1. A Transformer Network, like its predecessors RNNs, GRUs and LSTMs, can process information one word at a time. (Sequential architecture). **0 / 1 point**

- ○ False
- ◉ True

Expand

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

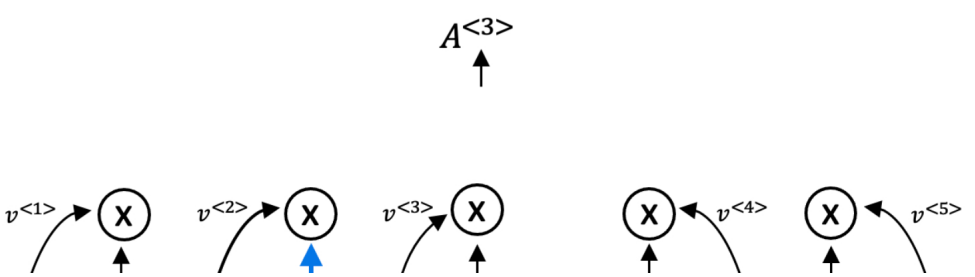
2. Transformer Network methodology is taken from: **1 / 1 point**

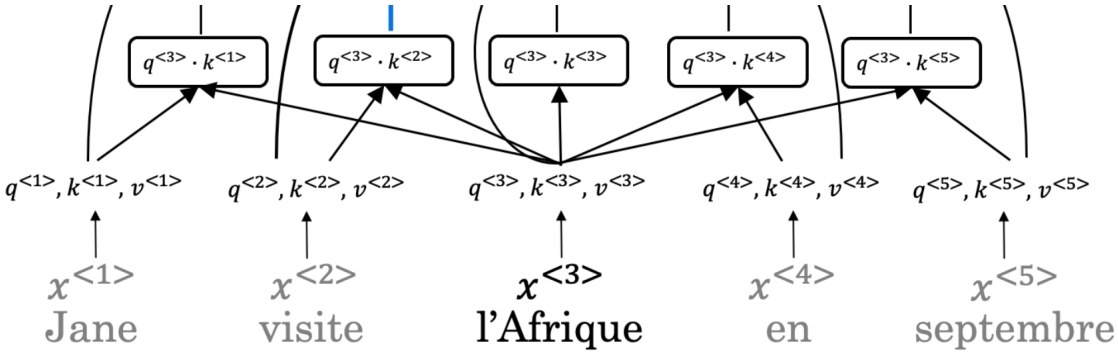
- ○ GRUs and LSTMs
- ◉ Attention Mechanism and CNN style of processing.
- ○ RNN and LSTMs
- ○ Attention Mechanism and RNN style of processing.

Expand

Correct
Transformer architecture combines the use of attention based representations and a CNN convolutional neural network style of processing.

3. The concept of *Self-Attention* is that: **1 / 1 point**





- (●) 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 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 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 taking the average of those word values to map the Attention related to that given word.

Expand

✓ Correct

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

- ( ) $${A(Q,K,V)} = {\sum}_i(\frac{\exp(q * v^{i})}{{\sum}_j\exp(q * v^{j})})* K^{i}$$
- (●) $${A(Q,K,V)} = {\sum}_i(\frac{\exp(q * k^{i})}{{\sum}_j\exp(q * k^{j})})* V^{i}$$
- ( ) $${A(Q,K,V)} = (\frac{\exp(q * k^{i})}{\exp(q * k^{j})})* V^{i}$$
- ( ) $${A(Q,K,V)} = {\sum}_i(\frac{\exp(q * k^{i})}{{\sum}_j\exp(q * k^{j})})* {\sum}_i{v}^{i}$$

Expand

✓ Correct
This is the correct Attention formula.

