In the transformer architecture, which of the following are true regarding cross attention in the decoder?

OPTIONS :

○ The query vectors come from the decoder stack, while the key and valuevectors come from the encoder stack.

○ The query and value vectors come from the decoder stack, while the keyvectors come from the encoder stack.

○ The query, key and value vectors come from the decoder stack.

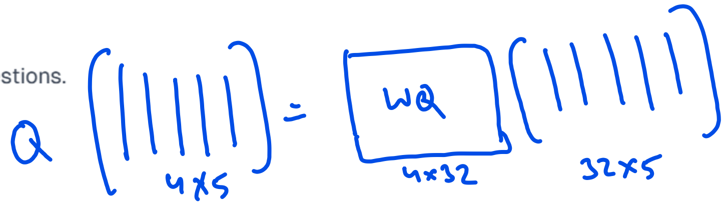
○ The query, key and value vectors come from the encoder stack.

Your score : 0

H. W

Consider the multi-head self-attention mechanism in the encoder of a transformer with 8 heads. The word embedding dimension is 32. In a given head, the query, key and value vectors have the same dimension and each of them is 4. The sequence length is 5.

Based on the above data, answer the given subquestions.



$(AB)^T = B^T A^T$

In any given head, what is the dimension of the WQ matrix?

OPTIONS :

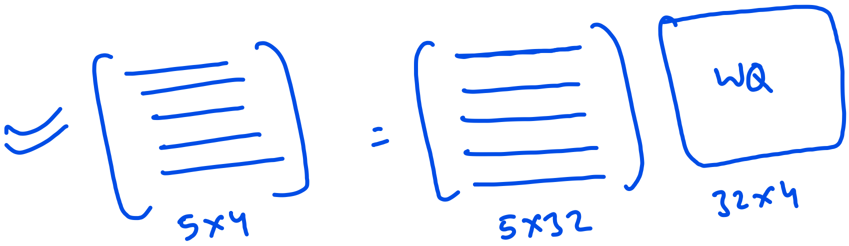
○ 32 × 4

○ 32 × 32

○ 4 × 4

○ 32 × 8

○ 5 × 5



$Q^T K = (4\times5)(5\times4) = 4\times4$

Your score : 0

In any given head, what is the dimension of the query-key product matrix?

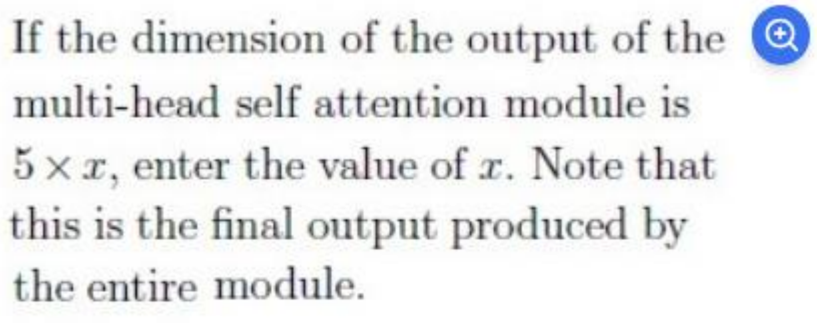
OPTIONS :

○ 5 × 5

○ 4 × 4

○ 8 × 8

○ 32 × 32



If the dimension of the output of the multi-head self attention module is $5 \times x$, enter the value of $x$. Note that this is the final output produced by the entire module.

