Congratulations! You passed!

Grade received 90% **Latest Submission Grade** 90% **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**

- False
- True

Expand

Correct

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

2. Transformer Network methodology is taken from:

1 / 1 point

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

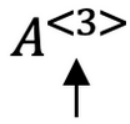
Expand

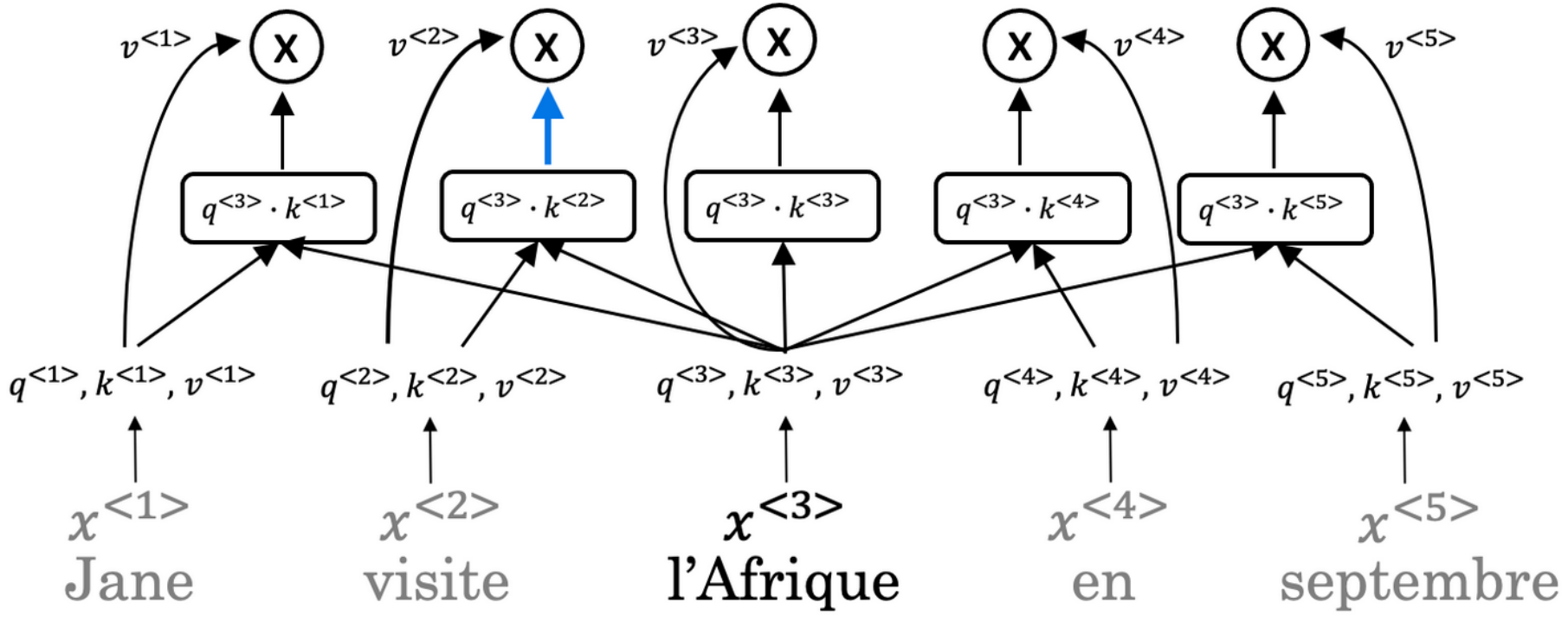
Correct

Transformer architecture combines 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




$v^{<1>}$
$v^{<2>}$
$v^{<3>}$
$v^{<4>}$
$v^{<5>}$
X
$q^{<3>} \cdot k^{<1>}$
$q^{<3>} \cdot k^{<2>}$
$q^{<3>} \cdot k^{<3>}$
$q^{<3>} \cdot k^{<4>}$
$q^{<3>} \cdot k^{<5>}$
$q^{<1>}, k^{<1>}, v^{<1>}$
$q^{<2>}, k^{<2>}, v^{<2>}$
$q^{<3>}, k^{<3>}, v^{<3>}$
$q^{<4>}, k^{<4>}, v^{<4>}$
$q^{<5>}, k^{<5>}, v^{<5>}$
$x^{<1>}$ Jane
$x^{<2>}$ visite
$x^{<3>}$ l'Afrique
$x^{<4>}$ en
$x^{<5>}$ septembre

- 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.
- 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.