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

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. $Attention(W_i^Q Q, W_i^K K, W_i^V V)$ 1 / 1 point

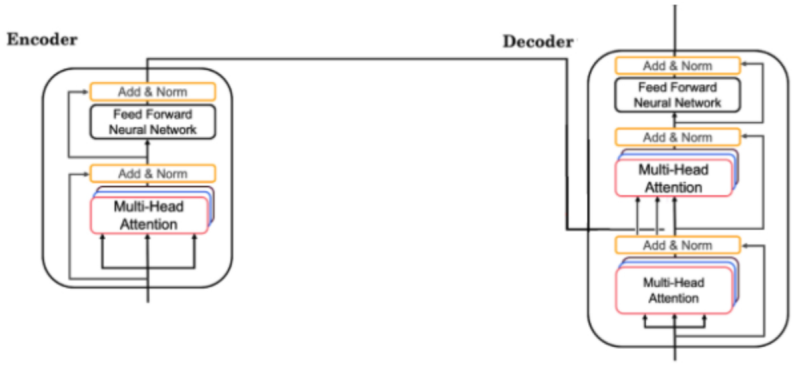
i here represents the computed attention weight matrix associated with the ith "word" in a sentence.

- False
- True

Expand

Correct
Correct! $$i$$ here represents 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 / 1 point



What is ***NOT*** necessary for the *Decoder's* second block of *Multi-Head Attention*?

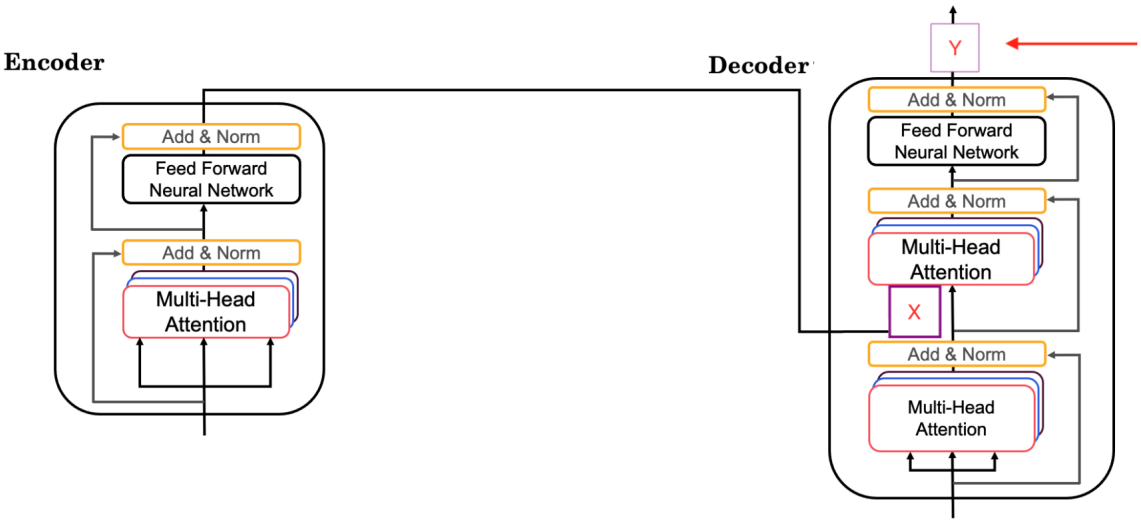
- K
- V
- Q
- All of the above are necessary for the Decoder's second block.

Expand

Correct

The first block's output is used to generate the Q matrix for the next Multi-Head Attention block. The Decoder also uses K and V from the Encoder for its second block of Multi-Head Attention.

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)

- Linear layer
- Softmax layer
- Softmax layer followed by a linear layer.
- Linear layer followed by a softmax layer.

Expand

Correct

9. Which of the following statements is true? 1 / 1 point

- The transformer network differs from the attention model in that only the transformer network contains positional encoding.
- The transformer network is similar to the attention model in that neither contain positional encoding.
- The transformer network is similar to the attention model in that both contain positional encoding.
- The transformer network differs from the attention model in that only the attention model contains positional encoding.

Expand

Correct
Positional encoding allows the transformer network to offer an additional benefit over the attention model.

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

- It must be nondeterministic.
- 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.

Expand

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