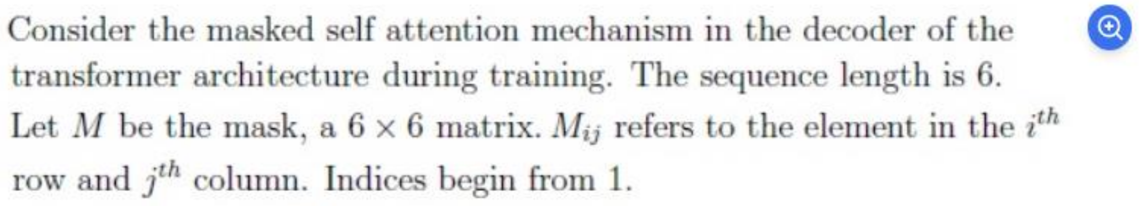
**Answer (Numeric):**

Answer

Accepted Answer : 32

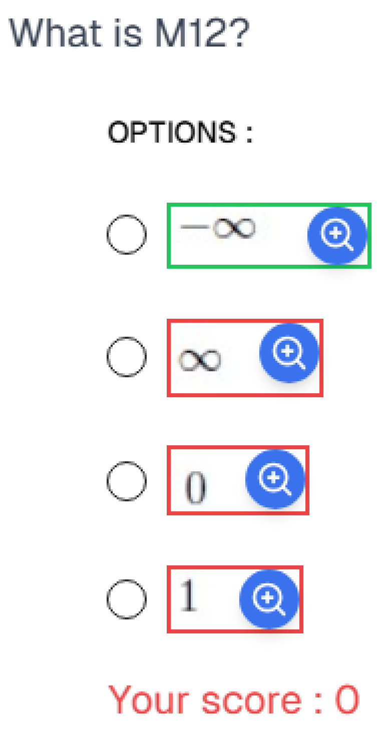
Your score : 0

Based on the above data, answer the given subquestions.



Consider the masked self attention mechanism in the decoder of the transformer architecture during training. The sequence length is 6. Let $M$ be the mask, a $6 \times 6$ matrix. $M_{ij}$ refers to the element in the $i^{th}$ row and $j^{th}$ column. Indices begin from 1.

Your score : 0



What is M12?

OPTIONS :

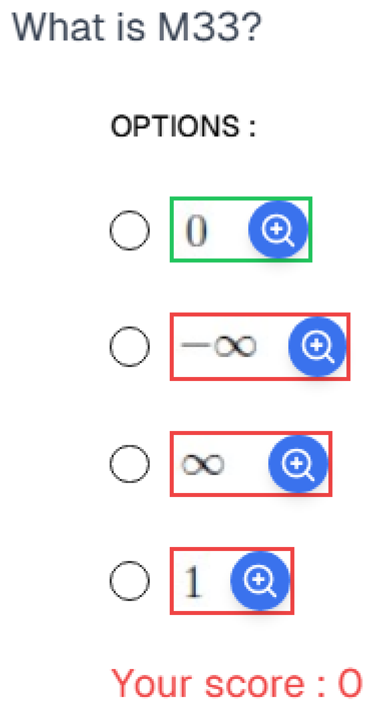
○ $-\infty$

○ $\infty$

○ $0$

○ $1$

Your score : 0



What is M33?

OPTIONS :

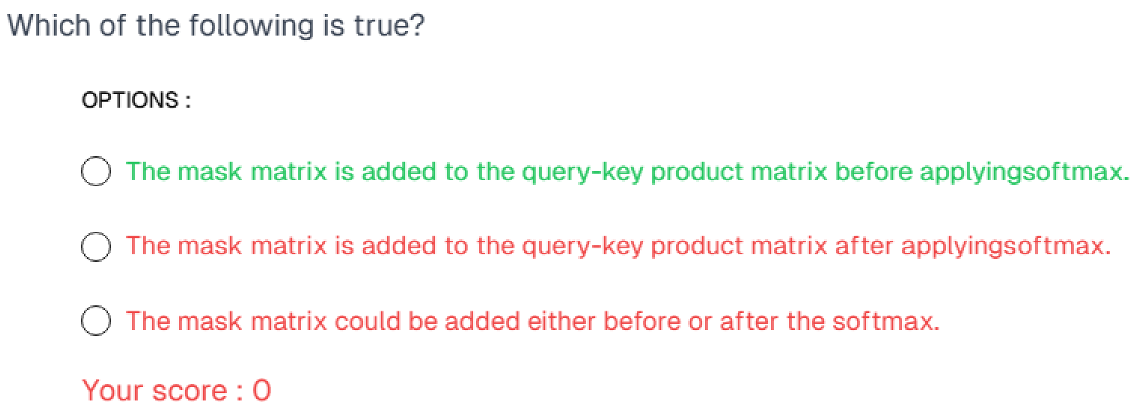
○ $0$

○ $-\infty$

○ $\infty$

○ $1$

Your score : 0



Which of the following is true?

