Bothayna Issam Hammad
202010812

Attention is all you need – Summary

Introduction

The paper generally talks about a model architecture called transformer which utilize self-attention mechanism to process input sequences in **parallel**.

Transformer Model Architecture:

1. Encoder and Decoder Stacks
   The transformer encoder comprises six identical layers, each containing a multi-head self-attention sub-layer and a position-wise fully connected feed-forward network. The decoder, also with six identical layers, includes three sub-layers: multi-head self-attention, position-wise feed-forward, and multi-head attention over the encoder's output, facilitating effective sequence processing and generation.
2. Attention
   An attention function maps a query and key-value pairs to an output, computing a weighted sum of values based on compatibility between the query and keys. The output represents the selective attention to different values determined by their alignment with the query.
3. Position-wise Feed-Forward Networks
4. Embeddings and Softmax
5. Positional Encoding

The paper also talks about why we should normalize the model .

The attention block is followed by norm layer that is good for the model to reduce training time and prevent weight explosion.

Normalization techniques:

- Layer normalization
- Batch normalization

The main difference between theses two methods is the way we calculate average and variance.

Conclusion

In this paper, the transformer architecture was introduced. It is the first sequence model based completely on attention, with multi-headed self-attention replacing the recurrent layers.

The transformer achieved new state-of-the-art on English to French and English to German translation.