RCOS AIHWKIT Status Update #3

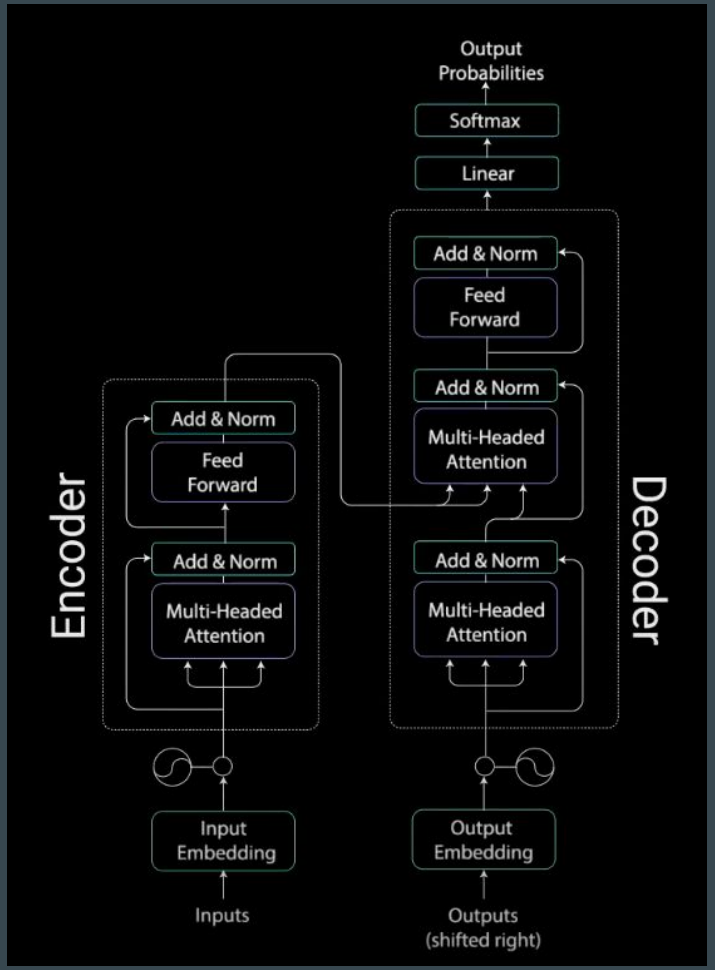
Aasiya Husain

Transformers

1) Input Embedding: Each word is mapped to a vector

2) Positional Encoding: Adds position information into the input vector.
   - For each odd index, add the cos function
   - For each even index, add he sin function

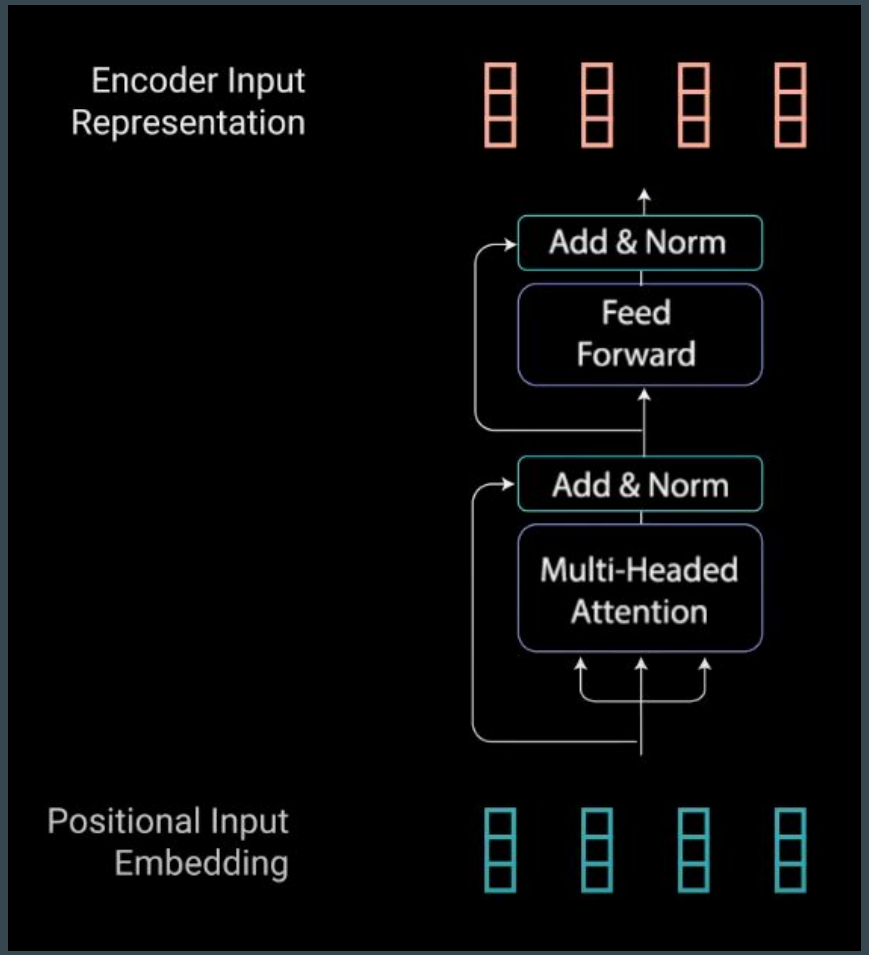
3) Query(Q), Key(K), Value(V) Vectors: These vectors are obtained by passing the input through 3 distinct linear layers.



Encoder

4) MultiHeaded Attention:

- Q, K, V vectors each passed through separate linear layers
- DotProduct(Q, K) -> Scale by square root of the dimension of Q and K -> softmax = Attention weights
- Output = Attention weights x value



Encoder Continued

5) Residual Connection: The output of attention is added to the original position input embedding

6) Normalization Layer

7) Feed-Forward network: =>