Congratulations! You passed!

Grade received 80%

Latest Submission Grade 80%

To pass 80% or higher

Go to next item

1. A Transformer Network, unlike its predecessors RNNs, GRUs and LSTMs, can process entire sentences all at the same time. (Parallel architecture). 1 / 1 point

- (•) True
- ( ) False

Expand

Correct
A Transformer Network can ingest entire sentences all at the same time.

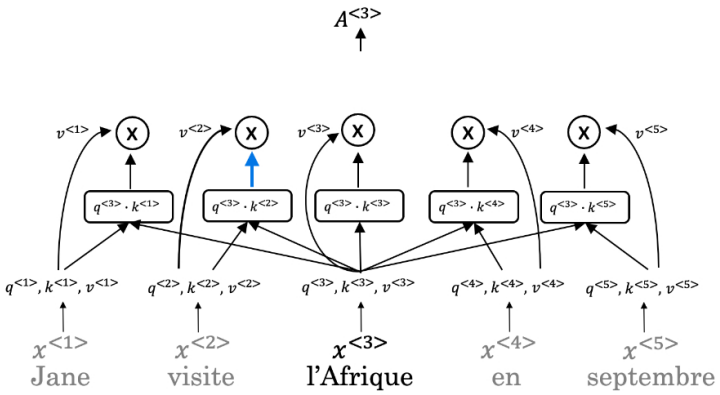
2. The major innovation of the transformer architecture is combining the use of LSTMs and RNN sequential processing. 1 / 1 point

- (•) False
- ( ) True

Expand

Correct
The major innovation of the transformer architecture is combining the use of attention based representations and a CNN convolutional neural network style of processing.

3. The concept of *Self-Attention* is that: 1 / 1 point



- (•) Given a word, its neighbouring words are used to compute its context by summing up the word values to map the Attention related to that given word.
- ( ) Given a word, its neighbouring words are used to compute its context by selecting the highest of those word values to map the Attention related to that given word.
- ( ) Given a word, its neighbouring words are used to compute its context by selecting the lowest of those word values to map the Attention related to that given word.
- ( ) Given a word, its neighbouring words are used to compute its context by taking the average of those word values to map the Attention related to that given word.

Expand

Correct

4. Which of the following correctly represents *Attention ?* 1 / 1 point

- $Attention(Q, K, V) = min(\frac{QV^T}{\sqrt{d_k}})K$
- $Attention(Q, K, V) = softmax(\frac{QV^T}{\sqrt{d_k}})K$
- $Attention(Q, K, V) = min(\frac{QK^T}{\sqrt{d_k}})V$
- $Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$

Expand

Correct

5. Are the following statements true regarding Query (Q), Key (K) and Value (V)? 0 / 1 point

Q = interesting questions about the words in a sentence

K = qualities of words given a Q

V = specific representations of words given a Q

- True
- False

Expand

Incorrect
To revise the concept watch the lecture ; Q = interesting questions about the words in a sentence, K = qualities of words given a Q, V = specific representations of words given a Q

6. $\boldsymbol{Attention(W_i^Q Q, W_i^K K, W_i^V V)}$ 1 / 1 point

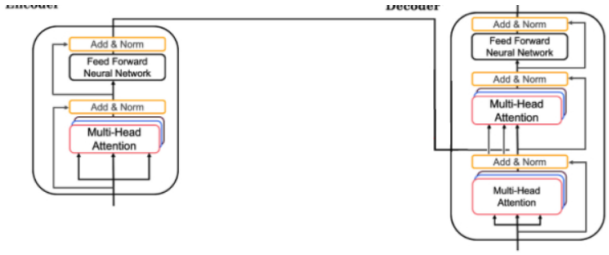
What does i represent in this multi-head attention computation?

- The computed attention weight matrix associated with the *ith* "head" (sequence)
- The computed attention weight matrix associated with specific representations of words given a Q
- The computed attention weight matrix associated with the order of the words in a sentence
- The computed attention weight matrix associated with the *ith* "word" in a sentence.

Expand

Correct
i here represents the computed attention weight matrix associated with the "head" (sequence).

7. Following is the architecture within a Transformer Network ***(without displaying positional encoding and output layers(s)).*** 1 / 1 point



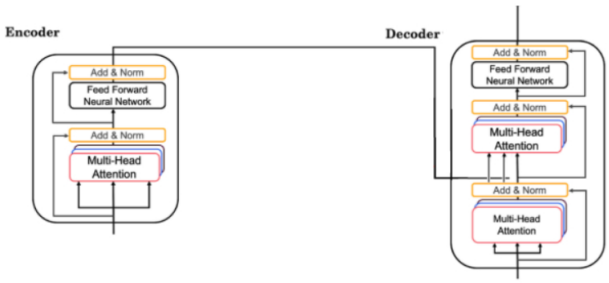
What is ***NOT*** necessary for the *Decoder's* second block of *Multi-Head Attention*?

- (●) All of the above are necessary for the Decoder's second block.
- ( ) Q
- ( ) V
- ( ) K

Expand

Correct
The first block's output is used to generate the Q matrix for the next Multi-Head Attention block. The Decoder also uses K and V from the Encoder for its second block of Multi-Head Attention.

8. Following is the architecture within a Transformer Network ***(without displaying positional encoding and output layers(s)).*** **1 / 1 point**



The output of the decoder block contains a softmax layer followed by a linear layer to predict the next word one word at a time.

- ( ) True
- (●) False

Expand

Correct
The output of the decoder block contains a linear layer followed by a softmax layer to predict the next word one word at a time.

9. Which of the following statements is true? **0 / 1 point**

- ( ) The transformer network differs from the attention model in that only the transformer network contains positional encoding.
- (●) The transformer network differs from the attention model in that only the attention model contains positional encoding.
- ( ) The transformer network is similar to the attention model in that neither contain positional encoding.

- The transformer network is similar to the attention model in that both contain positional encoding.

Expand

Incorrect
To revise the concept watch the lecture .

10. Which of these is a good criterion for a good positionial encoding algorithm? **1 / 1 point**

- The algorithm should be able to generalize to longer sentences.
- It must be nondeterministic.
- It should output a common encoding for each time-step (word's position in a sentence).
- Distance between any two time-steps should be inconsistent for all sentence lengths.

Expand

Correct
This is a good criterion for a good positional encoding algorithm.