# Attention is all you need

Shubham Gupta

January 28, 2020

## 1 Introduction

- This paper review is following the blog from Jay Alammar's blog on the **Illustrated Transformer**. The blog can be found here.

## 2 Paper Introduction

- New architecture based solely on attention mechanisms called **Transformer**. Gets rids of recurrent and convolution networks completely.
- Generally, RNN used to seq-to-seq tasks such as translation, language modelling, etc.
- Transformer allows for significant parallelization and relies only on attention.

## 3 Background

- *Self attention* Attention to different positions of a sequence in order to compute a representation of the sequence.

## 4 Model Architecture

- Transformer uses the following:
  - Encoder decode mechanism
  - Stacked self attention
  - Point wise fully connected layer for encoder and decoder

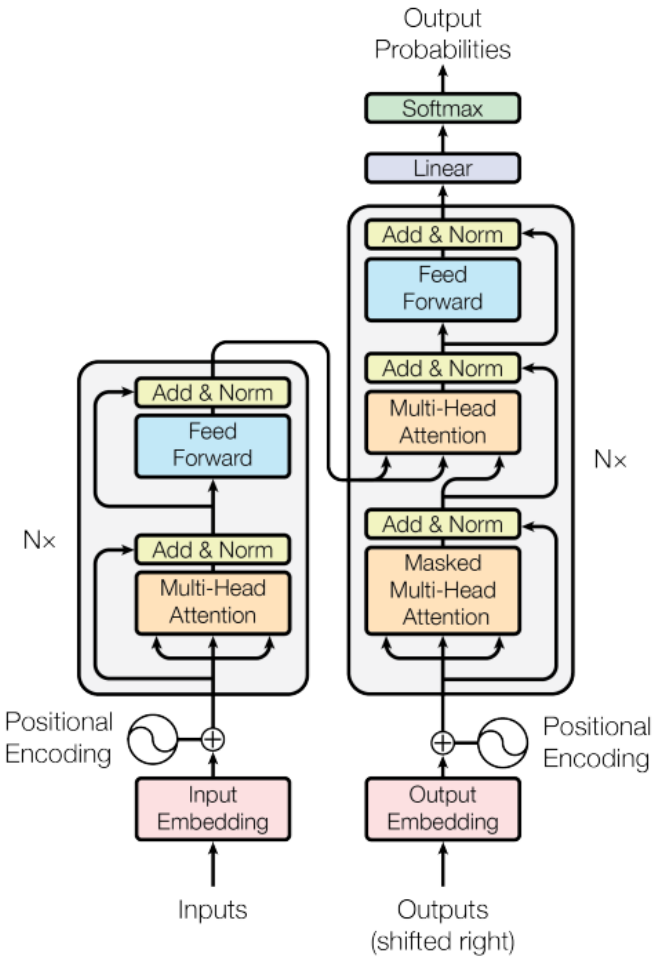


Figure 1: The Transformer - model architecture.

Figure 1: transformer

### 4.1 Encoder and decoder stacks

- **Encoder**: 6 identical layers. 2 sub layers per layer
- *First*: multi-head self attention mechanism
- *Second*: Fully connected feed forward network
- Apply residual connection for each of the two laters
- Apply layer normalization
- **Decoder**: 6 identical layers. 2 sub layers as above + 1 more which performs multi-head attention over output of encoder stack
- Resodual locks around all 3 sub layers

- Layer normalization
- Modify self-attention sub layer to prevent positions from attending to subsequent positions. Ensures that $i$ output depends only on words before $i$.

## 4.2 Attention

- 3 vectors: Query(Q), Key(K) and Value(V)
- Output = Weighted sum of values. Weights assigned as a function of query with key.
- Scaled dot-product attention and multi-head attention