BASMA MOSTAFA

Basma Mostafa

[COMPANY NAME] [Company address]

Attention is all you need

It talks about Transformer, based solely on attention mechanisms

Firstly, we must now what is Transformer:

Transduction is to convert (something, such as energy or a message) into another form essentially sense organs transduce physical energy into a nervous signal

What is Attention?

Attention mechanism was introduced to improve the performance of the encoder-decoder model for machine translation. attention mechanism is divided into the step-by-step computations of the alignment scores, the weights, and the context vector

Work of attention:

- attention function can be described as three vectors from each of the encoder's input vectors So for each word, we create a Query vector, a key vector, and a Value vector. The output is computed as a weighted sum of the values
- Scaled Dot-Product Attention: We compute the dot products of the query with all keys, divide each by √dk, and apply a SoftMax function to obtain the weights on the values.

$$\text{Attention (Q, K, V)} = \text{SoftMax } (\frac{\text{QK}^T}{\sqrt{dK}}) \text{ V}$$

- Multi-head attention: project the queries, keys and values h times with different, learned linear projections to dk, dk and dv dimensions

$$\text{MultiHead (Q, K, V)} = \text{Concat (head1, ..., headh)} W^0$$

where headi = Attention (QWQ, KW K, V WV)

Benefits of Transformer, based on attention mechanisms:

1. superior in quality
2. parallelizable
3. less time to train

Model Architecture:

The encoding component is a stack of encoders (six). The decoding component is a stack of decoders of the same number.

1. **Encoder:** the encoder maps an input sequence of symbol representations (x1, ..., xn) to a sequence of continuous representations z = (z1, ..., zn)
   The encoder is composed of a stack of N = 6 identical layers. Each layer has two sub-layers (a multi-head self-attention mechanism, fully connected feed-forward network)
   The encoder's inputs first flow through a self-attention layer – a layer that helps the encoder look at other words in the input sentence as it encodes a specific word. The outputs of the self-attention layer are fed to a feed-forward neural network.

2. **Decoder**: Given z from the encoder, the decoder then generates an output sequence (y1, ..., ym) of symbols
   The decoder is also composed of a stack of N = 6 identical layers. In addition to the two sub-layers in each encoder layer, the decoder inserts a third sub-layer that helps the decoder focus on relevant parts of the input sentence

   - fully connected feed-forward network is applied to each position separately and identically. This consists of two linear transformations with a ReLU activation in between.

Word Embedding: we use learned embeddings to convert the input tokens and output tokens to vectors