Applications of AI

Attention Is All you Need

Google Paper Summary

Student : Yazan Nabulsi Instructor : Dr. Abdulkareem Al-Banna

ID:201910022

* **Abstract** :

The dominant sequence transmission models are complex repeats or Convolutional neural network with encoder and decoder. The best Performance models also connect encoders and decoders through attention mechanism. We propose a new simple network architecture, Transformer. Based solely on attention mechanisms and avoids repetitions and twists completion. Experiments on two machine translation tasks show that these models work Better quality, at the same time more parallelizable and much more demanding Less time to train.

* **Introduction :**

Since then, efforts have continued to push the boundaries of recurrent language models and encoders/decoders. The order of entries is random. Jacob suggested and started replacing RNN with self-attention. Efforts to evaluate this idea. Ashish, along with Illia, designed and implemented the first Transformer model. It has helped me in every aspect of this job. Norm proposed multi-head scaled dot product attention attention and parameter-free position representations make almost everyone another participant detail. Niki has designed, implemented, optimized, and evaluated countless variants of her model in the original codebase. tensor2tensor. Llion also experimented with new model variants and integrated them with our initial codebase.

**Recurrent models** typically factor computation along the symbol positions of the input and output sequences. Aligning the positions to steps in computation time.

* **Background** :

The goal of reducing sequential computation also forms the foundation of the Extended Neural GPU [20], ByteNet [15] and ConvS2S [8], all of which use convolutional neural networks as basic building block, computing hidden representations in parallel for all input and output positions.

* **Model Architecture** :

Most of the competing **neural array transformation models have an encoder-decoder structure** [5, 2, 29]. Here, the encoder maps the input sequence of symbol representations (x1, ..., xn) into a sequence. Continuous representation z = (z1, ..., zn). Given z, the decoder produces an output. A sequence of symbols (y1, ..., ym), one element each. At each step the model is autoregressive [9]. Previously generated symbols are used as additional input when generating the next symbol. Transformer follows this overall architecture, building upon self-attention and integrity at each point. Connected layers of both encoder and decoder shown in the left and right halves of Figure 1 , One by one.

* The **Decoder consists** of a stack of N = 6 identical layers. In addition to those two**,** SublayersIn each encoder layer, the decoder inserts his third sublayer**,** which performs a multihead. Notice the output of the encoder stack. Similar to **e**ncoders**,** it uses residual connections Layer normalization continues around each sublayer**.** Change your sense of self Lower layers in the decoder stack so that positions do not take into account subsequent positions. of By combining the masking and the fact that the output embedding is offset by his one position, Predictions at location i can only rely on known outputs at locations smaller than i.
* **Applications of attention in our model :**

The Transformer employs three distinct methods of multi-head attention: In "encoder-decoder attention" layers, the encoder's output provides the memory keys and values, while the queries originate from the previous decoder layer. This enables each and every place to cover every position in the input sequence in the decoder. This resembles the typical sequence-to-sequence models' encoder-decoder attention mechanisms include [2, 9, 38]. There are layers of self-attention in the encoder. Every key, value, and layer in a self-attention layer and queries originate from the same location, which in this instance is the output of the layer before it in the encoder. Every position in the encoder has the ability to service every position in the layer above it encoder .In a similar vein, the decoder's self-attention layers enable every position.

* **Position-wise Feed-Forward Networks** In addition to attention sub-layers, each of the layers in our encoder and decoder contains a fully connected feed-forward network, which is applied to each position separately and identically. This consists of two linear transformations with a ReLU activation in between.
* **Positional Encoding** Since our model contains no recurrence and no convolution, in order for the model to make use of the order of the sequence, we must inject some
* **Conclusion :**

In this work, we presented the Transformer, the first sequence transduction model based solely on attention, which uses multi-headed self-attention in place of the recurrent layers that are typically found in encoder-decoder architectures.

Compared to architectures based on recurrent or convolutional layers, the Transformer can be trained much more quickly for translation tasks. We reach a new state of the art on the WMT 2014 English-to-German and WMT 2014 English-to-French translation tasks. Our best model beats all previously published ensembles in the former task .