Congratulations! You passed!
TO PASS 80% or higher

Keep Learning

GRADE
100%

Transformers

LATEST SUBMISSION GRADE
100%

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

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**

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

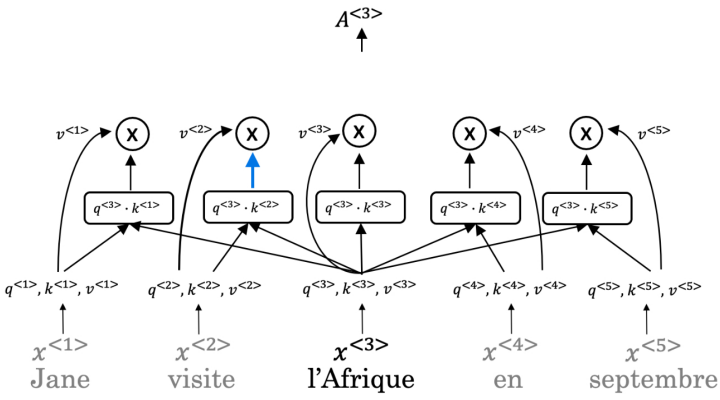
Correct

- [x] Attention mechanism.

Correct

- [ ] Convolutional Neural Network style of architecture.

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



- ( ) 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.
- (•) 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.

Correct

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

- ( ) $Attention(Q, K, V) = min(\frac{QV^T}{\sqrt{d_k}})K$
- (•) $Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$

- $Attention(Q, K, V) = min(\frac{QK^T}{\sqrt{d_k}})V$
- $Attention(Q, K, V) = softmax(\frac{QV^T}{\sqrt{d_k}})K$

Correct

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

- (selected) False
- True

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**

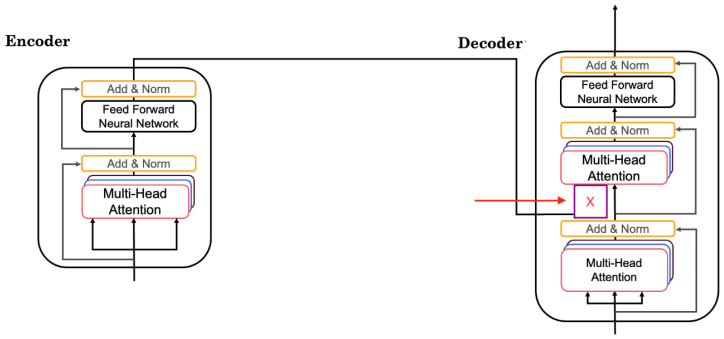
i here represents the computed attention weight matrix associated with the ith "word" in a sentence.

- (selected) False
- True

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 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

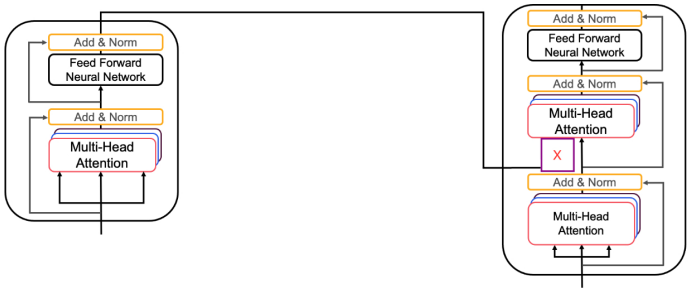
Correct

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

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

✓ **Correct**

9. Why is positional encoding important in the translation process? (Check all that apply) **1 / 1 point**

- ☑ Position and word order are essential in sentence construction of any language.

✓ **Correct**

- ☐ It helps to locate every word within a sentence.
- ☐ It is used in CNN and works well there.
- ☑ Providing extra information to our model.

✓ **Correct**

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

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

✓ **Correct**

- ☑ Distance between any two time-steps should be consistent for all sentence lengths.

✓ **Correct**

- ☑ The algorithm should be able to generalize to longer sentences.

✓ **Correct**

- ☐ None of the these.