Attention is All You Need

URL: https://arxiv.org/pdf/1706.03762.pdf

Vivek Krishnamurthy
UID: 005294257

Introduction

Most Sequence transduction models in today's world are based on Recurrent Neural Networks and/or Convolutional Neural Networks. The best performing models also use some type of encoder decoder mechanism in addition to Attention. In this paper, the authors have shown that the sequential nature can be deduced with just the Attention mechanism and that Recurrences and Convolutions are no longer necessary. Attention can be thought of as the importance of certain words or combinations of words with respect to the words of the sequence.

At the heart of the paper is the Transformer model. The transformer model consists of an encoder and decoder module which given an input sequence, encodes it to a context aware representation, which is then passed to the decoder. The decoder uses this encoding to create the output sequence.

Transformer Model

Positional Encoding

Figure 1 shows the transformer model. Let us assume that our sequence X is a concatenation of n tokens x_1, x_2, ….. x_n . As input to our model we will pass an embeddings sequence which contains word embeddings for each of these token, or maybe we might even pass a one hot encoding representation. Observe that in the encoding decoding architecture in Figure [1], unlike RNN's, there is no mechanism to take into account the sequential position of the tokens in the sequence. Hence the first operation performed in this transformer is a positional encoding. This is done so that the

model now has information regarding the relative positions of the tokens in the sequence. The positional encoding operation is done before the encoder and decoder.

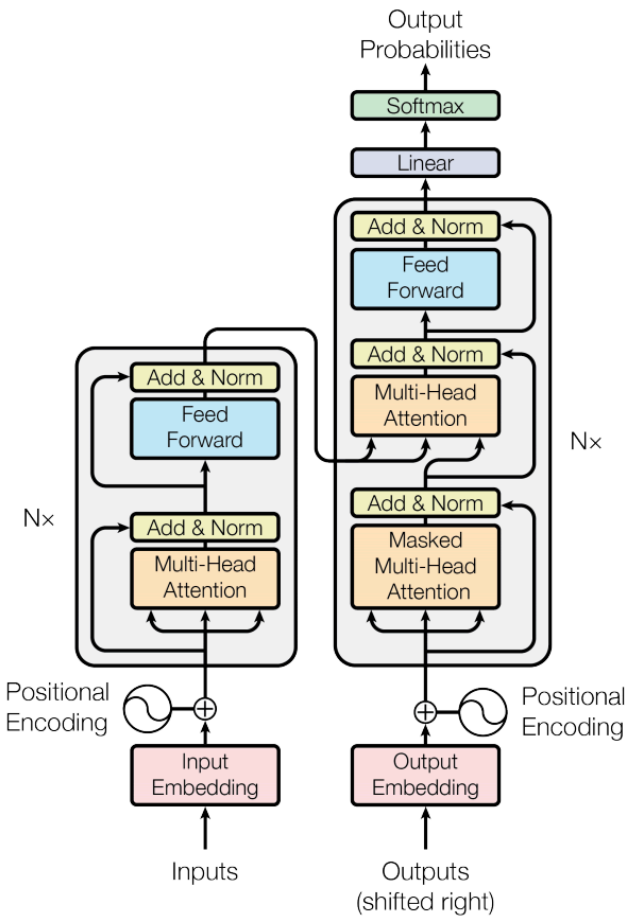

Figure 1: Transformer Model

Self Attention

The attention function can be thought of as mapping a query and set of key-value pairs to an output. The query, key, value and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key.

To Attention is calculated using a formula called the Scaled Dot Product Attention. Given query matrix Q, Key matrix K and Value matrix V, we calculate the Scaled Dot Product Attention as:

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

Here d_k is the dimension of Q and K.
The main aim in using attention is so that each word in the sentence can quantify how much every other word in the sentence values it. We attempt to relate different positions of a single sequence so that we can then calculate a representation of that relationship. For every token x, we create the above Q, K, V matrices. If x_1 wants to know how much it is valued by x_2, it queries the appropriate matrix.

