Your grade: 100%

Your latest: **100%** • Your highest: **100%** • To pass you need at least 80%. We keep your highest score.

Next item →

1. A Transformer Network, like its predecessors RNNs, GRUs and LSTMs, can process information one word at a time. (Sequential architecture). 1 / 1 point

- (•) False
- ( ) True

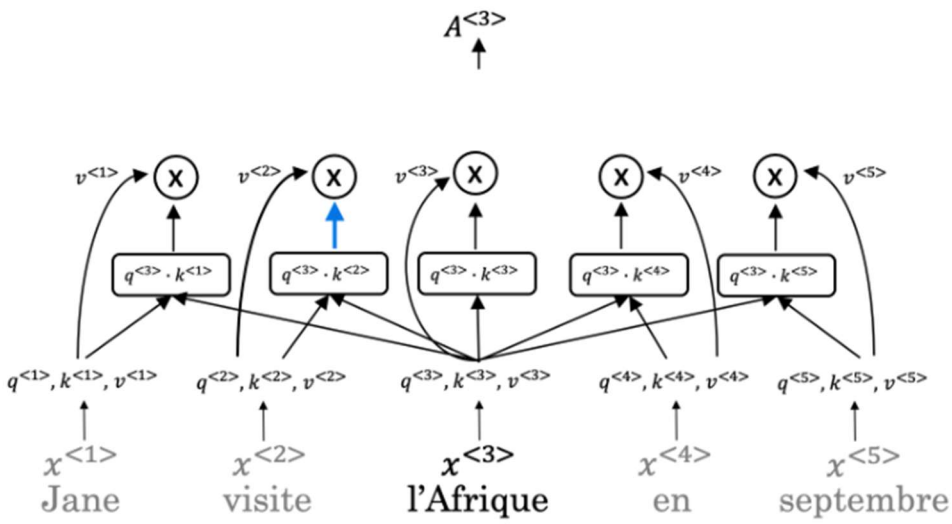
Correct
Correct! A Transformer Network can ingest entire sentences all at the same time.

2. Transformer Network methodology is taken from: 1 / 1 point

- ( ) GRUs and LSTMs
- ( ) RNN and LSTMs
- ( ) Attention Mechanism and RNN style of processing.
- (•) Attention Mechanism and CNN style of processing.

Correct
Transformer architecture combines the use of attention based representations and a CNN convolutional neural network style of processing.

3. What are the key inputs to computing the attention value for each word?



- ( ) The key inputs to computing the attention value for each word are called the quotation, knowledge, and value.
- (•) The key inputs to computing the attention value for each word are called the query, key, and value.
- ( ) The key inputs to computing the attention value for each word are called the quotation, key, and vector.
- ( ) The key inputs to computing the attention value for each word are called the query, knowledge, and vector.

Correct
The key inputs to computing the attention value for each word are called the query, key, and value.

4. What letter does the "?" represent in the following representation of *Attention*?

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_?}})V$$

- ○ t
- ◉ k
- ○ v
- ○ q

Correct
k is represented by the ? in the representation.

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

Q = interesting questions about the words in a sentence

K = specific representations of words given a Q

V = qualities of words given a Q

- ◉ False
- ○ True

Correct
Correct! Q = interesting questions about the words in a sentence, K = qualities of words given a Q, V = specific representations of words given a Q

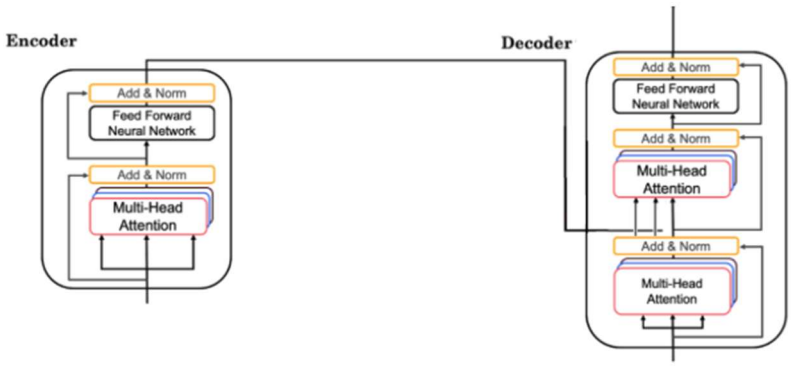
$$\boldsymbol{Attention(W_i^Q Q, W_i^K K, W_i^V V)}$$

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

- ◉ True
- ○ False

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

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

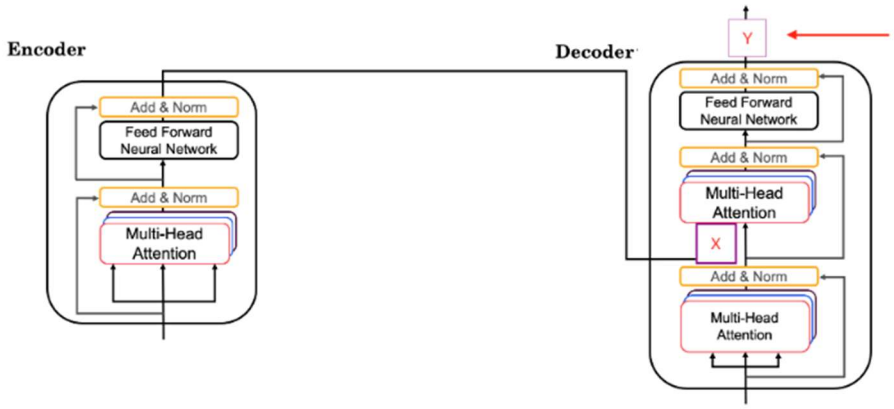


What is ***NOT*** necessary for the *Decoder's* second block of *Multi-Head Attention*?

- ○ K
- ○ V
- ◉ All of the above are necessary for the Decoder's second block.
- ○ Q

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))***



What is the output layer(s) of the *Decoder* ? (Marked Y, pointed by the independent arrow)

- ○ Softmax layer
- ○ Softmax layer followed by a linear layer.
- ○ Linear layer
- ◉ Linear layer followed by a softmax layer.

Correct

9. Which of the following statements is true?

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

Correct
Positional encoding allows the transformer network to offer an additional benefit over the attention model.

10. Which of these is a good criterion for a good positionial encoding algorithm?

- ( ) 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.
- (●) The algorithm should be able to generalize to longer sentences.
- ( ) It must be nondeterministic.

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