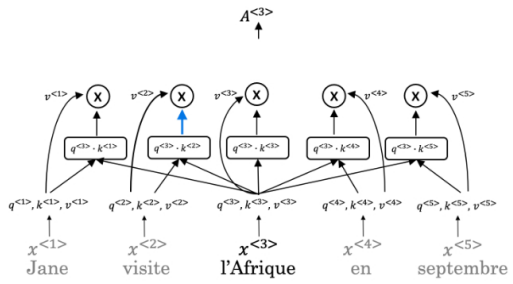
1. A Transformer Network, unlike its predecessors RNNs, GRUs and LSTMs, can process entire sentences all at the same time. (Parallel architecture). 1 point

- ○ False
- ◉ True

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

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

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 quotation, knowledge, 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 query, key, and value.

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

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

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 = specific representations of words given a Q

V = qualities of words given a Q

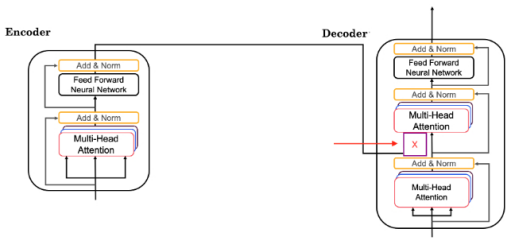
- ◉ False
- ○ True

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

6. 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 the *ith* "word" 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)

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

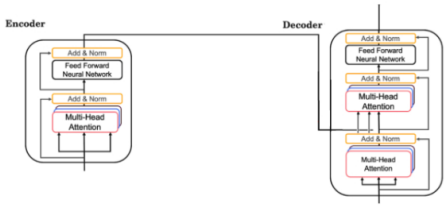


What information does the *Decoder*take from the *Encoder* for its second block of *Multi-HeadAttention* ? (Marked X, pointed by the independent arrow)

(Check all that apply)

- [x] V
- [ ] Q
- [x] 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. Which of the following statements is true about positional encoding? Select all that apply. 1 point

- [x] Positional encoding uses a combination of sine and cosine equations.
- [x] Positional encoding provides extra information to our model.
- [x] Positional encoding is used in the transformer network and the attention model.
- [x] Positional encoding is important because position and word order are essential in sentence construction of any language.

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

- ( ) Distance between any two time-steps should be inconsistent for all sentence lengths.
- ( ) It should output a common encoding for each time-step (word's position in a sentence).
- (•) The algorithm should be able to generalize to longer sentences.
- ( ) It must be nondeterministic.