Congratulations! You passed!

Grade received 100% **Latest Submission Grade** 100% **To pass** 80% or higher

Go to next item

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

- (•) False
- ( ) True

Expand

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

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

- [ ] Convolutional Neural Network style of architecture.
- [x] Convolutional Neural Network style of processing.

✓ Correct

- [x] Attention mechanism.

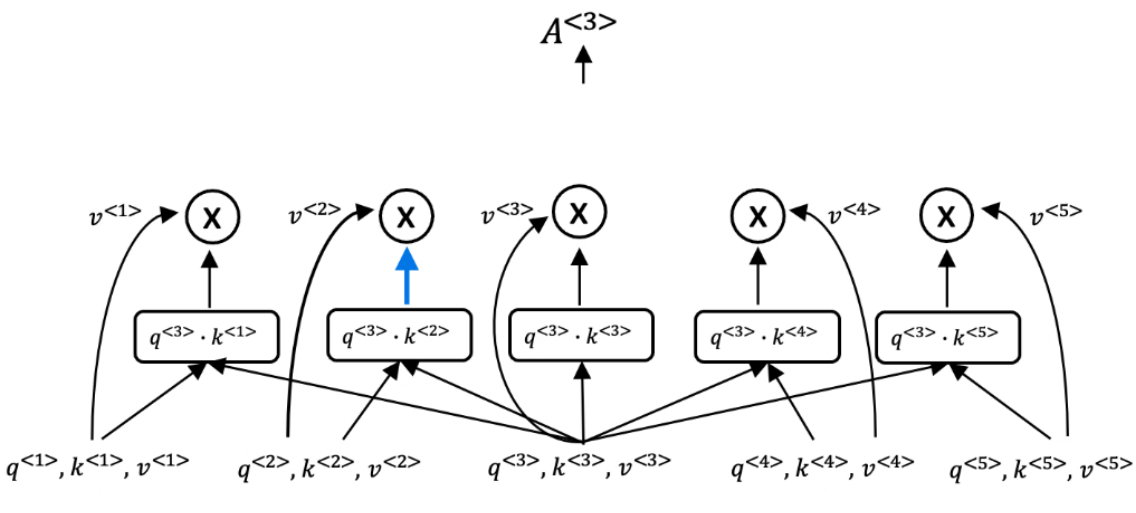
✓ Correct

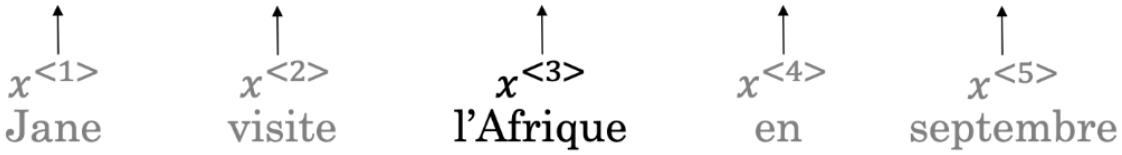
- [ ] None of these.

Expand

Correct
Great, you got all the right answers.

3. How does the Self-Attention mechanism of transformers use neighboring words to compute a word's context? 1 / 1 point





- Selecting the minimum word values to map the Attention related to that given word.
- Summation of the word values to map the Attention related to that given word.
- Selecting the maximum word values to map the Attention related to that given word.
- Multiplication of the word values to map the Attention related to that given word.

Expand

Correct
Given a word, its neighboring words are used to compute its context by summing up the word values to map the Attention related to that given word.

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

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

Expand

Correct
This is the correct Attention formula.

5. Which of the following statements represents Key (K) as used in the self-attention calculation? **1 / 1 point**

- K = qualities of words given a Q
- K = interesting questions about the words in a sentence
- K = specific representations of words given a Q
- K = the order of the words in a sentence

Expand

Correct
The qualities of words given a Q are represented by Key (K).

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

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

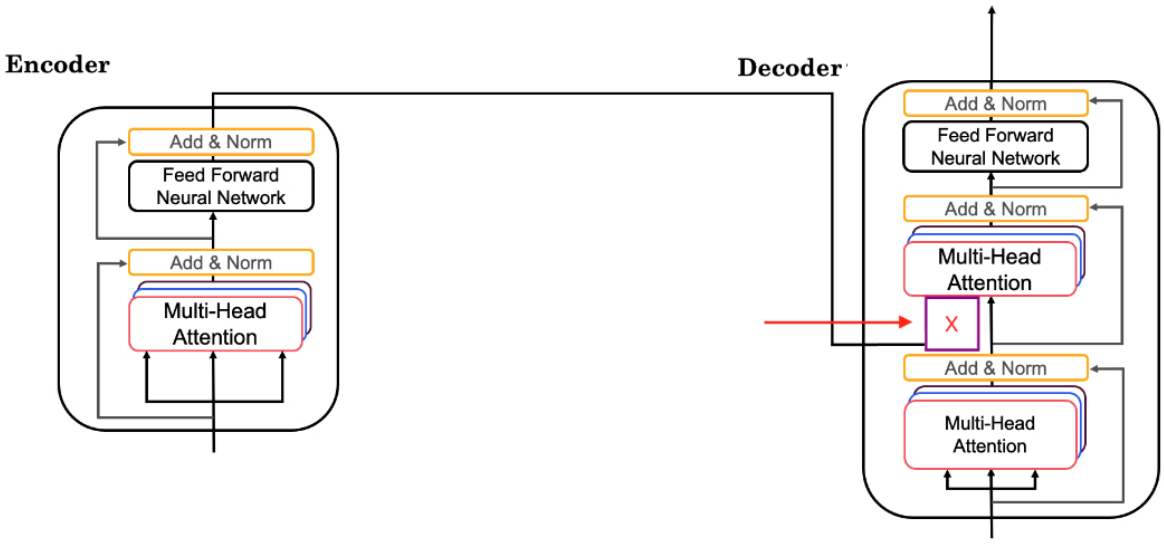
- False
- True

Expand

Correct

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

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



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
- [x] K

Correct

- [x] V

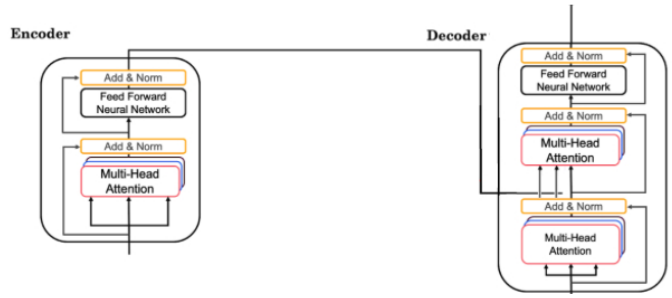
Correct

Expand

Correct

Great, you got all the right answers.

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

- ○ True
- ◉ False

Expand

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

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

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

Expand

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

10. Which of these is ***not*** a good criterion for a good positional encoding algorithm? **1 / 1 point**

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

Expand

✓ **Correct**
This is not a good criterion for a good positional encoding algorithm.