate positional information into the output embeddings
- To attend to the encoded representations from the encoder ✔
- To mask future tokens during training

Your answer is correct.

The cross-attention mechanism is specifically designed to enable the decoder to use the context provided by the encoder's representations. By attending to these encoded representations, the decoder can generate output tokens based on a comprehensive understanding of the entire input sequence, integrating relevant information from the encoder

The correct answer is:
To attend to the encoded representations from the encoder