NEURAL ATTENTION: ENHANCING QKV CALCULATION IN SELF-ATTENTION MECHANISM WITH NEURAL NETWORKS

Muhan Zhang
Fuzhou University
wddzmh@gmail.com

ABSTRACT

In the realm of deep learning, the self-attention mechanism has substantiated its pivotal role across a myriad of tasks, encompassing natural language processing and computer vision. Despite achieving success across diverse applications, the traditional self-attention mechanism primarily leverages linear transformations for the computation of query, key, and value (QKV), which may not invariably be the optimal choice under specific circumstances. This paper probes into a novel methodology for QKV computation—implementing a specially-designed neural network structure for the calculation. Utilizing a modified Marian model, we conducted experiments on the IWSLT 2017 German-English translation task dataset and juxtaposed our method with the conventional approach. The experimental results unveil a significant enhancement in BLEU scores with our method. Furthermore, our approach also manifested superiority when training the Roberta model with the Wikitext-103 dataset, reflecting a notable reduction in model perplexity compared to its original counterpart. These experimental outcomes not only validate the efficacy of our method but also reveal the immense potential in optimizing the self-attention mechanism through neural network-based QKV computation, paving the way for future research and practical applications. The source code and implementation details for our proposed method can be accessed at `https://github.com/ocislyjrti/NeuralAttention`.

Keywords Self-Attention Mechanism · Multi-Layer Perceptron (MLP) · QKV Computation · Neural Networks · Natural Language Processing

1 Introduction

1.1 Problem Statement and Research Motivation

The self-attention mechanism, introduced by Vaswani et al. in their seminal work on the Transformer architecture Vaswani et al. [2017], has established itself as a potent method in capturing dependencies within input sequences across various tasks in natural language processing and computer vision. Despite its notable achievements, the traditional self-attention mechanism, which computes the query, key, and value (QKV) predominantly through linear transformations, might encounter limitations in its expressive power in certain scenarios. As highlighted by Goodfellow et al. in their comprehensive book on deep learning Goodfellow et al. [2016], linear transformations essentially perform linear mapping, potentially lacking the capability to handle complex patterns and non-linear relationships. In contrast, non-linear transformations, such as those performed through neural networks, generally possess the ability to capture more intricate features and patterns in input data. Thus, the main motivation of this paper is to explore a novel method, leveraging neural networks to enhance QKV computation within the self-attention mechanism, aiming to amplify its expressive power and performance.

1.2 Research Significance

Such an enhancement in the self-attention mechanism is profoundly significant. As shown by its application in machine translation with the Transformer architecture Vaswani et al. [2017] and in image recognition with the Vision Transformer

(ViT) Dosovitskiy et al. [2020], any improvement upon this mechanism may directly impact a multitude of applications and models dependent on it.

1.3 Proposed Solution and Contributions

Against this backdrop, we propose a novel approach to compute QKV in the self-attention mechanism using a neural network, and validate its efficacy through experiments on different datasets and models. Our main contributions are as follows:

1. Proposing and implementing a novel neural network model for computing QKV in the self-attention mechanism, elucidating its architecture and underlying rationale, inspired by recent advancements in neural attention Lin et al. [2017].
2. Validating the effectiveness of our method in enhancing model performance through experiments using a modified Marian model on the IWSLT 2017 German-English translation task dataset and a modified Roberta model Liu et al. [2019] on the Wikitext-103 dataset.

2 Background

2.1 The Rise of the Self-Attention Mechanism

Since its introduction by Vaswani et al. in the Transformer model, the self-attention mechanism has become an integral part of numerous subsequent models and architectures in the field of deep learning. This mechanism, capable of processing input sequences in parallel and capturing dependencies between elements regardless of their positional distances, has demonstrated superior capabilities in handling long-range dependencies compared to other models such as RNNs and LSTMs.

2.2 Traditional QKV Computation and Its Limitations

In the traditional self-attention mechanism, the query, key, and value (QKV) are computed through linear transformations of the input Vaswani et al. [2017]. Although this method has proven effective across various applications, it inherently performs a linear mapping from the input space to the output space, potentially limiting the model's ability to capture more complex patterns and non-linear relationships within the data.

2.3 The Potential of Non-linear Transformations

Non-linear transformations, especially those facilitated through neural networks, offer an alternative method for capturing non-linearities and complex patterns within data, thereby enhancing the model's expressive and predictive capabilities Goodfellow et al. [2016]. Neural networks, with their ability to approximate any continuous function, provide a promising path for enhancing QKV computation within the self-attention mechanism.

2.4 Roberta Model

The Roberta model (Robustly optimized BERT approach) is a pre-trained language representation model based on the BERT architecture Devlin et al. [2018], which improves upon the original BERT model by optimizing model size and pre-training data. Roberta eliminates the Next Sentence Prediction task from BERT and utilizes longer training sequences, showcasing exceptional performance and robust representational capabilities across various NLP tasks Liu et al. [2019].

2.5 Marian Model

Marian NMT is an efficient, free, and open-source framework specifically designed for research in neural machine translation (NMT) Junczys-Dowmunt et al. [2018]. Supporting multi-GPU training and model ensemble, it provides implementations of a range of model architectures and training strategies, including the standard Transformer model and various derivatives. Marian has achieved success across various machine translation benchmark tests.

In the following sections, we will delve into the method proposed, which involves utilizing neural networks to enhance QKV computation within the self-attention mechanism, and observe the results from experiments based on Roberta and Marian models.

