Review of Attention Is All You Need

Summary

The paper introduces a new architecture called "The Transformer", which is based on attention mechanisms. As opposed to traditional state-of-the-art methods like recurrent neural networks and convolutional neural networks that were used for natural language processing, this architecture enables better parallelization while achieving state-of-the-art performance. To demonstrate the efficiency of the Transformer, the authors apply it to various translation tasks (English-to-German, English-to-French) as well as English constituency parsing to evaluate its generalization properties.

Reasons to accept the paper

- The major strength of the paper lies in its approach to handling natural language processing tasks in an innovative way. The results presented prove the effectiveness and innovative approach of handling sequence modeling tasks.
- The authors detail every aspect of their architecture by providing all the parameters needed to reproduce their results. They mention the hidden size between the layers, the number of layers used as well as their hardware configuration and the optimizer's parameters used. The authors are very transparent with their method.
- The explanation of the overall architecture is very clear. Figures show really well the overall architecture as well as explaining every operation done in every layer. It is really easy to understand what is happening to the data in each layer.
- Obtained results are outstanding both in performance and parallelization.

Reasons to reject the paper

- The paper does not present a "Limitations" section. It might be hard to evaluate what are the drawbacks of the authors' method with respect to other methods. For example, RNNs, as opposed to Transformers, handle sequential dependencies in a more natural way making them better for task implying temporal dependencies like speech processing.
- The authors only compare their result based on one metric (BLEU) which might not be enough to represent the overall effectiveness of a method. They could have used the ROUGE metric too to compare their results.

Comments

My recommendation is to accept the paper (Strong accept), because of the novelty and the efficiency of the approach. The results speak for themselves and the overall explanation of the method is very clear.

Major Comments

Section 1 Introduction : The authors mention that the Transformer "can reach a new state of the art in translation quality after being trained for as little as twelve hours on eight P100 GPUs". The number of hours and GPUs might not be important in the introduction and could be mentioned later. It might be useful to instead explain that the Transformer achieves state of the art performance while needing less training than other methods.

Section 3 Model Architecture : The authors could add a little bit more explanation on how a text sequence is processed before going into the architecture (before the input and output embedding). It might be useful to mention if there are begin and end tokens. Is there a way to mask certain tokens ?

Section 3.2.3 Applications of Attention in our Model : This section mentions that the information flow of the decoder has to be leftward and that the authors implemented it using a mask, but it could be useful to add more explanation on how the mask is applied in the layer. It is challenging to understand (a) what positions is the mask applied to, (b) how the mask is computed based on the input and (c) how the mask relates to the fact that the input of decoder is the output shifted right ?

Section 4 Why Self-Attention : Providing a more in-depth explanation of why and how the Transformer architecture is more parallelizable than classical architecture like recurrent neural networks would enhance clarity and benefit the authors. A paragraph could also be added to explain how exactly this property can be utilized effectively on GPUs.

Section 4 Why Self-Attention : It might be clearer to put the theoretical complexities of every model's path length in a table to improve readability

Section 5.1 Training Data and Batching : There could be a table in this section describing the datasets to improve readability

Section 5.4 Regularization : This section mentions that the authors used three types of regularization, but only two methods are discussed. Is there a third method ? It might be more useful to remove this section and transfer the dropout part to the section 3.1 since the dropout is linked to the layer normalisation. The part on label smoothing could be left in that section as it is only related to the training. This would also allow the Table 2 to be next to the results section.

Section 6.2 Model Variations : The Table 3 shows clearly every single parameter, but a graph in this section could show in an easier way what the impacts of each parameters are. For example, a graph of the BLEU score based on the size of the dimensions of the key and value vectors could be useful.

Section 6.3 English Constituency Parsing It might be clearer to put the description of the datasets used in a table.

Limitations : The authors do not discuss the drawbacks of the method based on other methods and what potential drawbacks the Transformer architecture could have in some tasks.

Minor comments

Section 3.3 Position-wise Feed-Forward Networks : The authors could add the acronym "FFN" in parenthese when mentionning a feed-forwawrd network to make it clearer to understand equation 2.

Section 3.5 Positional Encoding : It might be usefull to put the fraction into the form $\frac{a}{b}$ to improve readability

Section 5.4 Regularization : If there are two regularization methods, change it to two instead of three.

References

This review was based on the 7th version of the paper Attention is all you need last revised 2 Aug 2023 :

A. Vaswani *et al.*, "Attention Is All You Need." arXiv, Aug. 01, 2023. Accessed: Mar. 09, 2024. [Online]. Available: http://arxiv.org/abs/1706.03762