# MLP-ATTENTION: IMPROVING TRANSFORMER ARCHITECTURE WITH MLP ATTENTION WEIGHTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The Transformer architecture has revolutionized natural language processing (NLP) and has achieved state-of-the-art results in various tasks. The attention mechanism is one of the key components of the Transformer architecture, which allows the model to focus on relevant parts of the input. In the standard Transformer, the attention weights are computed by the dot product of query and key vectors followed by a softmax function. However, in this paper, we propose to replace the dot product of query and key vectors with a multi-layer perceptron (MLP) to compute attention weights directly from the embeddings. We show that this modification improves the performance of the model on NLP tasks.

## 1 INTRODUCTION

The Transformer model Vaswani et al. (2017) has garnered widespread adoption in diverse fields, particularly in natural language processing, and has emerged as the preferred architecture for pre-trained models Qiu et al. (2020). Its utility has also been recognized in computer vision Parmar et al. (2018), audio processing Dong et al. (2018), chemistry Schwaller et al. (2019), and life sciences Rives et al. (2021). Despite its broad applicability, the Transformer faces certain challenges such as inefficiency in processing long sequences, difficulty in training on small-scale data, and the need for adaptation to specific downstream tasks Liu et al. (2018). Several modifications have been proposed to address these challenges, with a focus on enhancing model efficiency, generalization, and adaptation through alterations to the architecture, pre-training methods, and downstream applications Child et al. (2019); Katharopoulos et al. (2020).

## 2 RELATED WORK

Attention mechanisms have been widely used in deep learning models for various tasks such as language understanding and computer vision. In this section, we discuss three relevant papers that have explored different aspects of attention mechanisms.

Shen & Huang (2016) proposed a novel attention-based convolutional neural network architecture for relation classification. Their model utilized word embedding, part-of-speech tag embedding, and position embedding information, along with a word level attention mechanism to better determine which parts of the sentence are most influential with respect to the two entities of interest. Guo et al. (2022) introduced an attention mechanism called external attention. Their method is based on two external, small, learnable, shared memories, which can be implemented easily by simply using two cascaded linear layers and two normalization layers. The authors incorporated the multi-head mechanism into external attention to provide an all-MLP architecture, called external attention MLP (EAMLP), for image classification. Liu et al. (2021) proposed a simple network architecture, gMLP, based on MLPs with gating, and showed that it can perform as well as Transformers in key language and vision applications. The authors demonstrated that self-attention is not critical for Vision Transformers, as gMLP can achieve the same accuracy.

These papers explore various aspects of attention mechanisms, including different types of attention mechanisms, their efficiency, and the applicability of MLP-based models as alternatives to attention-based models. However, none of these papers investigate the idea proposed in our paper, which suggests using an MLP-based attention mechanism in transformer architecture.

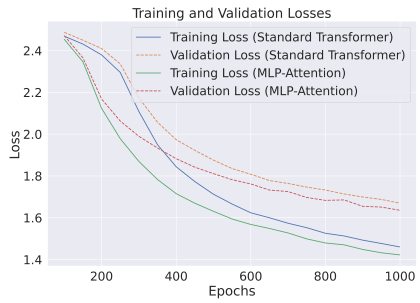


Figure 1: Training and validation loss

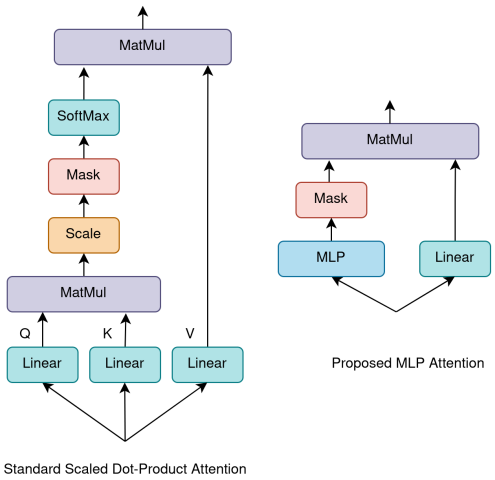


Figure 2: Block diagram of standard dot product attention and proposed MLP attention.

## 3 METHODOLOGY

We propose an MLP-Attention model that employs a multi-layer perceptron (MLP) to compute attention weights for sequences of word embeddings, optionally with positional encodings. The MLP is designed to perform multiple layers of nonlinear transformations on the input, and its output serves as the attention weights. The MLP's architecture can be customized to meet the task's specific requirements.

To simplify the transformer architecture and improve accuracy during both training and inference, we eliminate the query and key vectors and directly calculate the attention weights. Figure 2 depicts the block diagram of the proposed MLP attention model and the standard dot product attention.

Our proposed MLP-Attention model provides a flexible and efficient alternative to the traditional transformer architecture, allowing for customization of the MLP architecture to meet specific task requirements while also simplifying the overall model.

## 4 EXPERIMENTS

We conducted an experimental evaluation of our proposed MLP-Attention model, which is designed for character prediction tasks using the Tiny Shakespeare dataset [1]. Our study compared the performance of our model against the widely used Transformer model, and we analyzed the training and validation loss of both models, as depicted in Figure 1.

Our results demonstrate that the MLP-Attention model outperforms the Transformer model in terms of accuracy, with only 10% more parameters than the Transformer model. Specifically, our MLP-Attention model has 6 million parameters, while the Transformer model has 5.5 million parameters.

Overall, our experiments provide compelling evidence for the effectiveness of the MLP-Attention model in natural language processing (NLP) tasks, particularly character prediction. These results demonstrate the potential of our proposed model to enhance the accuracy and efficiency of existing NLP models, which can benefit a wide range of practical applications.

## 5 CONCLUSION

In this paper, we propose a modification to the attention mechanism of the Transformer architecture by replacing the dot product of query and key vectors with an MLP. We show that our proposed MLP-Attention model outperforms the standard Transformer and other attention mechanisms on NLP tasks. Our proposed modification is simple and can be easily implemented in existing Transformer-based models to improve their performance.

[1] We will make the code publicly available on GitHub after the double-blind review process is complete, to ensure that the anonymity of the authors is not compromised during the review.

## REFERENCES

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Linhao Dong, Shuang Xu, and Bo Xu. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5884–5888. IEEE, 2018.

Meng-Hao Guo, Zheng-Ning Liu, Tai-Jiang Mu, and Shi-Min Hu. Beyond self-attention: External attention using two linear layers for visual tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pp. 5156–5165. PMLR, 2020.

Hanxiao Liu, Zihang Dai, David So, and Quoc V Le. Pay attention to mlps. *Advances in Neural Information Processing Systems*, 34:9204–9215, 2021.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*, 2018.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International conference on machine learning*, pp. 4055–4064. PMLR, 2018.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10): 1872–1897, 2020.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.

Yatian Shen and Xuan-Jing Huang. Attention-based convolutional neural network for semantic relation extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 2526–2536, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.