Your grade: 100%

Your latest: **100%** • Your highest: **100%** • To pass you need at least 80%. We keep your highest score.

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

- (•) True
- ( ) False

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

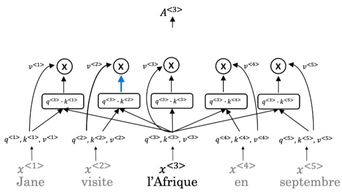
- [ ] None of these.
- [ ] Convolutional Neural Network style of architecture.
- [x] Attention mechanism.

Correct

- [x] Convolutional Neural Network style of processing.

Correct

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



- (•) 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 maximum 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.

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

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

- ( ) q
- ( ) t
- (•) k
- ( ) v

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 = qualities of words given a Q

V = specific representations of words given a Q

- ( ) False
- (•) True

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

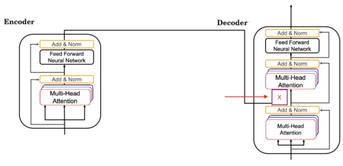
$$\boldsymbol{Attention(W_i^Q Q, W_i^K K, W_i^V V)}$$

6. What does i represent in this multi-head attention computation? 1 / 1 point

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

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 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] K

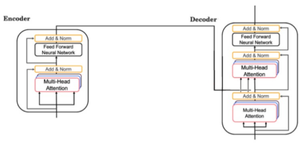
Correct

- [x] V

Correct

- [ ] Q

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?

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

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

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

- ( ) 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.
- ( ) 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.

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.

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