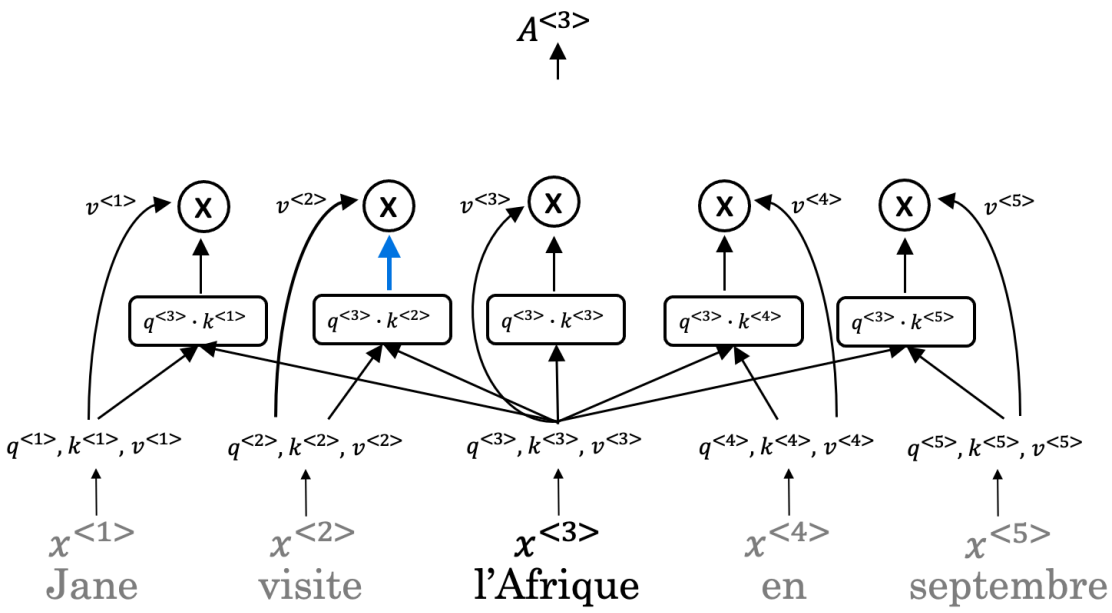
A Transformer Network processes sentences from left to right, one word at a time. 1 point

- False
- True

2.Transformer Network methodology is taken from: (Check all that apply) 1 point

- [ ] None of these.
- [ ] Convolutional Neural Network style of processing.
- [ ] Attention mechanism.
- [ ] Convolutional Neural Network style of architecture.

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



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

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

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

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

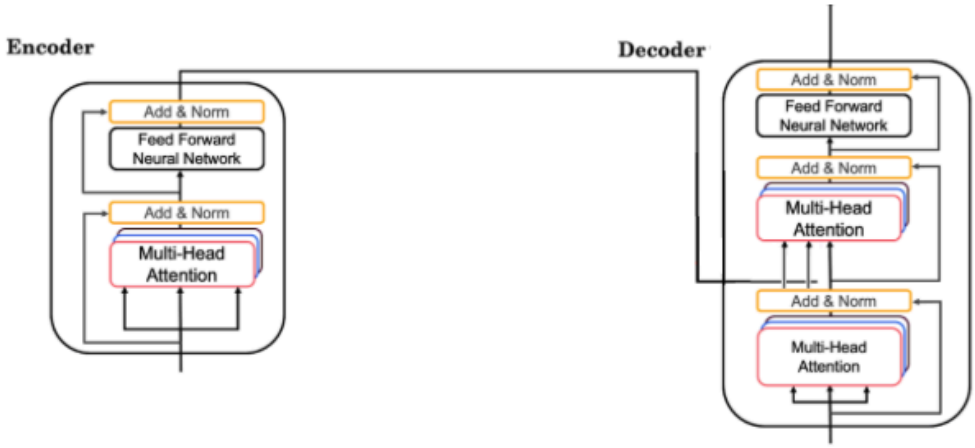
- True
- False

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

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

- False
- True

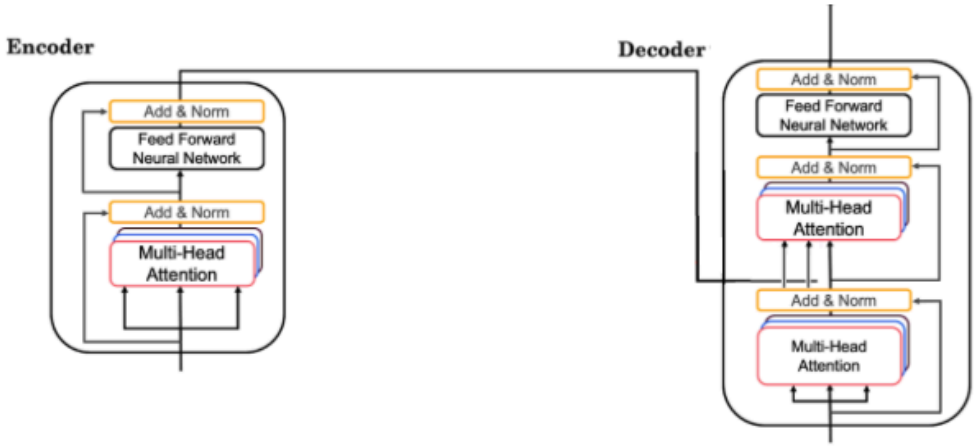
7.Following is the architecture within a Transformer Network *(without displaying positional encoding and output layers(s)).* 1 point



What is generated from the output of the *Decoder's* first block of *Multi-Head Attention*?

- Q
- V
- K

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



The output of the decoder block contains a softmax layer followed by a linear layer to predict the next word one word at a time.

- ○ False
- ○ True

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

- ☐ Position and word order are essential in sentence construction of any language.
- ☐ It helps to locate every word within a sentence.
- ☐ It is used in CNN and works well there.
- ☐ Providing extra information to our model.

10.Which of these is a good criterion for a good positional encoding algorithm? 1 point

- ☐ It should output a unique encoding for each time-step (word's position in a sentence).
- ☐ Distance between any two time-steps should be consistent for all sentence lengths.
- ☐ The algorithm should be able to generalize to longer sentences.
- ☐ None of these.