Transformers

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

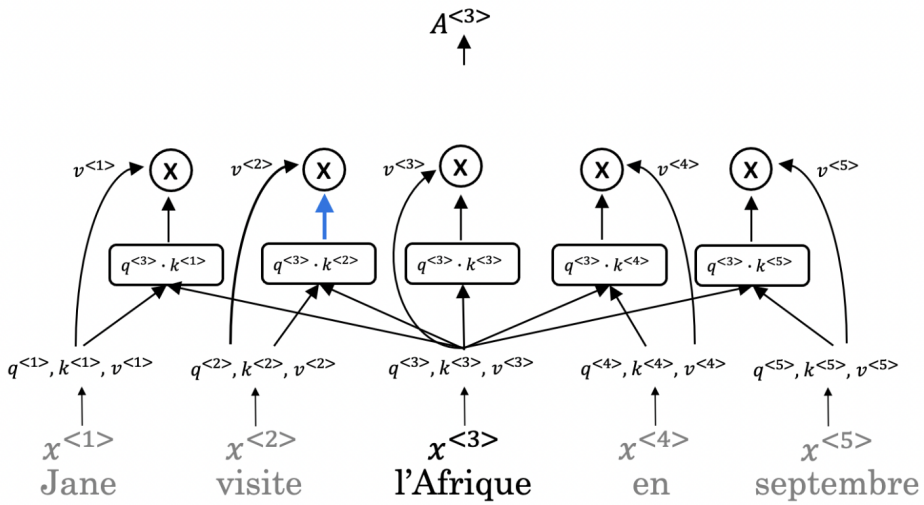
Correct
Correct! A Transformer Network can ingest entire sentences all at the same time.

2. The major innovation of the transformer architecture is combining the use of LSTMs and RNN sequential processing. **1 / 1 point**

- (•) False
- ( ) True

Correct
The major innovation of the transformer architecture is combining the use of attention based representations and a CNN convolutional neural network style of processing.

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



- 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.
- 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.

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. What letter does the "?" represent in the following representation of *Attention*? **1 / 1 point**

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_?}})V$$

- t
- k
- v
- q

Expand

Correct
k is represented by the ? in the representation.

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**

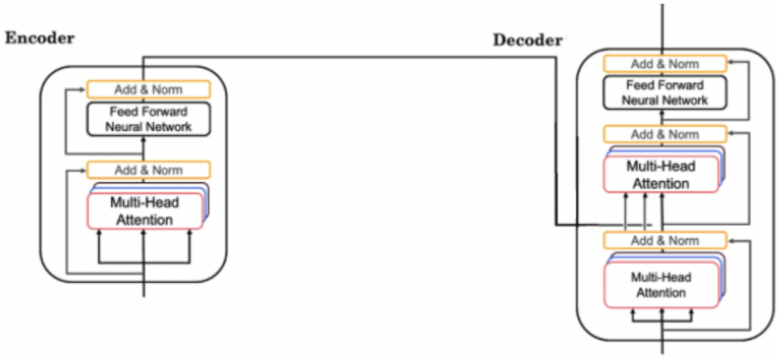
What does i represent in this multi-head attention computation?

- (●) The computed attention weight matrix associated with the *ith* "head" (sequence)
- ( ) The computed attention weight matrix associated with the *ith* "word" in a sentence.
- ( ) The computed attention weight matrix associated with the order of the words in a sentence
- ( ) The computed attention weight matrix associated with specific representations of words given a Q

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**



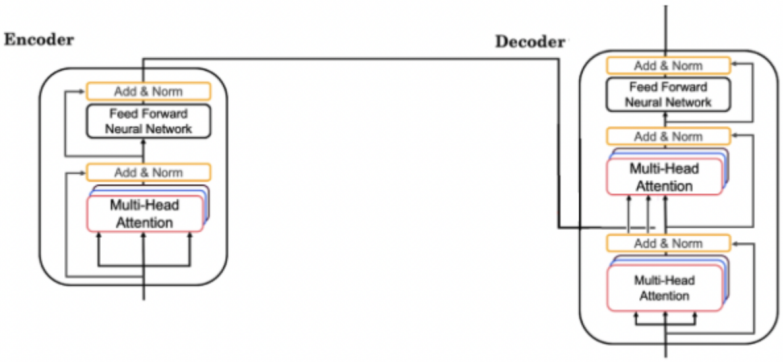
What is generated from the output of the *Decoder's* first block of *Multi-Head Attention*?

- K
- V
- Q

Expand

Correct
This first block's output is used to generate the Q matrix for the next Multi-Head Attention block.

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



What does the output of the *encoder* block contain?

- Softmax layer followed by a linear layer.
- Contextual semantic embedding and positional encoding information
- Linear layer followed by a softmax layer.
- Prediction of the next word.

Expand

Correct
The output of the *encoder* block contains contextual semantic embedding and positional encoding information.

9. Which of the following statements is true about positional encoding? Select all that apply. **1 / 1 point**

- [ ] Positional encoding is used in the transformer network and the attention model.
- [x] Positional encoding uses a combination of sine and cosine equations.

Correct
This is a correct answer, but other options are also correct. To review the concept watch the lecture *Transformer Network*.

- [x] Positional encoding is important because position and word order are essential in sentence construction of any language.

Correct
This is a correct answer, but other options are also correct. To review the concept watch the lecture *Transformer Network*.

- [x] Positional encoding provides extra information to our model.

Correct
This is a correct answer, but other options are also correct. To review the concept watch the lecture *Transformer Network*.

Expand

Correct
Great, you got all the right answers.

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

- (•) The algorithm should be able to generalize to longer sentences.
- ( ) 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).

Expand

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