Multi Head Attention

The above section describes one Self Attention unit. The authors of the paper have decided to use multiple self attention units. This is formally known as the Multi Head Attention. Each of these different self attention units will learn different projection matrices which can be used on the incoming sequence and then concatenated together. We then take this concatenated matrix and multiply it with another matrix whose values we initialize and learn iteratively during the training procedure using back propagation.

Layer Normalization

The output of the Multi Head Attention unit is then Layer Normalized. Layer Normalization normalizes the input across all features. Statistics are calculated for every feature and are independent of other examples.

Neural Network

Both the encoder and decoder components have a Neural Network in them. The architecture is fairly straightforward. They consist of two linear layers with a ReLU activation function between them. The input is a set of embeddings x_1', x_2' …., x_n' and the output is another set of embeddings x_1", x_2" …., x_n". Both input and output are of the same dimensions.

Encoder Decoder Stack

The above describes the basic components and functionality of the encoder decoder unit. In the paper however, the authors have stacked up multiple encoders. An output of one encoder serves as the input of the next one and so on. They have also stacked up multiple decoders. The paper uses 6 encoders and 6 decoders. The Figure 2 below describes the stack. The encoder does not function in a sequential manner. The decoder however functions sequentially. It outputs one word at a time until it is finished.

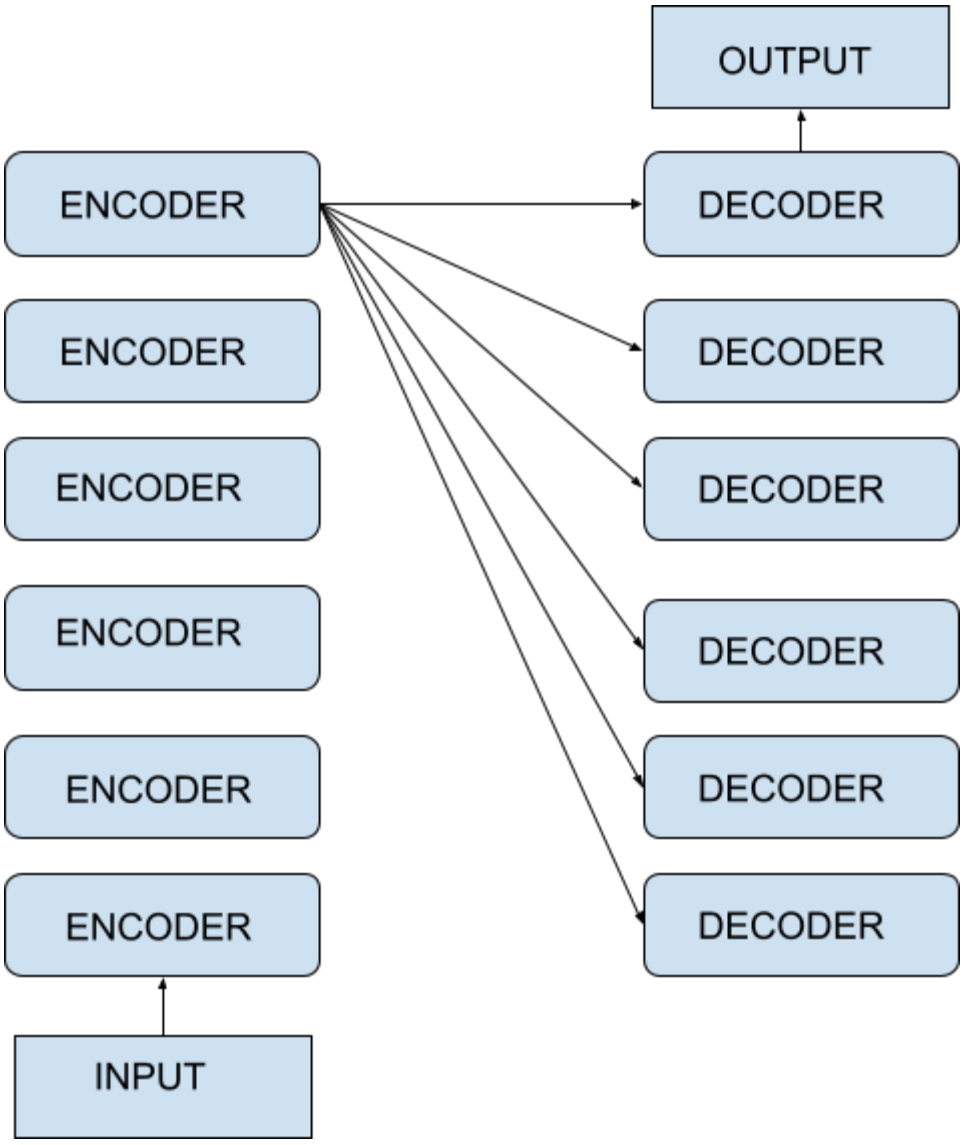

Figure 2: Encoder Decoder Stack

Results

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

Figure 3: Results [1]

The above figure shows us the comparative performance of the Transformer model vs other state of the art techniques. As we can see, the BLEU score of the transformer is higher than the other models. We can also see that the training cost in the case of the transformer model is a fraction of the cost.

Related Work

Image Transformer
https://arxiv.org/pdf/1802.05751.pdf

The image transformer paper is the natural extension to the “Attention is All You Need” paper. The original paper dealt with textual data using the self attention mechanism. This paper talks about using self attention for image generation. The attention mechanism is different from the more commonly used Convolutional Neural Network (CNN) in that it has larger receptive fields per layer, while applying the self attention to local neighborhoods thus allowing it to process larger images. The authors also perform an additional experiment where they use the Transformer model to increase the

resolution of blurred images and find that it is able to outperform current benchmark algorithms.

Structured Attention Networks
https://arxiv.org/pdf/1702.00887.pdf

