C5W4-Quiz-Transformers

1. A Transformer Network, like its predecessors RNNs, GRUs and LSTMs, can process information one word at a time. (Sequential architecture).

   - ☑ ~~False~~
   - ☐ **True**

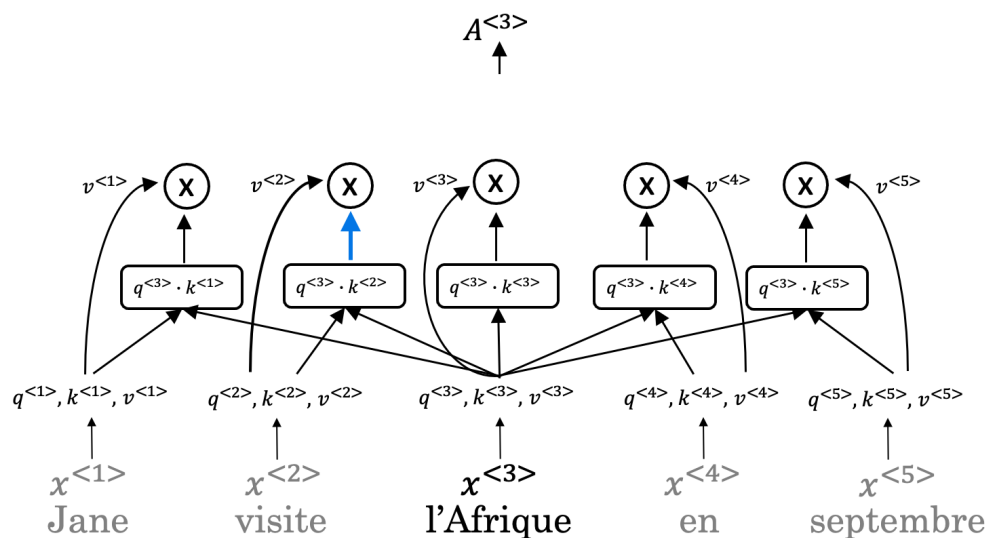2. Transformer Network methodology is taken from: (Check all that apply)

   - ☑ ~~Convolutional Neural Network style of processing.~~
   - ☐ **None of these.**
   - ☐ **Convolutional Neural Network style of architecture.**
   - ☑ ~~Attention mechanism.~~

3. The concept of *Self-Attention* is that:

$$A^{<3>}$$
$$\uparrow$$

$v^{<1>}$ Ⓧ $v^{<2>}$ Ⓧ $v^{<3>}$ Ⓧ Ⓧ $v^{<4>}$ Ⓧ $v^{<5>}$

$q^{<3>} \cdot k^{<1>}$ $q^{<3>} \cdot k^{<2>}$ $q^{<3>} \cdot k^{<3>}$ $q^{<3>} \cdot k^{<4>}$ $q^{<3>} \cdot k^{<5>}$

$q^{<1>}, k^{<1>}, v^{<1>}$ $q^{<2>}, k^{<2>}, v^{<2>}$ $q^{<3>}, k^{<3>}, v^{<3>}$ $q^{<4>}, k^{<4>}, v^{<4>}$ $q^{<5>}, k^{<5>}, v^{<5>}$

$x^{<1>}$ $x^{<2>}$ $x^{<3>}$ $x^{<4>}$ $x^{<5>}$
Jane visite l'Afrique en septembre

   - ☐ **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.**

☑ ~~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 taking the average of those word values to map the Attention related to that given word.**

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

   ☐ *Attention(Q,K,V)=softmax(dkQVT)K*

   ☑ ~~*Attention(Q,K,V)=softmax(dkQKT)V*~~

   ☐ *Attention(Q,K,V)=min(dkQKT)V*

   ☐ *Attention(Q,K,V)=min(dkQVT)K*

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

   Q = interesting questions about the words in a sentence

   K = specific representations of words given a Q

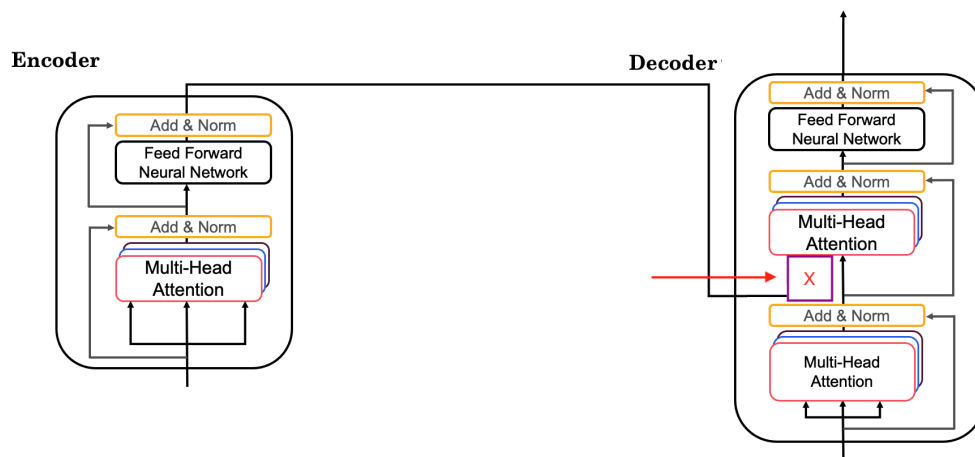   V = qualities of words given a Q

   ☑ ~~False~~

   ☐ **True**

6.

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

   i here represents the computed attention weight matrix associated with the ith "word" in a sentence.

   ☐ **True**

   ☑ ~~False~~

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)
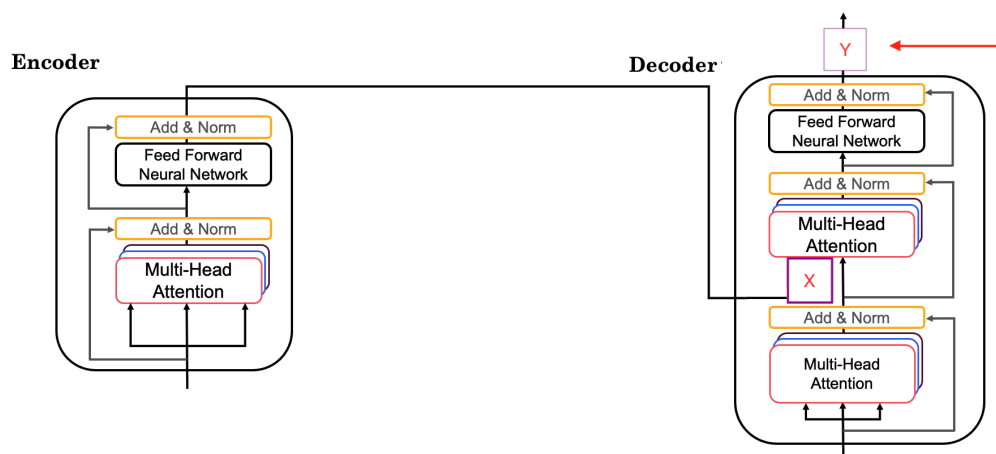
(Check all that apply)

- ☐ **Q**
- ☑ K
- ☑ V

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

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

- ☐ **Softmax layer followed by a linear layer.**

☐ **Softmax layer**

☑ ~~Linear layer followed by a softmax layer.~~

☐ **Linear layer**

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

☑ ~~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 criteria for a good positionial encoding algorithm?

☑ ~~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 the these.**