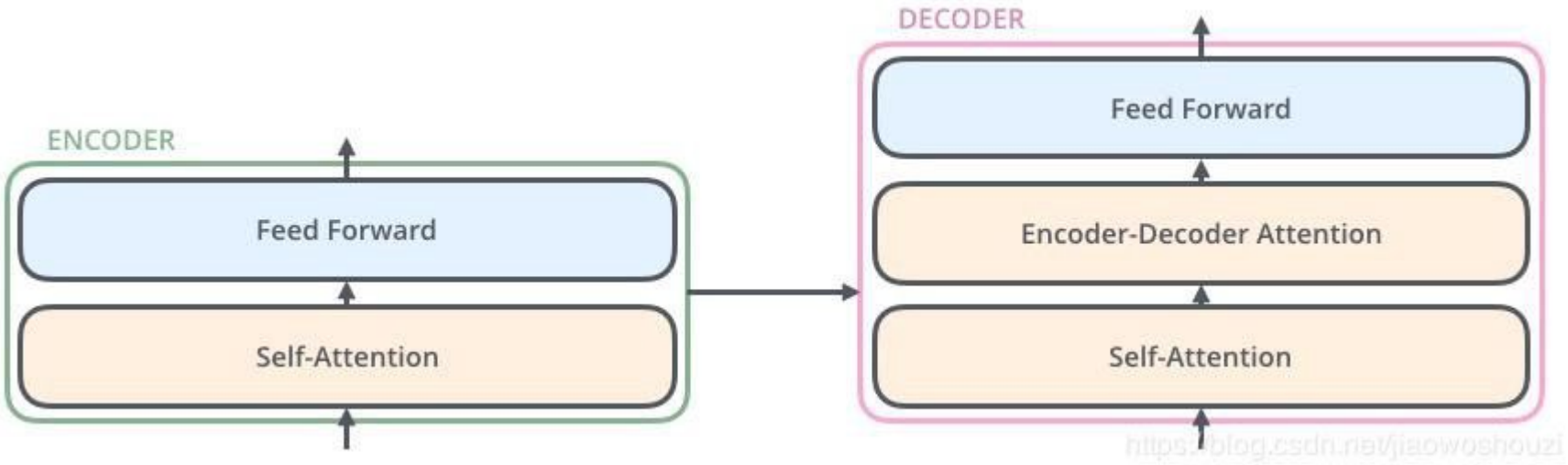
Attention is all you need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin

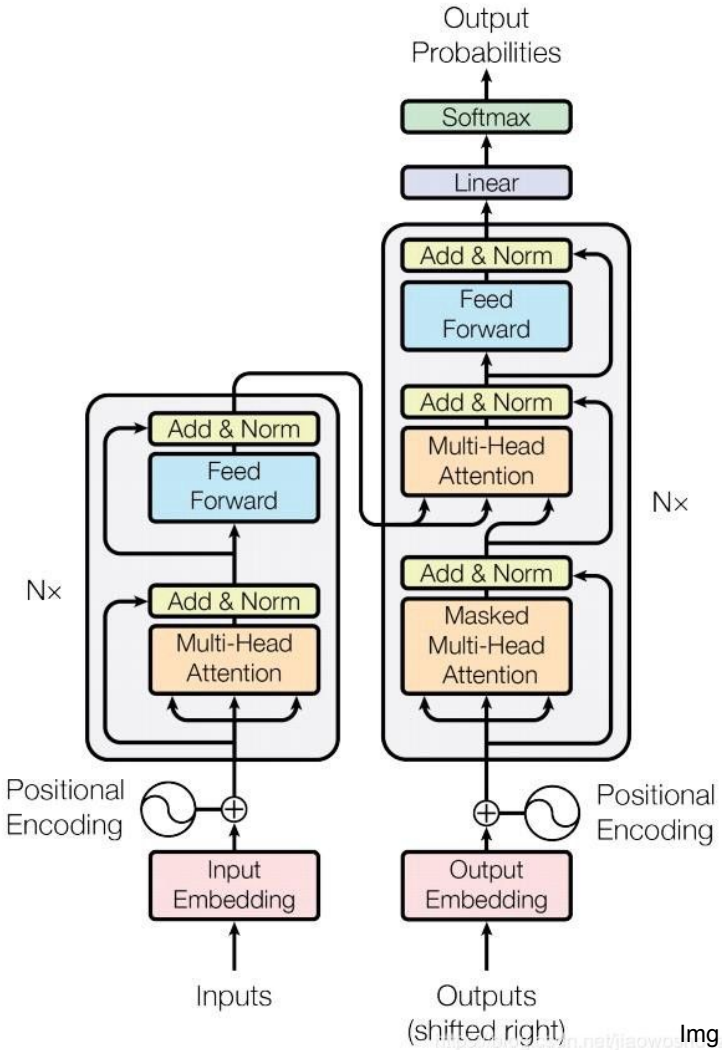
Architecture

Uses encoder-decoder architecture. The encoder layer is stacked by 6 encoders, and the decoder layer is the same.



Img source: Breaking BERT Down

Architecture



Img source: [Breaking BERT Down]

Self attention

- Self-attention is a way for Transformer to convert the “understanding” of other related words into the word we are dealing with.
- Self-attention calculates three new vectors. We call these three vectors Query, Key, and Value respectively.
- We compute the dot products of the query with all keys, divide each by constant and apply a softmax function to obtain the weights on the values.

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T/{d_k}^{\frac{1}{2}})V$$

Multi-headed attention

- Similar to self-attention, where Q, K and V are initialised.
- Instead, multiple groups are initialized, and transformer uses 8 groups, so the final result is 8 matrices.
- The result of these 8 matrices are concatenated and multiplied by another matrix to get final matrix.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(head_1, \ldots, head_h)W^O$$
$$\text{where } head_i = \text{Attention}(QW_i{}^Q, KW_i{}^K, VW_i{}^V)$$

Positional Encoding

- We don’t have a way of interpreting the order of words in an input sequence in the transformer model.
- To handle this problem, the transformer adds an additional vector Positional Encoding to the input of the encoder and decoder layers.
- The value of this Positional Encoding is added to the value of embedding and sent to the next layer as input.

Decoder

- Similar to encoder, but has masked multi-headed attention.
- Some values are masked so as to not affect certain parameters during update.
- Padding mask and sequence mask are the two types.

Padding mask

- If input sequence is small 0 is padded.
- If it is long, the excess the discarded. This is done making the values to negative infinity, so that softmax would give probabilities close to 0.

Sequence mask

- Sequence mask is only used in the decoder’s self-attention.
- A sequence mask is designed to ensure that the decoder is unable to see future information.
- That is, for a sequence, at time_step t, decoded output should only depend on the output before t, not the output after t.

Training

- Trained the models on machine with 8 NVIDIA P100 GPUs.
- For our base models each training step took about 0.4 seconds. So for a total of 100,000 steps or 12 hours.
- For our big models, step time was 1.0 seconds. They were trained for 300,000 steps (3.5 days).

Performance

- On the WMT 2014 English-to-French translation task, big model achieves a BLEU score of 41.0
- Outperforming all of the previously published single models, at less than 1/4 the training cost of the previous state-of-the-art model.