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

- ○ True
- ◉ False

Expand

Correct

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

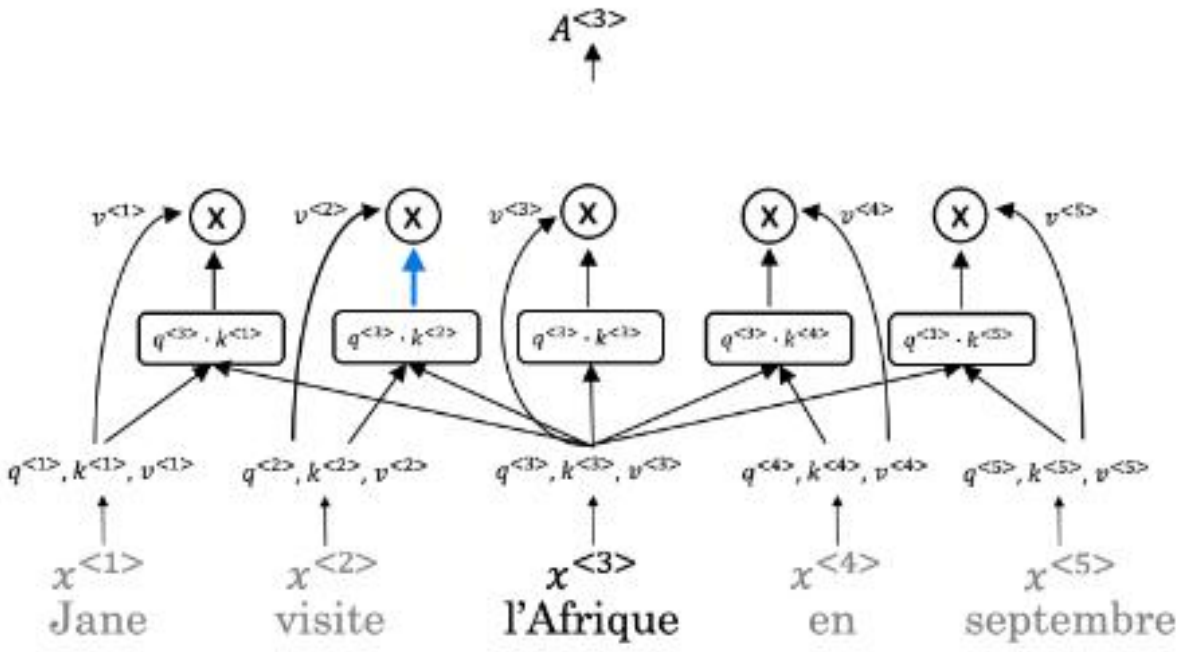
2. The major innovation of the transformer architecture is combining the use of LSTMs and RNN sequential processing. 1 / 1 point

- (●) 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. How does the Self-Attention mechanism of transformers use neighboring words to compute a word's context? 1 / 1 point




- Selecting the maximum word values to map the Attention related to that given word.
- Selecting the minimum word values to map the Attention related to that given word.
- Summation of the word values to map the Attention related to that given word.
- Multiplication of the word values to map the Attention related to that given word.

Expand

Correct
Given a word, its neighboring words are used to compute its context by summing up the word values to map the Attention related to that given word.

4. What letter does the "?" represent in the following representation of *Attention*? 1 / 1 point

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

- (●) k
- ( ) q
- ( ) t
- ( ) v

Expand

Correct
k is represented by the ? in the representation.

5. Which of the following statements represents Key (K) as used in the self-attention calculation? 1 / 1 point

- (●) K = qualities of words given a Q
- ( ) K = specific representations of words given a Q
- ( ) K = the order of the words in a sentence
- ( ) K = interesting questions about the words in a sentence

Expand

Correct
The qualities of words given a Q are represented by Key (K).

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

1 / 1 point

$i$ here represents the computed attention weight matrix associated with the $ith$ "word" in a sentence.