The main goal of this paper is to model richer structural dependencies within Deep Neural Networks. The authors achieve this by encoding the structural distribution using graphical models.The standard attention model can be viewed as the expectation of an annotation function over a single latent variable. In this paper, a graphical model with multiple latent variables is specified, whose edges encode the desired structure of the dependencies. This mechanism is a natural extension of the basic attention procedure. The authors focus on two types of structured attention problems: Linear chain Conditional Random Fields and first order graph based dependency parsers.

A Structured Self-Attentive Sentence Embedding
https://arxiv.org/pdf/1703.03130.pdf

This paper proposes a new model for extracting sentence embeddings by using self attention. The authors propose the self attention model as a replacement for the max pooling or average pooling step. The mechanism enables the extraction of different aspects of the sentence into multiple vector representations. These representations are then used to perform a wide variety of text/language tasks. The authors evaluate their model on three different tasks: author profiling, sentiment classification and textual entailment. The results show that this proposed model yields better results in all three tasks compared to other sentence embedding methods. This is similar to the “Attention is All You Need Paper” in that both the papers try to create a representation of the input sequence using an attention mechanism.

Self-Attention Generative Adversarial Networks
https://arxiv.org/pdf/1805.08318.pdf

This paper discusses the usage of self attention by Generative Adversarial Networks during image generation. The key improvement of this paper over regular GAN’s is that details can be generated using features from non local/farther away locations in the image as opposed to the earlier case, which involved convolutional GANs, where only

spatially local points were used. In the case of convolutional GAN's the filter sizes could be increased to allow for better representation, but this would result in computational inefficiencies. The Self Attention model is better suited for long range dependencies and is more computationally efficient. The proposed model performs better than previous models on the ImageNet dataset.

State-of-the-Art Speech Recognition with Sequence-to-Sequence Models
https://ieeexplore.ieee.org/abstract/document/8462105

This paper describes the use of the attention mechanism within the domain of speech recognition. The authors build upon an existing single head attention model, by adding multi head attention, which is shown to improve performance. By using the multi head architecture as opposed to the single head architecture, each head can play a different role while attending to the encoder output. The authors hypothesize that this makes it easier for the decoder to retrieve information from the encoder. They also hypothesize that the Multi headed model will perform better in noisy environments. Both this paper and "Attention is All You Need" use multi headed architectures which allow for better representation. The authors test their results on 12 500 hour voice search tasks and dictation tasks and find that their model performs better than the conventional model.