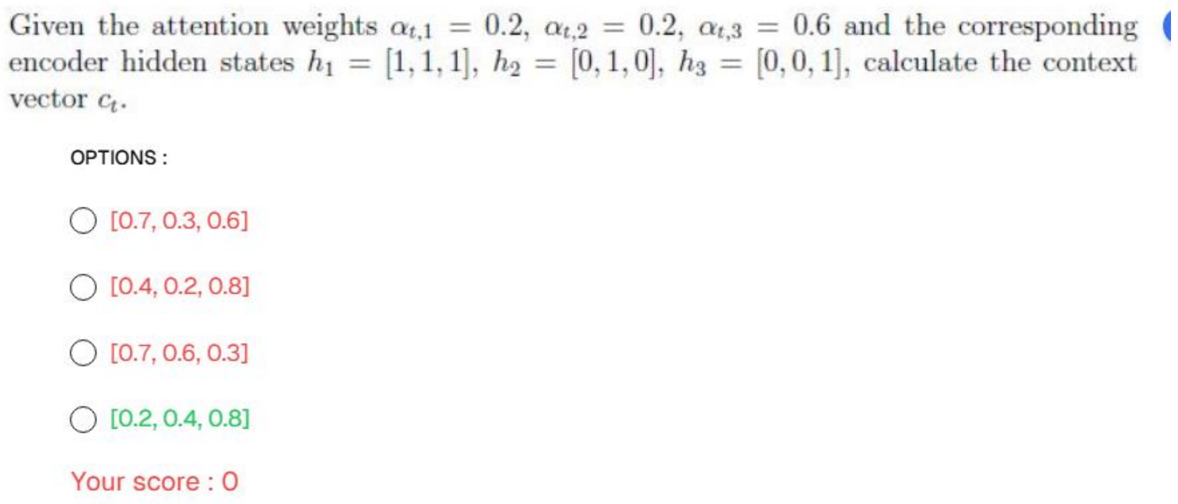
OPTIONS :

○ The mask matrix is added to the query-key product matrix before applyingsoftmax.

○ The mask matrix is added to the query-key product matrix after applyingsoftmax.

○ The mask matrix could be added either before or after the softmax.

Your score : 0



Given the attention weights $\alpha_{t,1} = 0.2$, $\alpha_{t,2} = 0.2$, $\alpha_{t,3} = 0.6$ and the corresponding encoder hidden states $h_1 = [1, 1, 1]$, $h_2 = [0, 1, 0]$, $h_3 = [0, 0, 1]$, calculate the context vector $c_t$.

OPTIONS :

○ [0.7, 0.3, 0.6]

○ [0.4, 0.2, 0.8]

○ [0.7, 0.6, 0.3]

○ [0.2, 0.4, 0.8]

Your score : 0

In the transformer model, how are self-attention weights calculated for a given token?

OPTIONS :

○ By passing the token’s embeddings to softmax function.

○ Using a feedforward neural network applied to the token’s embeddings.

○ Through a convolutional operation applied to the token’s neighborhood.

○ As the dot product between the token’s embeddings and those of all othertokens, followed by a softmax operation.

Your score : 0

In the context of the Transformer model’s encoder-decoder architecture, which of the following statements are accurate?

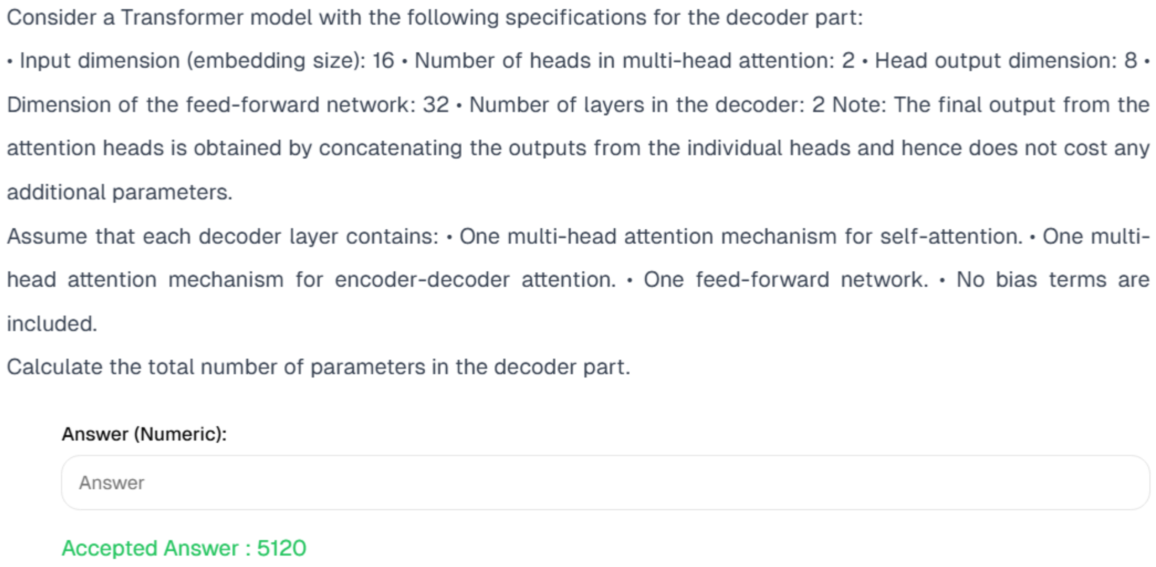
OPTIONS :

