Review of “Attention Is All You Need”

Riyad Bin Rafiq

1. Paper summary

The paper [1] proposed a new architecture called Transformer to solve transduction problems such as language modeling and machine translation. Previously, recurrent neural networks, long-short-term-memory and gated recurrent neural networks were used for these tasks. But Transformer has replaced the idea of recurrence with self-attention mechanism. Experiments were done on two translation tasks such as English-to-German and English-to-French and the newly proposed architecture outperformed all of the past (best) results.

2. Contribution

2.1 New model architecture

The authors implemented a new model architecture called Transformer. The model avoids using recurrence and relies on an attention mechanism that draws global dependencies between input and output. Moreover, the model performs the training by more parallelization.

2.2 Self-attention

In my opinion, the most significant contribution of the paper is implementing self-attention. Self attention helps to provide where to put attention in a sequence and thus it removes the problem of learning long-range dependencies. The authors also explained three needs of self-attention layers.

2.3 Model variations

Another important aspect of the paper is the authors experimented with different components of the Transformer on the English-to-German dataset. For example, they varied the number of attention-heads and the attention key and value dimensions and so on. These kinds of experiments provide the proper understanding of the model architecture and its components.

3. Critique

3.1 Parameter matrice/weights (W)

In the paper, the authors mentioned the parameter matrices for the keys, queries, values but they didn’t clarify how these weights were initialized. Were those initialized randomly or followed by any distribution rule? In my opinion, this should be clear as these weights are an important part of the model.

3.2 Query, key, value

The origin of the query, key and value was not described. How did the idea of these three parameters come up? Moreover, it’s very hard to interpret how these work in the model architecture. As the research applied the model on two datasets, a couple of examples should be provided to get the intuition of the query, key and value.

3.3 Model variations based on only one dataset

The authors varied the base model with different experiments to understand the components of the Transformer. But they used only one dataset, English-to-German for the evaluation. So a question arises why they didn’t use the other dataset (English-to-French). Is there any particular reason for this? If they experimented with the two datasets, the generalization of the varying matrices would be more robust.

Reference

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017. pp. 5998–6008.