Transformers

Monday, March 16, 2020 10:22 PM

→ Get rid of recurrence operation

RNN →

So, how to represent the relation between words?

By using attention mechanism on itself. (self-attention.)

How would transformer know word order?

RNN: → sequential order

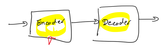
Transformer: Add something, but encode position

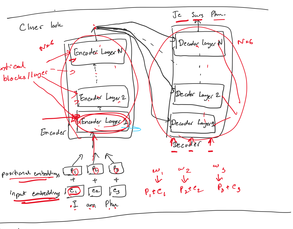
P → position embedding. So number → 1 → 2 → 3

Is there other ways to do that?

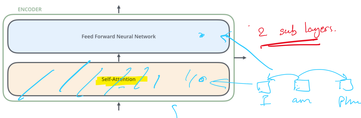
How to encode position?

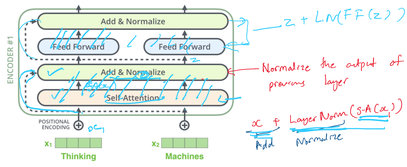
Transformer in NMT



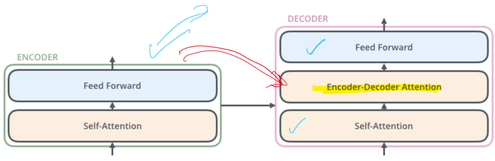


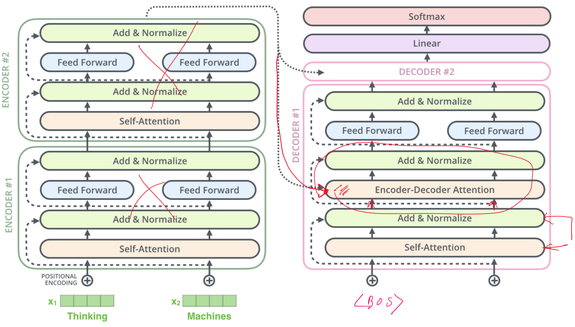
Encoder

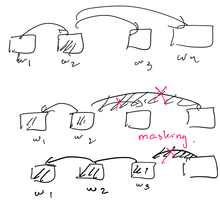




Decoder







How do we compute attention?

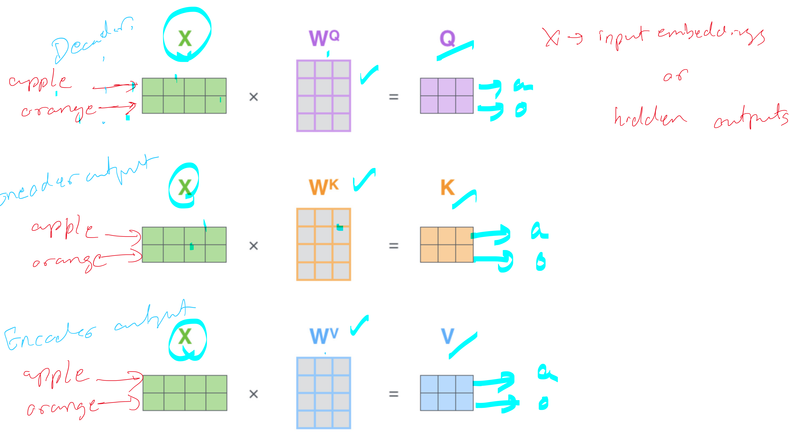
$$\text{Attention}(Q, K, V) = \boxed{\text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)} V$$

$(h_2 \cdot h_2)$ → d_k $\quad (h_1 \cdot h_2)$

① For ~~self-attention~~, query, key, and value are the ~~output of previous layer~~

② For ~~encoder-decoder attention~~, query is ~~previous decoder layer~~, ~~key and value are from the output of encoder.~~

③ Self-attention in decoder do not ~~attend to future positions.~~



Every row in the X matrix corresponds to a word in the input sentence. We again see the difference in size of the embedding vector (512, or 4 boxes in the figure), and the q/k/v vectors (64, or 3 boxes in the figure)