Rethinking Attention: Exploring Shallow Feed-Forward Neural Networks as an Alternative to Attention Layers in Transformers (Student Abstract)

Danilo Dordevic *, Vukasin Bozic *, Joseph Thommes, Daniele Coppola, Sidak Pal Singh

ETH Zurich

{ddordevic, vbozic, jthommes, dcoppola}@student.ethz.ch, sidak.singh@inf.ethz.ch

Abstract

This work presents an analysis of the effectiveness of using standard shallow feed-forward networks to mimic the behavior of the attention mechanism in the original Transformer model, a state-of-the-art architecture for sequence-to-sequence tasks. We substitute key elements of the attention mechanism in the Transformer with simple feed-forward networks, trained using the original components via knowledge distillation. Our experiments, conducted on the IWSLT2017 dataset, reveal the capacity of these "attentionless Transformers" to rival the performance of the original architecture. Through rigorous ablation studies, and experimenting with various replacement network types and sizes, we offer insights that support the viability of our approach. This not only sheds light on the adaptability of shallow feed-forward networks in emulating attention mechanisms but also underscores their potential to streamline complex architectures for sequence-to-sequence tasks.

Introduction

The seminal paper (Vaswani et al. 2017) which introduced the Transformer model has fundamentally altered the landscape of sequence-to-sequence modeling tasks. It set new benchmarks for language translation, measured by the BLEU score (Papineni et al. 2002). The Transformer's attention mechanism enables the establishment of long-term dependencies in sequential data, allowing it to attend to every element in a sequence, a feat prior network architectures struggled to achieve without significant computational overheads.

Inspired by prior work (Ba and Caruana 2014), (Urban et al. 2017) which explore the feasibility of training shallow feed-forward networks to emulate the behavior of deep convolutional networks with deep networks as teachers, we conduct a similar investigation on the original Transformer presented in (Vaswani et al. 2017). Our focus is on language translation, utilizing the IWSLT2017 dataset (Cettolo et al. 2017). We aim to assess the extent to which standard shallow feed-forward networks can model attention mechanisms by substituting key attention components with feed-forward networks trained to replicate their behavior.

This work provides empirical evidence supporting the notion that shallow feed-forward networks can effectively learn the behaviors of Transformer attention modules and replace them without significantly impacting its overall performance. While it does not introduce a competitive advantage over established methods, it offers a conceptual analysis of existing techniques and potential alternatives.

Models and Methods

The Transformer architecture is composed of stacked encoder and decoder blocks, which use attention to process input data. The encoder layer features one self-attention block, while the decoder layer encompasses both self-attention and cross-attention blocks, fusing the data processed by the encoder and itself. This model was used as the baseline, i.e. the teacher model, where the intermediate activations of its blocks were used for knowledge distillation (Hinton, Vinyals, and Dean 2015) in the training of the feed-forward networks.

Encoder self-attention replacement. In the proposed approach, a thorough ablation study of the potential replacement methods was conducted. The experiments were done on self-attention layers in all 6 encoder blocks.

We introduced four different levels of abstraction for replacing the original encoder attention:
Attention Layer Replacement (**ALR**), Attention Layer with Residual Connection Replacement (**ALRR**), Attention Separate Heads Layer Replacement (**ASLR**), and Encoder Layer Replacement (**ELR**), as depicted in Figure 1. Furthermore, all of these architectures were trained in 5 different sizes, ranging from "XS" to "L".

Full Transformer attention replacement. As ALR was found to be the most effective approach in the case of encoder attention replacement, featuring both high performance and a small number of parameters, the whole procedure was recreated for decoder self-attention and cross-attention replacement. This required adaptations of the previously introduced architectures, caused by different types of attention in the decoder.

*These authors contributed equally.

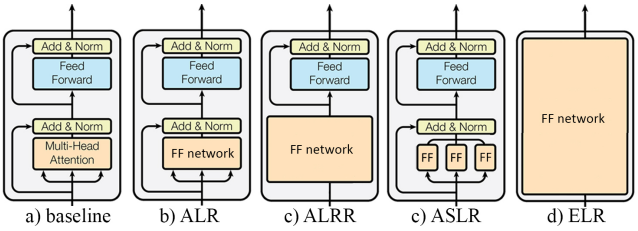

Figure 1: Different encoder self-attention replacement approaches presented.

Results

BLEU metric was used for evaluation purposes in this work, as it represents a standard metric for language translation tasks. The results for both encoder self-attention and full Transformer replacement studies span on 4 subsets of the IWSLT2017 dataset. Furthermore, BLEU scores relative to the baseline (vanilla Transformer) score of every experiment were calculated and then averaged over the datasets. The most important experimental results are presented in Figures 2 and 3. We provide the implementation code on Github[1].

Discussion

In the case of encoder replacement, all of the proposed methods achieve competitive results compared to the baseline, as seen in Figure 2. Out of the four approaches, ELR performs the worst, which is caused by the simplicity of the replacement model, which discards all of the encoder structures that aid training.

