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: (Check all that apply) 0 / 1 point

- [x] Convolutional Neural Network style of architecture.

! This should not be selected

- [ ] None of these.
- [x] Attention mechanism.

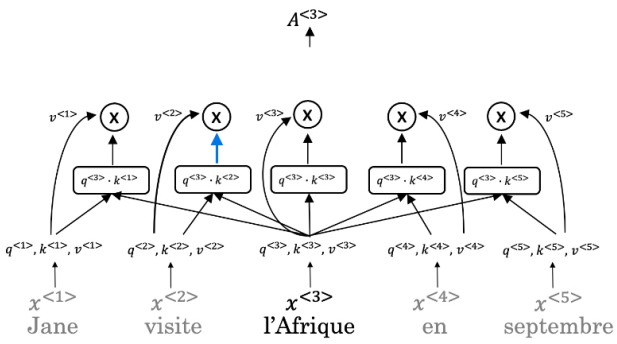
Correct

- [ ] Convolutional Neural Network style of processing.

Expand

Incorrect
You didn't select all the correct answers

3. What are the key inputs to computing the attention value for each word? 1 / 1 point



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

Expand

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*? **1 / 1 point**

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

- v
- q
- k
- 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 = specific representations of words given a Q

V = qualities of words given a Q

- False
- True

Expand

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

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

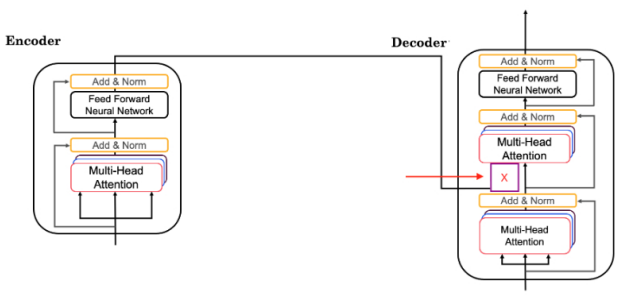
What does i represent in this multi-head attention computation?

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

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

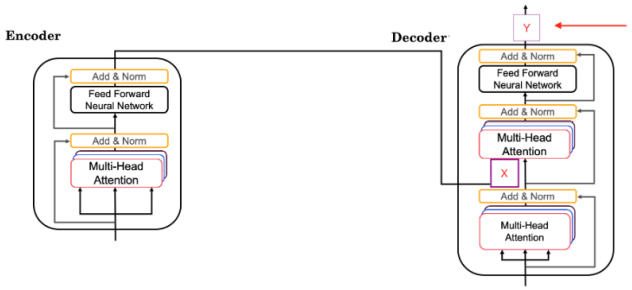
- [ ] Q
- [x] K

Correct

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

Expand

Correct

9. Why is positional encoding important in the translation process? (Check all that apply) 1 / 1 point

- [x] Position and word order are essential in sentence construction of any language.

Correct

- [ ] It helps to locate every word within a sentence.
- [ ] It is used in CNN and works well there.
- [x] Providing extra information to our model.

Correct

Expand

Correct
Great, you got all the right answers.

10. Which of these is ***not*** a good criterion for a good positional encoding algorithm? 1 / 1 point

- ( ) Distance between any two time-steps should be consistent for all sentence lengths.

It should output a common encoding for each time-step (word's position in a sentence).

It must be deterministic.

The algorithm should be able to generalize to longer sentences.

Expand

Correct

This is not a good criterion for a good positional encoding algorithm.