Congratulations! You passed!

Grade received 90% | Latest Submission Grade 90% | To pass 80% or higher

Go to next item

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

- False
- True

Expand

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] Attention mechanism.

Correct

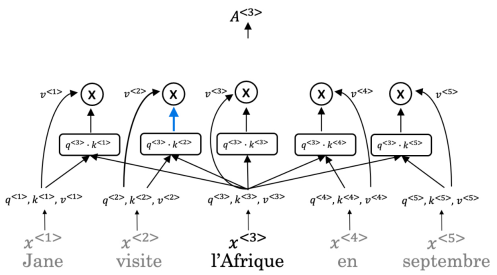
- [ ] None of these.
- [x] Convolutional Neural Network style of processing.

Correct

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 maximum 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.
- 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.

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

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

- k
- q
- v
- t

Expand

Correct

k is represented by the ? in the representation.

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

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

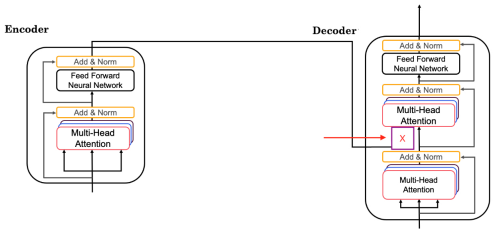
What does i represent in this multi-head attention computation?

- ( ) The computed attention weight matrix associated with the *ith* "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
- (•) The computed attention weight matrix associated with the *ith* "head" (sequence)

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)

- [x] K

Correct

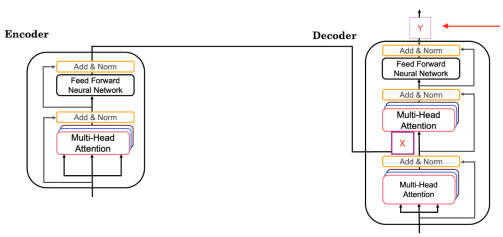
- [ ] Q
- [x] V

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**



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

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

Expand

Correct

9. Which of the following statements is true? **0 / 1 point**

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

Expand

Incorrect
To revise the concept watch the lecture .

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

- [x] It should output a unique encoding for each time-step (word's position in a sentence).

Correct

- [x] Distance between any two time-steps should be consistent for all sentence lengths.

Correct

- [x] The algorithm should be able to generalize to longer sentences.

Correct

- [ ] None of these.

Expand

Correct
Great, you got all the right answers.