Furthermore, the full Transformer replacement approach, where only the ALR method is utilized, yielded results showcasing the potential of the feed-forward networks to successfully replicate the decoder self-attention behavior, while the performance on decoder cross-attention is comparatively worse, as presented in Figure 3. The potential reason for this behavior could be the lack of the expressiveness of the feed-forward network needed to describe the more complex mapping and interaction between sequences used in the cross-attention block, which also influences final evaluation scores for the fully "attentionless" Transformer.

However, all of the replacement approaches come at a significant cost of having more parameters. Another downside of our replacement of the attention with a fixed-size feed-forward network is the imminent lack of flexibility of the model in terms of the length of sequences the model can operate with.

Conclusion

Empirical evidence suggests that the proposed approaches are capable of achieving comparable performance to that of the original Transformer, demonstrating that Transformers do not necessarily need to have attention. These conclusions also point out the deficiencies of the current optimization methods, which are not able to train these "attentionless Transformers" from scratch but need more advanced techniques, such as knowledge distillation to converge into desired parameter configurations. This conclusion emphasizes that with the advancements in optimization techniques, less specialized architectures such as feed-forward networks could be used for advanced tasks, currently reserved for highly specialized architectures.

[1]https://github.com/vulus98/Rethinking-attention.git

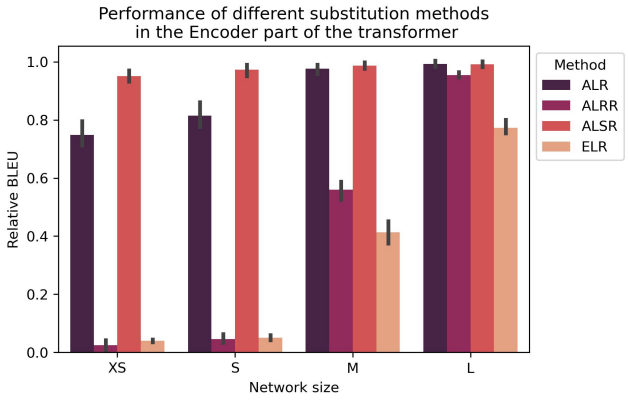

Figure 2: Relative BLEU scores [%] (relative to the baseline Transformer), depending on the FF network size. Encoder self-attention is replaced using different replacement methods.

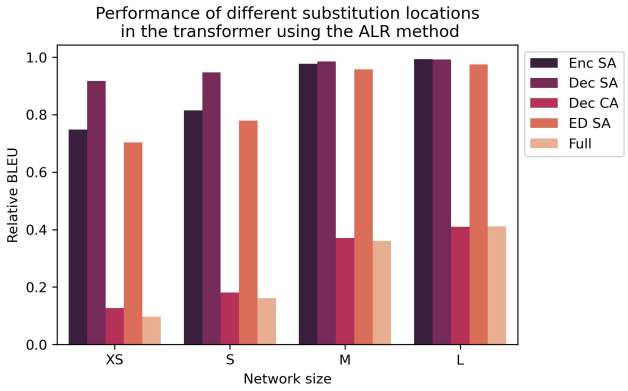

Figure 3: Relative BLEU scores [%] (relative to the baseline), depending on the FF network size. ALR method is used to replace different attention parts of the transformer.

Future Work

By matching the performance of the original Transformer, it is highly probable that the further optimization of the FF networks' hyperparameters using advanced parameter search (e.g. using Bayesian optimization (Snoek, Larochelle, and Adams 2012)) could yield even better results in terms of translation quality and possibly even enable the usage of smaller FF networks for the replacement, as the size of the networks represents one of the major bottlenecks for the deployment of these 'attentionless' Transformers in practice.

Acknowledgements

We would like to express our sincere gratitude to the Data Analytics lab of ETH Zurich for providing the necessary resources and support during the course of this project.

References

Ba, L. J.; and Caruana, R. 2014. Do Deep Nets Really Need to be Deep? ArXiv:1312.6184 [cs].

Cettolo, M.; Federico, M.; Bentivogli, L.; Niehues, J.; Stüker, S.; Sudoh, K.; Yoshino, K.; and Federmann, C. 2017. Overview of the IWSLT 2017 Evaluation Campaign. In Sakti, S.; and Utiyama, M., eds., *Proceedings of the 14th International Conference on Spoken Language Translation*, 2–14. Tokyo, Japan: International Workshop on Spoken Language Translation.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. ArXiv:1503.02531 [cs, stat].

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P.; Charniak, E.; and Lin, D., eds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Snoek, J.; Larochelle, H.; and Adams, R. P. 2012. Practical Bayesian Optimization of Machine Learning Algorithms. arXiv:1206.2944.

Urban, G.; Geras, K. J.; Kahou, S. E.; Aslan, O.; Wang, S.; Caruana, R.; Mohamed, A.; Philipose, M.; and Richardson, M. 2017. Do Deep Convolutional Nets Really Need to be Deep and Convolutional? ArXiv:1603.05691 [cs, stat].

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. arXiv:1706.03762.