- ○ True
- ◉ False

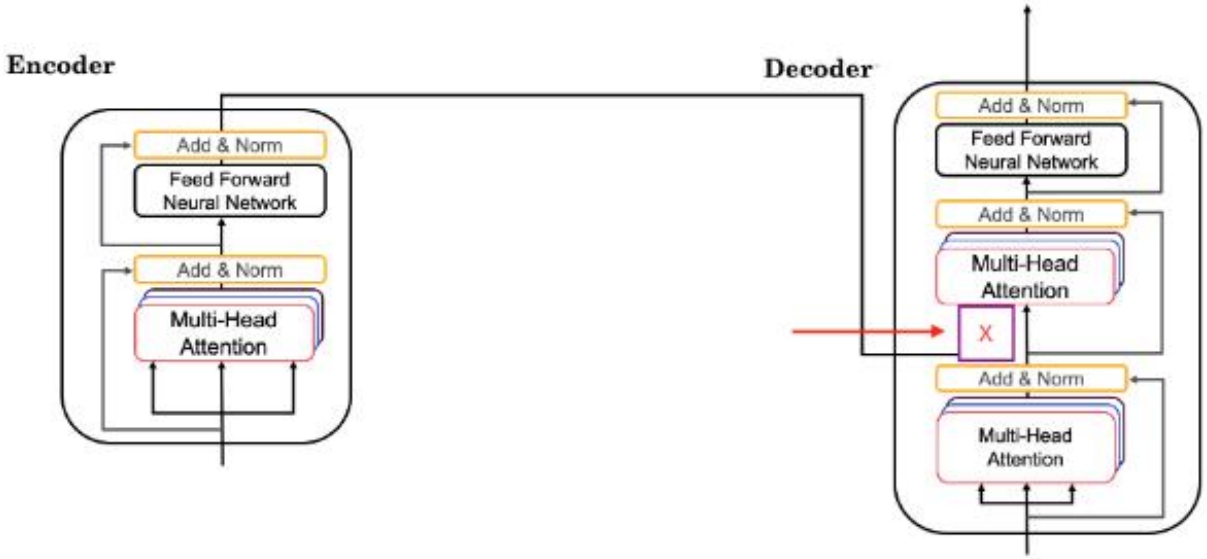
Expand

Correct

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)).***

1 / 1 point




What information does the *Decoder* take from the *Encoder* for its second block of *Multi-Head Attention*? (Marked $X$, pointed by the independent arrow)

(Check all that apply)

- [x] K

  ✓ Correct

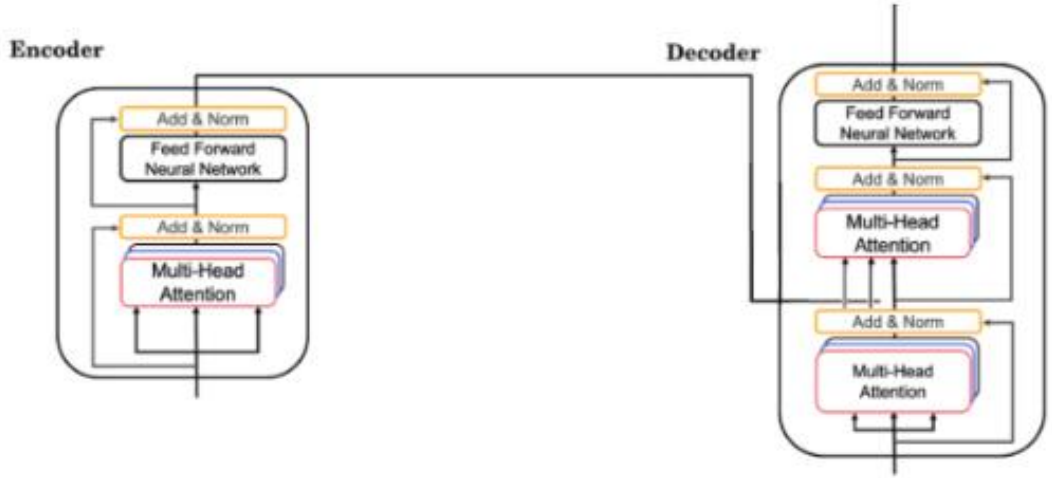
- [ ] Q

- [x] V

  ✓ Correct

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

1 / 1 point




What does the output of the *encoder* block contain?

- (●) Contextual semantic embedding and positional encoding information
- ( ) Linear layer followed by a softmax layer.
- ( ) Prediction of the next word.
- ( ) Softmax layer followed by a linear layer.

Expand

✓ Correct

The output of the  block contains contextual semantic embedding and positional encoding information.

9. Which of the following statements is true? 1 / 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 both contain positional encoding.
- ( ) The transformer network is similar to the attention model in that neither contain positional encoding.

Expand

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 positional encoding algorithm? 1 / 1 point

- [x] It should output a unique encoding for each time-step (word's position in a sentence).

✓ Correct

- [x] Distance between any two time-steps should be consistent for all sentence lengths.

✓ Correct

- [x] The algorithm should be able to generalize to longer sentences.

✓ Correct

- [ ] None of these.

Expand

Correct
Great, you got all the right answers.