□ The encoder in a Transformer model is responsible for converting the inputsequence into a continuous representation that the decoder can then use to generate the output sequence.

□ The decoder in a Transformer model only attends to the encoder’s outputwithout any form of self-attention on its own inputs.

□ Multi-head attention in the encoder allows the model to jointly attend toinformation from different representations at different positions.

□ The decoder’s self-attention mechanism includes a masking component toprevent attending to future positions, ensuring the model generates outputs one step at a time.



Consider a Transformer model with the following specifications for the decoder part:

• Input dimension (embedding size): 16 • Number of heads in multi-head attention: 2 • Head output dimension: 8 • Dimension of the feed-forward network: 32 • Number of layers in the decoder: 2 Note: The final output from the attention heads is obtained by concatenating the outputs from the individual heads and hence does not cost any additional parameters.

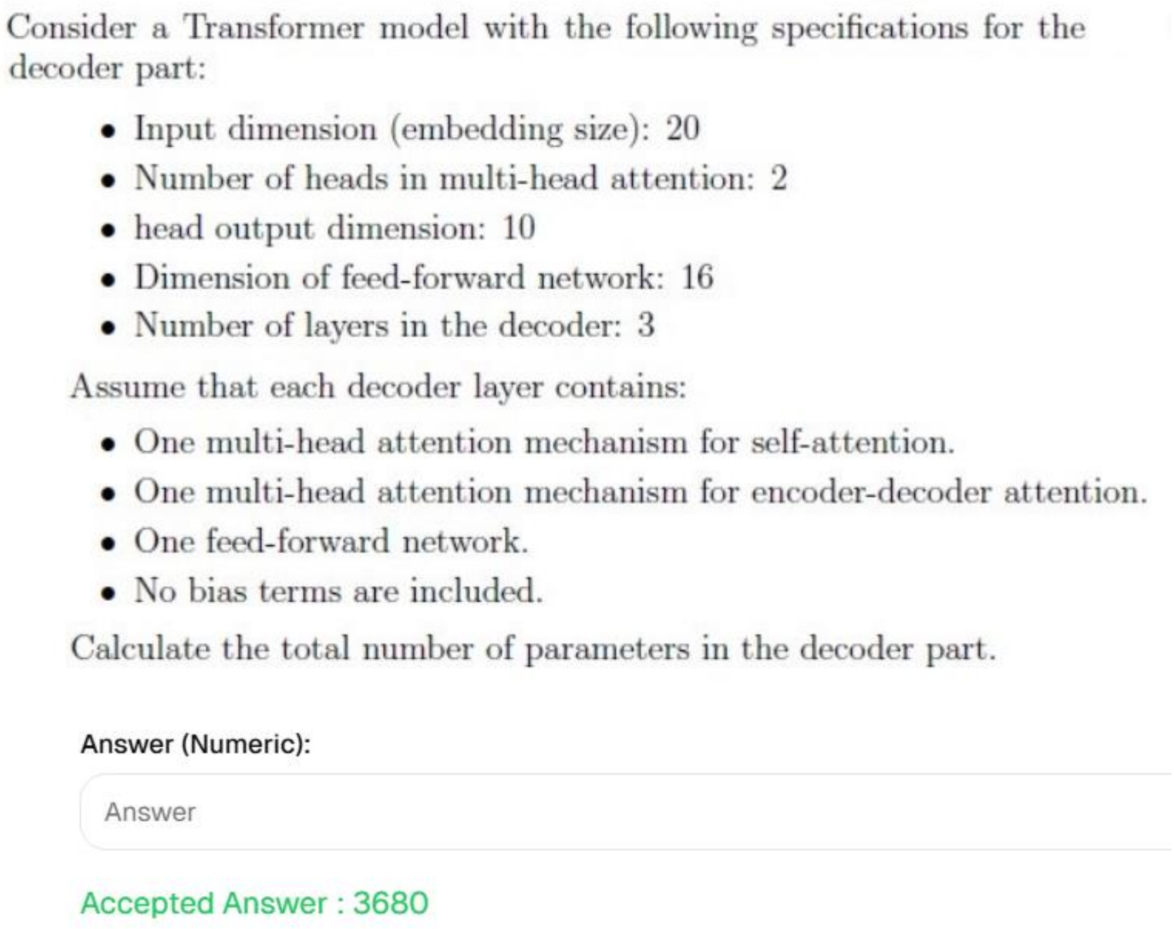
Assume that each decoder layer contains: • One multi-head attention mechanism for self-attention. • One multi-head attention mechanism for encoder-decoder attention. • One feed-forward network. • No bias terms are included.