Expand

Correct

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
- v
- t

Expand

Correct

k is represented by the ? in the representation.

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

- (●) True
- ( ) False

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

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

1 / 1 point

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

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

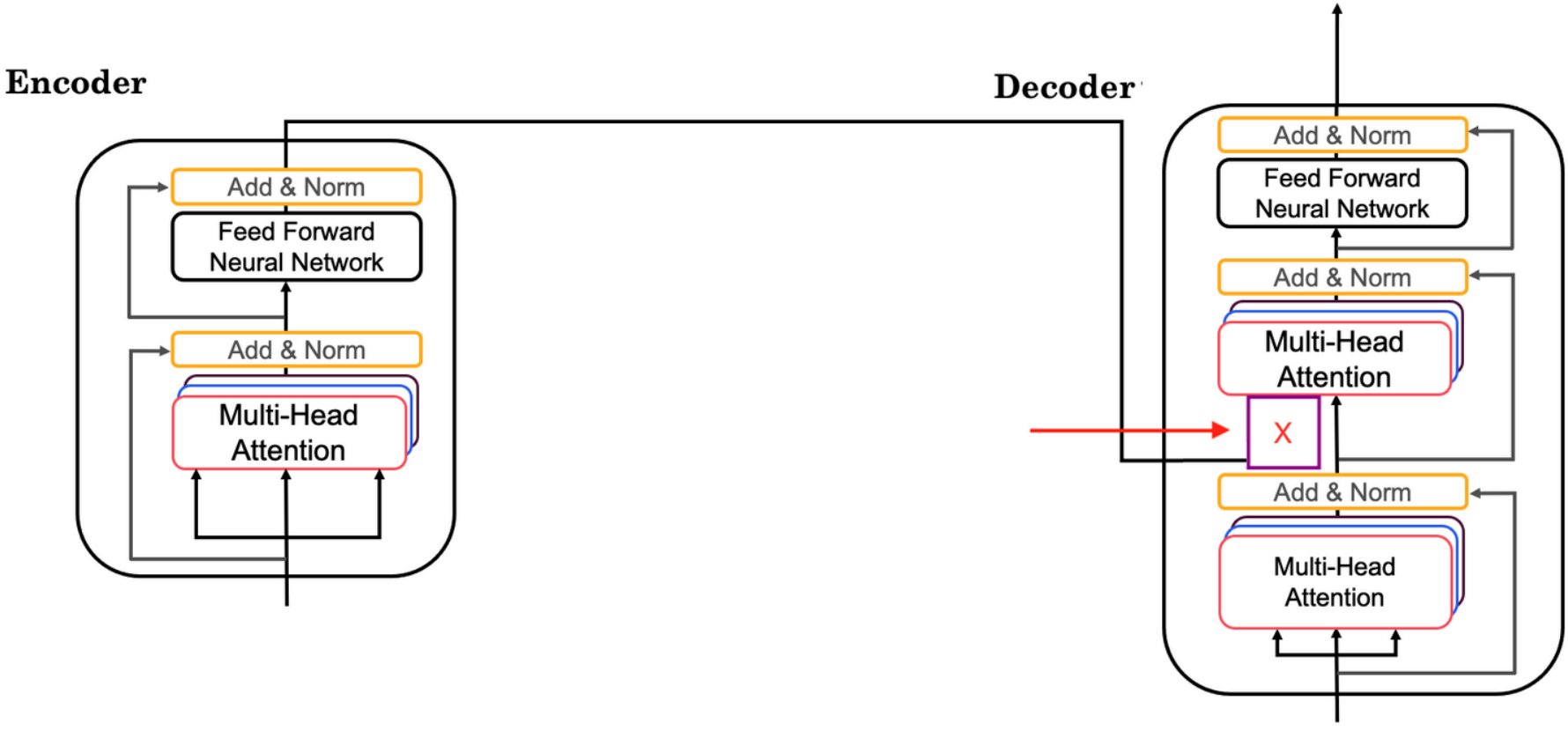
Expand

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

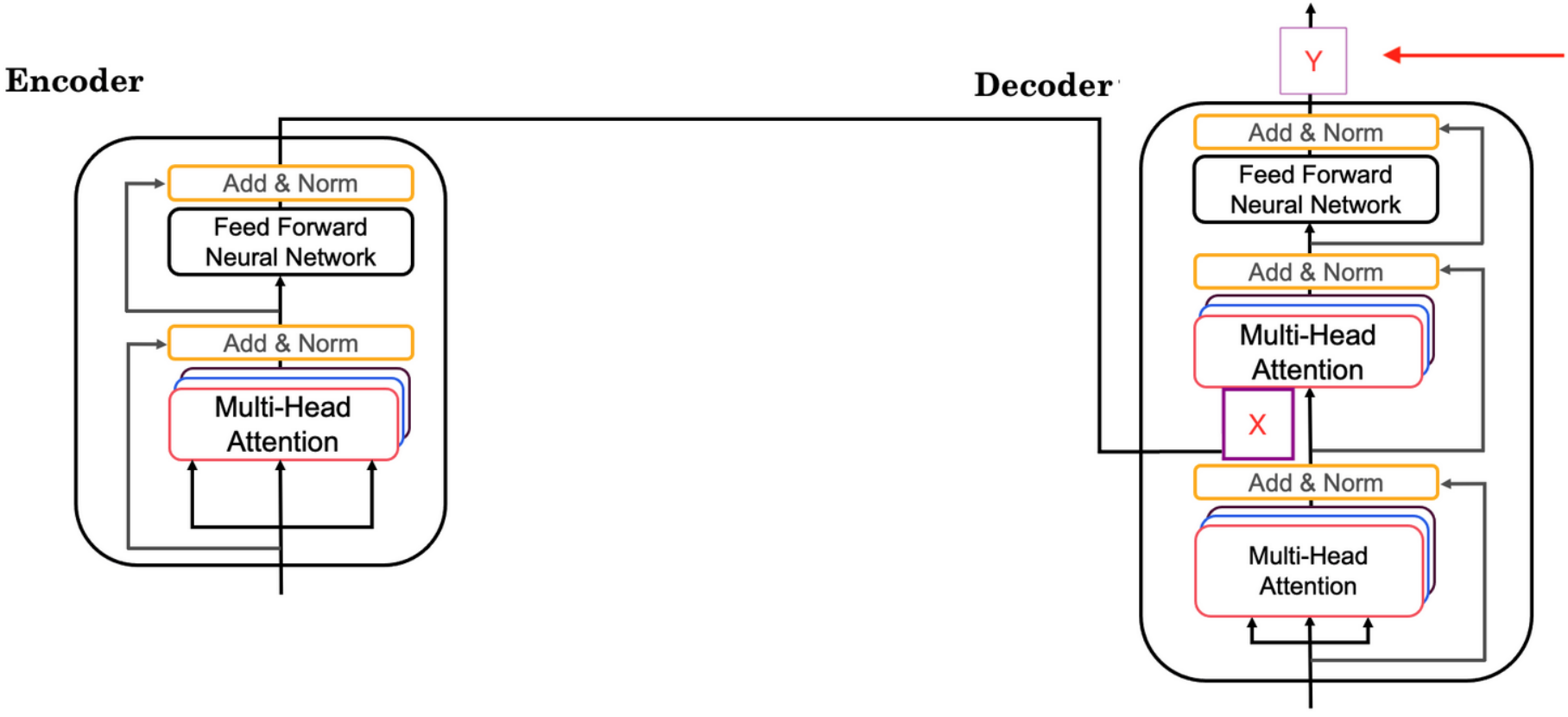
Expand

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)

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

Expand

Correct

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

0 / 1 point

- [ ] Positional encoding is used in the transformer network and the attention model.
- [ ] Positional encoding uses a combination of sine and cosine equations.
- [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*.

- [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*.

Expand

Incorrect

You didn't select all the correct answers

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.