1. A Transformer Network processes sentences from left to right, one word at a time.

- (•) False
- ( ) True

Expand

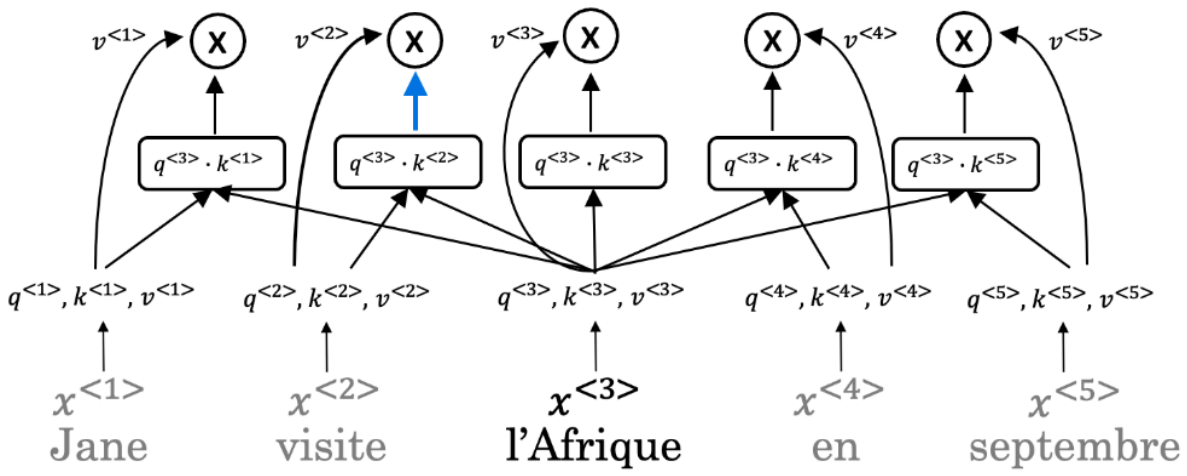
Correct
A Transformer Network can ingest entire sentences all at the same time.

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

- (•) 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



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

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

- ( ) v
- (●) k
- ( ) q
- ( ) t

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?

- ○ 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

✗ **Incorrect**

To revise the concept watch the lecture ; V = specific representations of words given a Q

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

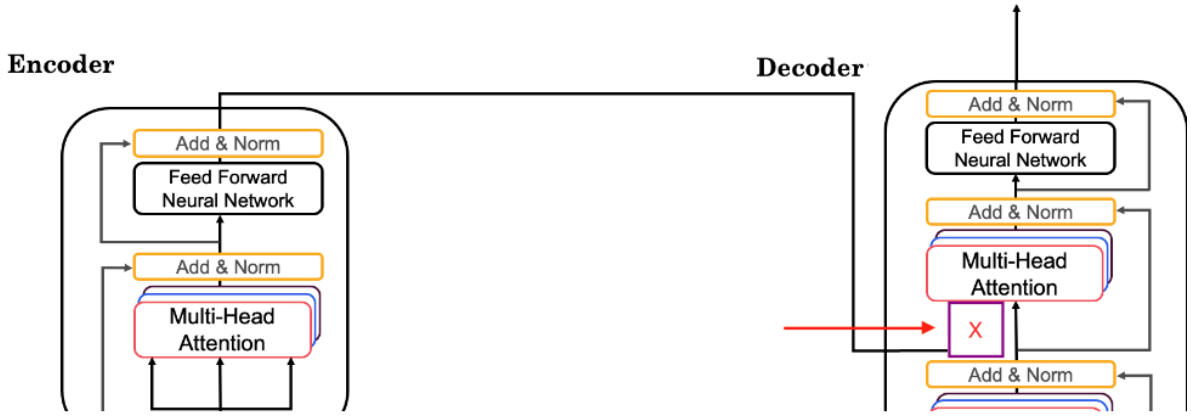
What does i represent in this multi-head attention computation?

- ○ 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)
- ○ The computed attention weight matrix associated with the *ith* "word" in a sentence.

Expand

✓ **Correct**

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



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)

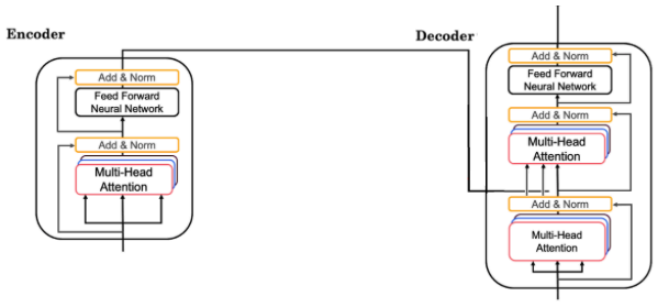
- [ ] Q
- [ ] V
- [x] K

✓ Correct

Expand

✗ **Incorrect**
You didn't select all the correct answers

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



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

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

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

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

- [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**

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

- (●) 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).
- ( ) It must be nondeterministic.
- ( ) Distance between any two time-steps should be inconsistent for all sentence lengths.

Expand

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