Summary of "Attention Is All You Need"

Isaac Akintaro

Abstract

This paper represents a major shift in the network architectures used to predict when given machine learning examples. It presents the the Transformer architecture, which uses attention mechanisms to make such a prediction. In this paper the focus is on language tasks. This new architecture takes less time to train because it is more parallelizable, taking advantage of GPUs.

1 Introduction

- The Transformer architecture outperforms recurrent neural networks (RNNs), long short-term memory (LSTMs) and gated recurrent neural networks (GRUs) which were the state of the art in prediction when it came to language modelling and machine translation.

- The innovative approach to deliver this state of the art architecture was the sole use of the attention mechanism.

2 Understanding the Attention Mechanism

- The attention mechanism in simple terms is the process by which weights are derived between what you are trying to predict (query, Q) and your given input.

- Our input can be broken down to a key-value pair. Why do we do this? The Key, K is a way to reference, the content held in Value, Key helps to figure out which parts of input data are important, whilst Value holds the actual information in the input data.

- The paper presents the scaled dot-product attention mechanism which is a faster mechanism than previous versions due to

- This is done first using the dot product which is a measure of how vectors align.

- Then we divide by a scaling factor just before using the softmax function to avoid vanishing gradients (this is when are gradients are close to zero, we use gradients to updates our weights in back propagation, so if zero, it means we will not have any weight updates).

- And then applying the softmax function turns our previous output to weights or probability distributions or importance of what input we should pay attention to for predicting the next output word.
- We then apply the dot product with respect to values, V. This allows us to distribute our attention across the different values.

3 Multi-Head Attention

- A multi-head attention layer consists of attention heads.
- An attention head is a single instance of an attention mechanism with its own set of weights/parameters.
- Each head has its own set of weights, we can think of this as it attends to different parts of the input sequence. These heads can learn specialise in learning different patterns which aggregated help us to make better predictions.
- Each head learns these different patterns, due to random initialization of weights.
- The vectors from each head created are then passed on to other layers in the architecture after the addition and normalization of the attention heads together.

4 Conclusion

- This work is foundational in the the creation of the Large-Language Models we now have, as well impact in other areas such as localisation of objects in computer vision and focusing on specific nodes in graphical neural networks and many other areas.