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, 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.

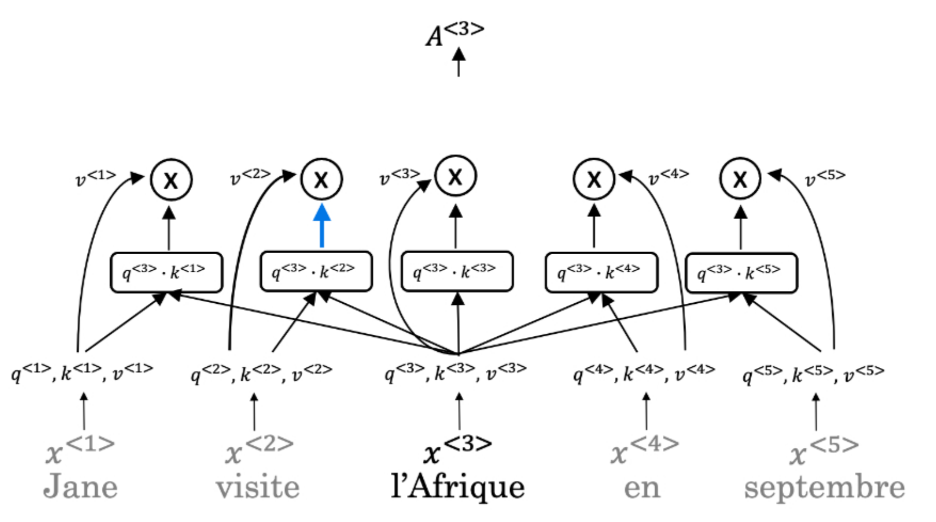
2. Transformer Network methodology is taken from: — 1 / 1 point

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

Expand

Correct
Transformer architecture combines 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



- (•) 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.
- ( ) 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.

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

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

Expand

Correct

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

Expand

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

$$Attention(W_i^Q Q, W_i^K K, W_i^V V)$$

1 / 1 point

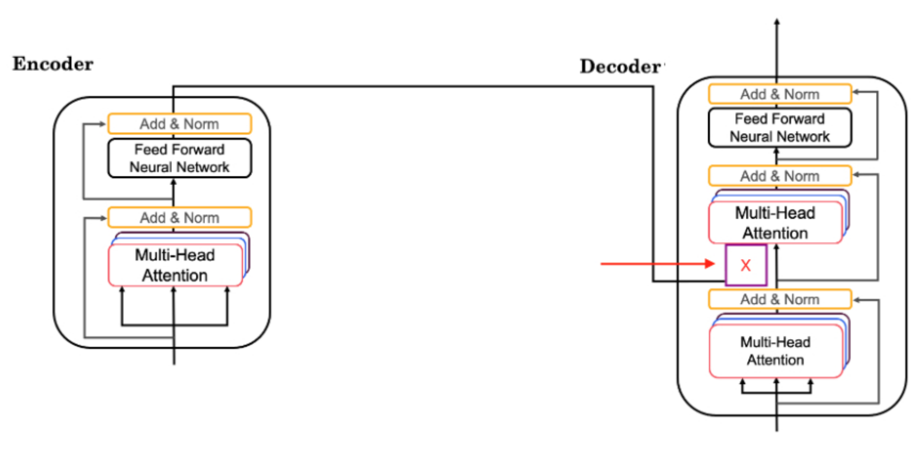
6. i here represents the computed attention weight matrix associated with the ith "head" (sequence).

- (•) True
- ( ) False

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 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] V

  ✓ Correct

- [x] K

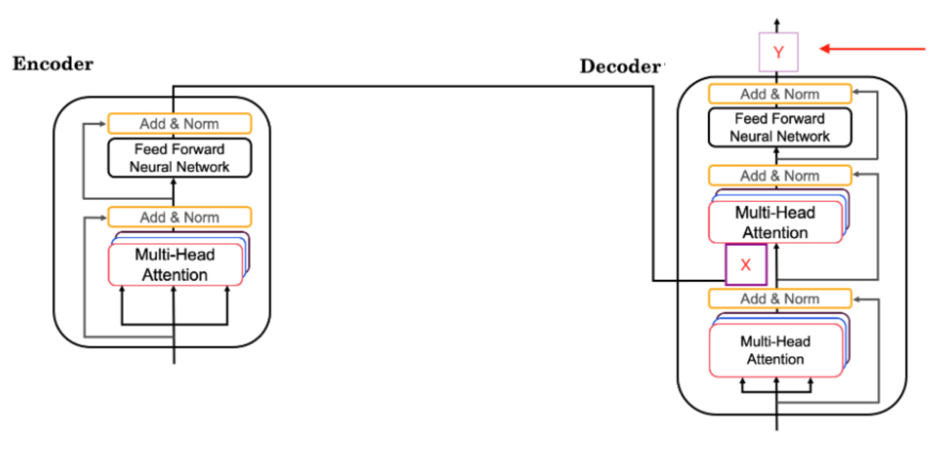
  ✓ Correct

- [ ] Q

Expand

Correct
Great, you got all the right answers.

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



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

- ( ) Softmax layer
- (•) Linear layer followed by a softmax layer.
- ( ) Softmax layer followed by a linear layer.
- ( ) Linear layer

Expand

Correct

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

- ( ) The transformer network is similar to the attention model in that both contain 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 differs from the attention model in that only the transformer network contains 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

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

Expand

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