Thesis Proposal: Attention Is All You Need

Pascal Sager*, Ricardo Chavarriaga*

*ZHAW Datalab
Winterthur, Switzerland
{sage, chav}@zhaw.ch

***Abstract*—The paper "Attention Is All You Need" [1] introduced a new network architecture called Transfomer, which has the potential to sustainable change the field of deep learning. Although this architecture was originally developed for NLP, it recently finds its way into other fields such as ASR [2] and computer vision [3]. Recent research showed that transformers are Turing-complete [4], meaning that they can learn nearly any sequence-to-sequence function. This promising architecture could also be relevant for my first project thesis in the field of sequence modelling. I see this architecture as one of the most promising developments in the field of deep learning in recent years and thus it should be familiar to any aspiring data scientist.**

***Index Terms*—deep learning, sequence modeling, natural language processing**

I. Content

Before the novel Transformer network, recurrent neural networks such as long short-term memory [5] and gated recurrent neural networks [6] have been established as state-of-the-art approaches in sequence modeling [7], [8]. Many of the best performing models used attention mechanism to model dependencies between inputs and outputs [9], [10]. Vaswani et al. [1] then proposed a new network architecture called Transfomer, which relies entirely on an attention mechanism. This architecture outperformed state-of-the-art sequence models in translation tasks while using a fraction of the training costs.

The Transformer network consists of $N = 6$ stacked encoder and decoder layers. The input sequence is fed into the encoder, which maps them to a sequence of continuous representations z. The decoder then generates an output sequence based on z and the previously generated symbols. The recurrent layers which are commonly used in encoder-decoder architectures were replaced with multi-headed self-attention. This self-attention reduces the computational complexity per layer and increases the amount of computation that can be parallelized and thus leads to a more efficient training. Another advantage is that self-attention layers can more efficient learn the dependencies between input and output sequences than recurrent layers.

II. Motivation

The Transformer network achieved state-of-the-art performance not only in translation tasks but in various fields of deep learning [2]–[4]. Besides the good performance, also the resource-efficient training is highly relevant for the applicability in practice.

My personal impression is that the Transformer networks will replace recurrent neural networks in many application. Therefore, I think this paper is very influential and should be known to anyone working in the field of deep learning. The transformer network could also be used in my first project thesis, which deals with sequence models in the area of ASR. With this paper, I hope not only to have an exciting seminar literature but also to expand my technical knowledge. I want to understand how Transformer networks work and to be able to assess in which applications they could be promising.

III. Expectations & Goals

In this AI seminar, I would like to improve my scientific communication skills. This includes consuming and providing information. Consuming information consists of different subtasks such as searching, evaluating and analysing scientific texts. Providing information involves oral and written communication in a way that provides a benefit to the audience. After this seminar, I hope to be well prepared for communication tasks in upcoming projects theses, in the Master thesis as well as in my further career. By studying this paper, I hope to understand how Transformers work and to become familiar with the latest advances of this architecture.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[2] H. Miao, G. Cheng, C. Gao, P. Zhang, and Y. Yan, "Transformer-based online ctc/attention end-to-end speech recognition architecture," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6084–6088.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020.

[4] J. Pérez, J. Marinković, and P. Barceló, "On the turing completeness of modern neural network architectures," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=HyGBdo0qFm

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[6] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[7] R. Józefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," *ArXiv*, vol. abs/1602.02410, 2016.

[8] M.-T. Luong, H. Pham, and C. Manning, "Effective approaches to attention-based neural machine translation," 08 2015.

[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2015.

[10] Y. Kim, C. Denton, L. Hoang, and A. M. Rush, "Structured attention networks," *ArXiv*, vol. abs/1702.00887, 2017.