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

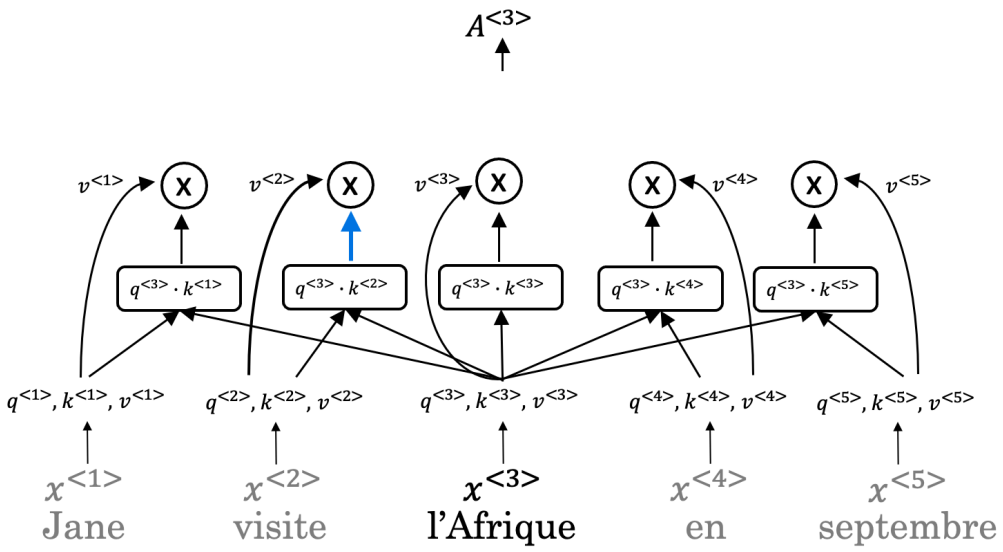
Correct
Correct! A Transformer Network can ingest entire sentences all at the same time.

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

- (●) False
- ( ) True

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

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

- ○ $A(Q,K,V) = \left(\frac{\exp(q*k^{<i>})}{\exp(q*k^{<j>})}\right) * V^{<i>}$
- ◉ $A(Q,K,V) = \sum_i \left(\frac{\exp(q*k^{<i>})}{\sum_j \exp(q*k^{<j>})}\right) * V^{<i>}$

- $A(Q,K,V) = \sum_i (\frac{\exp(q*v^{<i>})}{\sum_j \exp(q*v^{<j>})}) * K^{<i>}$
- $A(Q,K,V) = \sum_i (\frac{\exp(q*k^{<i>})}{\sum_j \exp(q*k^{<j>})}) * \sum_i v^i$

Correct
This is the correct Attention formula.

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

- False
- True

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

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

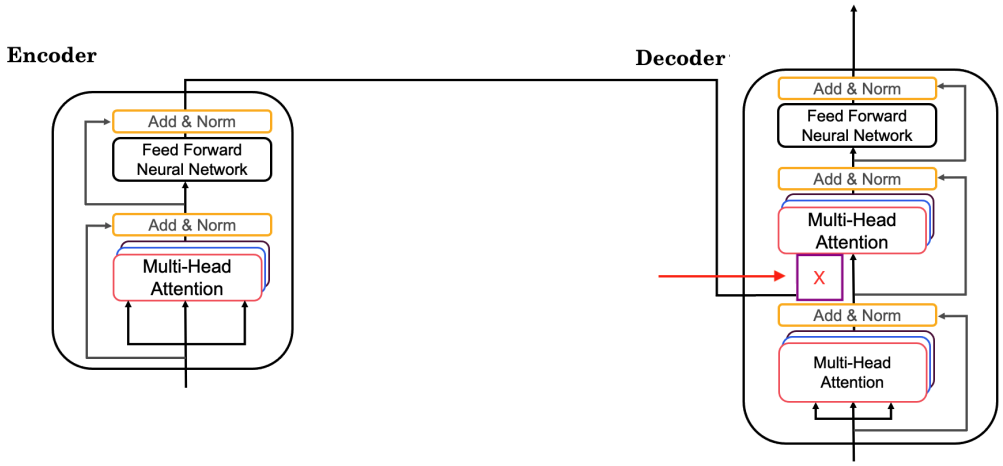
6. i here represents the computed attention weight matrix associated with the ith "head" (sequence).

- 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)).*** **1 / 1 point**



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

(Check all that apply)

- [x] K

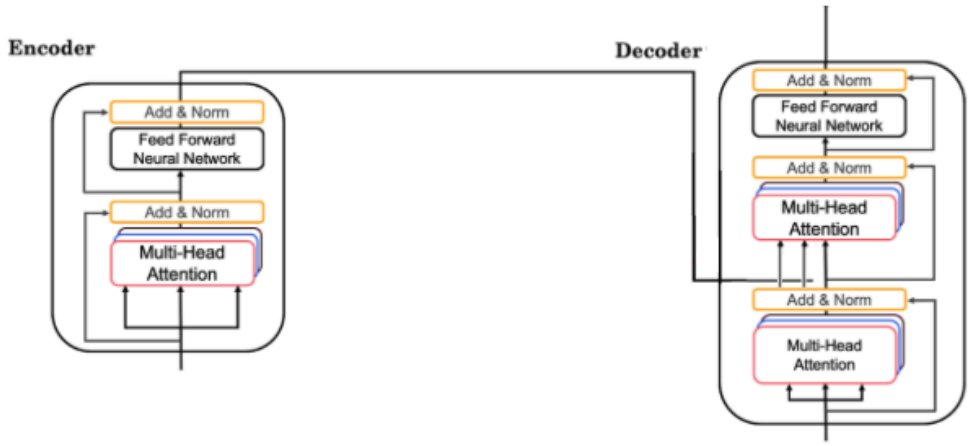
Correct

- [x] V

Correct

- [ ] Q

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
- ( ) Prediction of the next word.
- ( ) Softmax layer followed by a linear layer.
- ( ) Linear layer followed by a softmax layer.

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

9. Which of the following statements is true about positional encoding? Select all that apply. 1 / 1 point

- [x] Positional encoding uses a combination of sine and cosine equations.

Correct

This is a correct answer, but other options are also correct. To review the concept watch the lecture *Transformer Network*.

- [x] Positional encoding is important because position and word order are essential in sentence construction of any language.

Correct
This is a correct answer, but other options are also correct. To review the concept watch the lecture *Transformer Network*.

- [x] Positional encoding provides extra information to our model.

Correct
This is a correct answer, but other options are also correct. To review the concept watch the lecture *Transformer Network*.

- [ ] Positional encoding is used in the transformer network and the attention model.

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

- [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.