**Demystifying Self Attention in the Transformer Neural Network Architecture**

# Introduction

This is the third blog in my ongoing series related to the Transformer Neural Network Architecture – the purpose of the whole series is to demystify the different units/components of the Transformer Neural Network Architecture in order to develop a deeper and intuitive understanding of the overall working of a language model. The first 2 articles corresponding to the series provided a deeper understanding of the Foundational principles of Deep Learning [] and the evolution of language models starting from the N-gram to Recurrent Neural Networks to Long Short-Term Memory Units and then into Transformers/BERT/GPT[]. The article on the evolution of language models also went into the finer details of the Attention mechanism highlighting about the Attention Scores, Self-Attention.

The purpose of this article is to cover – or uncover – more details of Self Attention, Multi Headed as well as code the mathematics in a Colab Notebook! The article is organized as follows:

* Firstly, we revisit the motivation behind Attention – laying emphasis on the concept of “Attention Scores” which very intuitively illustrate the related words in a sentence thus capturing the contextual / semantic meaning of the attention.
* In the next section of the article, we discuss an overview of the Transformer architecture at a very high level.
* And finally, the most highlight of this article – we go into the Colab notebook wherein we code the mathematics involved in Self Attention and Multi-headed attention.

# Motivation behind Attention

Recurrent Neural Networks – as discussed in my article on the evolution of language models [] were the state-of-the-art for problems involving sequent to sequence modelling. They have been used in many applications involving sequence modelling. Some examples of the problems involving sequence data where RNNs have been used include the following:

1. Text can be split into sequence of characters or words and each of the individual character or word can be thought of as a time step or sequence.
2. Audio like wave forms from a speech can be split into sequence of sound waves. RNNs have been used for problems involving speech recognition.