Calculate the total number of parameters in the decoder part.

Answer (Numeric):

Answer

Accepted Answer : 5120



Consider a Transformer model with the following specifications for the decoder part:

- Input dimension (embedding size): 20
- Number of heads in multi-head attention: 2
- head output dimension: 10
- Dimension of feed-forward network: 16
- Number of layers in the decoder: 3

Assume that each decoder layer contains:

- One multi-head attention mechanism for self-attention.
- One multi-head attention mechanism for encoder-decoder attention.
- One feed-forward network.
- No bias terms are included.

Calculate the total number of parameters in the decoder part.

Answer (Numeric):

Answer

Accepted Answer : 3680

Given the attention weights αt,1 = 0.3, αt,2 = 0.4, αt,3 = 0.3 and the corresponding encoder hidden states h1 = [2, 1, 0], h2 = [1, 2, 1], h3 = [0, 1, 2], calculate the context vector ct.

OPTIONS :

○ [0.7, 1.4, 0.9]

○ [1.1, 1.6, 1.3]

○ [0.6, 1.3, 0.9]

○ [1.1, 1.4, 1.1]

Your score : 0

In the Transformer model, what is the purpose of the multi-head attention mechanism?

OPTIONS :

○ To allow the model to focus on different parts of the input sequence usingdifferent sets of attention weights.

○ To average the attention weights across multiple heads for more stabletraining.

○ To reduce the dimensionality of the input sequence before applying attention.

○ To apply attention in parallel across multiple layers of the Transformer model.

Your score : 0

In the context of the Transformer model's encoder-decoder architecture, which of the following statements are correct?

OPTIONS :

□ The encoder processes the entire input sequence at once at a particular timestep, and its output serves as the context for the decoder during generation.

□ The multi-head attention mechanism in the decoder allows the model to focuson different parts of the encoder's output while generating the sequence.

□ The decoder applies self-attention over its entire sequence of inputs withoutany restrictions, allowing it to consider all future tokens at once.

□ The decoder's self-attention mechanism includes a masking component toprevent attending to future positions, ensuring the model generates outputs one step at a time.

Your score : 0

Suppose we divide the available training samples into mini batches of size 32 to train a model with mini-batch gradient descent. Assume that we have 33 different machines to train the model. One out of 33 machines acts as a master machine. The actual weight update happens in the master machine. The master machine can send one sample for the rest of the machines along with a copy of the model in its current state to compute the gradients. We call this entire set-up parallelization. Which of the following deep learning architectures can be trained in parallel then?

OPTIONS :

□ Fully connected Feed forword neural network

□ Convolutional Neural network

□ Reccurent Neural Network

□ Transformers

Suppose we divide the available training samples into mini batches of size 32 to train a model with mini-batch gradient descent. Assume that we have 33 different machines to train the model. One out of 33 machines acts as a master machine. The actual weight update happens in the master machine. The master machine can send one sample for the rest of the machines along with a copy of the model in its current state to compute the gradients. We call this entire set-up parallelization. Which of the following deep learning architectures can be trained in parallel then?

OPTIONS :

□ Fully connected Feed forword neural network

□ Convolutional Neural network

□ Reccurent Neural Network

□ Transformers

Your score : 0