Paper Review 7

Attention is All You Need

Authors: Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

Introduction

In this work, the authors propose the Transformer architecture as a replacement to recurrent neural networks, long short-term memory (LSTM) and gated recurrent networks for the purpose of sequential modeling tasks like language modeling and machine translation. The recurrent networks model the input as a sequence of hidden states, with each state depending on the previous hidden states, and the corresponding input for that state. However, this leads to a constraint over memory consumption due to the lack in parallelization. The attention mechanisms address this problem by drawing global dependencies between the input and the output.

A transformer architecture relies on the attention mechanism for this purpose, but significantly improves the parallelization. In the transformer architecture, the number of operations required to relate signals between two arbitrary input or output positions is constant. The reduction in effective resolution as a result of this is handled by Multi-Head attention, which maps a query and a set of key-value pairs to an output.

Model Architecture

Transformers have an encoder-decoder structure, wherein, the encoder maps an input sequence of symbol representations to a sequence of continuous representations, and the decoder generated the output sequence of symbols one element at a time. The encoder has multiple layers, each of which has two layers, the first containing the multi-head attention mechanism, and the second, a position-wise fully connected network. The decoder also has multiple layers, with each layer having one additional layer as compared to the encoder, which performs multi-head attention mechanism on the output of the encoder stack.

In this architecture, the authors use the scaled dot product attention mechanism, which computes the softmax of the dot product of the query with all the keys, the output of which is scaled by a factor of $\sqrt{d_k}$. While attention can be computed by the addition mechanism and the dot product mechanism, the latter is preferred because it can be parallelized and made computationally faster by using highly optimized matrix multiplication techniques. Multi-head attention linearly projects queries, keys and values to d_k, d_k, d_v dimensions respectively. Thus, due to multi-head attention, the model learns information from different representation subspaces.

The multi-head attention can be implemented in the transformers in the following ways: where the queries come from the previous decoder layer and the key-value pairs from the output of encoder; where the encoder contains a self-attention layer in which all three, queries, keys and values come from this layer; where self-attention layers in decoders allow each position to attend to all positions in the decoder up to and including that position. In order to make use of the order of the sequence, the authors inject information about the relative or absolute position of the tokens, the authors use the input embeddings at the bottoms of encoder and decoder stacks.

Experimentation & Results

The authors test this model architecture on two datasets: WMT 2014 English-German dataset consisting on 4.5 million sentence pairs for the English-German translation task, and the WMT 2014 English-French dataset consisting of 36 million sentence pairs for the English-French translation task. A training batch size of roughly 25,000 was used, containing the source and target sentence pairs. The former were trained on a total of 100,000 steps and the latter on 300,000 steps, owing to the bigger size of the WMT 2014 English-French dataset.

The Adam optimizer was used for the task, with a learning rate, which was increasing for the initial 4000 training steps, and then decreasing proportionally to the inverse square root of the number of steps. In total, ”Residual Dropout” and ”Label Smoothing” were employed for the purpose of regularization. The former applies dropout to the output of each sub-layer before it is added to the sub-layer input. In the latter, a label smoothing value of 0.1 is employed which hurts the perplexity as the model learns to be more unsure, but improves the accuracy. In the WMT English-German translation task, the Transformer outperforms the previous best performance by 2.0 BLEU points, with a much lesser training cost. In the WMT English-French translation task, the Transformer architecture outperforms the previous best performance, and more significantly, at one-fourth the training cost. In Table 2, the authors record the BLEU scores and the training costs for all of the state-of-the-art models.

Furthermore, the authors evaluate the importance of different aspects of the Transformer by tweaking the different components in the architecture and recording the change in performance on the English-German translation task. The authors vary the number of attention heads, and the attention key and value dimensions and record the change in performance. They note that the single-head transformer is 0.9 BLEU points worse than the best setting achieved for this architecture. Also, a reduction in attention size hurts the model quality, thus proving that the dot product is not a very effective function for measuring the compatibility.

Strengths & Weaknesses

This paper introduces the Transformer architecture, which is the first sequence-transduction model based completely on attention. One of the major strengths of this architecture, apart from the superior performance, is the significant reduction in training costs - to the magnitude of one-fourth of the training cost of the next best-performing model for the English-French translation task. One other advantage of this architecture is the possiblity to compute all the output sequences at once, which in the case of sequential models is a distant possibility. One of the major disadvantages of the

Transformer architecture is that attention can only deal with fixed-length text strings, and the text has to be split into a certain number of segments or chunks before being fed into the system as an input, thus causing context fragmentation. Thus, the text is split without respecting the semantic boundary in the sentence.

Further Work

The introduction of the Transformers architecture has led to the self-attention paradigm which led to many other architectures based on Transformers such as BERT [1], which is a breakthrough in the field of machine learning and natural language modeling. Recently, in the Fourth Conference on Machine Translation, Facebook introduced RoBERTa, [2] as an extension to BERT, which has now outperformed the human performance in the fields of machine translation and question answering.

References

[1] Devlin, Jacob, et al. ”Bert: Pre-training of deep bidirectional transformers for language understanding.” arXiv preprint arXiv:1810.04805 (2018)

[2] Liu, Yinhan, et al. ”Roberta: A robustly optimized bert pretraining approach.” arXiv preprint arXiv:1907.11692 (2019).