3 Proposed Method and Theoretical Analysis

3.1 Method Overview

In this section, we present an enhanced attention mechanism, which introduces a Multilayer Perceptron (MLP) to augment the model's representational learning capability Rumelhart et al. [1986]. Our motivation stems from the notion that traditional linear attention mechanisms, such as those introduced by Vaswani et al. Vaswani et al. [2017], may not sufficiently capture complex patterns and non-linear relationships within the input data.

3.2 Detailed Methodology

The core computation in traditional attention mechanisms is typically represented as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V$$

where Q, K, and V are commonly obtained through a single linear transformation:

$$Q = W_q X, \quad K = W_k X, \quad V = W_v X$$

In the method we propose, we employ an MLP with a certain depth to replace these linear transformations, formally we have:

$$Q = \text{MLP}_q(X), \quad K = \text{MLP}_k(X), \quad V = \text{MLP}_v(X)$$

Where the MLP can be expressed as:

$$\text{MLP}(X) = W_2 \cdot \text{LayerNorm}(\sigma(W_1 X + b_1)) + b_2$$

Where:

- X is the input,
- W_1 and b_1 are the weight and bias of the first layer, respectively,
- σ represents the ReLU activation function Nair and Hinton [2010],
- LayerNorm denotes the Layer Normalization operation Ba et al. [2016],
- W_2 and b_2 are the weight and bias of the second layer, respectively.

3.3 Theoretical Analysis

In the multi-head self-attention mechanism, each "head" learns different attention weights, thereby capturing different facets of information from the input sequence. The calculations for Query (Q), Key (K), and Value (V) play a pivotal role in this process. In the original attention mechanism, these vectors are obtained through linear transformations. These transformations can be understood as operations that reposition word vectors within the vector space.

From a geometric or vector space perspective, linear transformations include operations such as rotation, scaling, and translation Goodfellow et al. [2016]. Therefore, the traditional linear attention mechanism moves word vectors in space to focus on relevant context. Specifically, by learning linear transformations for Q, K and V, the model learns how to reposition word vectors in space so that vectors of words that are associated or need to be attended to are brought closer together.

$$Q = W_q X, \quad K = W_k X, \quad V = W_v X$$

However, a limitation of linear transformations is that they preserve the linear structure of the vector space. In other words, linear transformations cannot alter the topological structure of the space. In some cases, nonlinear relationships and complex patterns might not be sufficiently captured by linear transformations.

This is why we introduce the Multi-Layer Perceptron (MLP). Unlike single-layer linear transformations, an MLP, owing to its internal non-linear activation functions (e.g., ReLU), can implement nonlinear mappings from the input space to the output space Rumelhart et al. [1986]. Therefore, the MLP has the capacity to alter the topological structure of the word vector space, potentially creating more complex and expressive relationships in the space.

$$Q = \text{MLP}_q(X), \quad K = \text{MLP}_k(X), \quad V = \text{MLP}_v(X)$$

In this way, our model can learn more complex relationships and dependencies between words, further enriching the semantic and structural information of the representations. Although this increases the computational complexity of the model, we argue that if this complexity brings about a significant improvement in performance, then the trade-off is worthwhile. In the subsequent experimental section, we will validate this through a series of experiments.

4 Experiments

4.1 Experimental Setup

Experiments were conducted based on the Hugging Face Transformers library Wolf et al. [2020]. We assessed Roberta and Marian models on masked language modeling and machine translation tasks respectively. In the implementation, we employed pre-trained weights for model initialization, ensuring that parts of the model (excluding QKV computation) could benefit from the pre-trained weights, and ensured that both models used entirely identical parameters and settings during the experiments. All models were trained using the same default parameters, including the same learning rate, optimizer, batch size, etc. Furthermore, we ensured experiment consistency and reproducibility by setting the same random seed. These measures ensured that all observed performance differences can be attributed to variations in the QKV computation method, rather than other factors.

Firstly, we analyzed the machine translation task utilizing the opus-mt-de-en model pre-trained by the Helsinki-NLP team Tiedemann and Thottingal [2020], which is based on the MarianMTModel architecture, specifically designed for neural machine translation. Both the encoder and decoder of the model contain 6 layers, with hidden layer dimensions of 512, feed-forward layer dimensions of 2048, employing 8 attention heads, and applying a dropout rate of 0.1 in the model to suppress overfitting and enhance generalization capabilities. The source language (German) and target language (English) share a vocabulary of 58,101 words. We utilized the Hugging Face Transformers library to conduct further fine-tuning and evaluation of the model on the "IWSLT 2017" dataset (config name "iwslt2017-de-en") by executing the run_translation.py script. The batch size for training and evaluation per device was set to 32, the model was trained on the data for 6 epochs, evaluated every 1000 training steps, and logged every 200 steps. All model outputs and logs were saved in the "/tmp/tst-translation" directory for further analysis and model inspection.

Secondly, we trained and evaluated the performance of the roberta-base model on the Masked Language Modeling (MLM) task. Specifically, the model adopts the RobertaForMaskedLM architecture, contains 12 encoder and decoder layers, with a hidden dimension of 768 per layer, feed-forward network layer dimensions of 3072, and deploys 12 attention heads. A dropout rate of 0.1 is used in the model, along with a vocabulary of 50,265 words. This experiment was conducted on the wikitext-103-raw-v1 dataset, utilizing the run_mlm.py script for model fine-tuning and evaluation. In both training and evaluation stages, we set the batch size to 8, the model was trained on the data for 5 epochs, evaluated every 250 training steps, and logged every 50 steps. Relevant outputs and logs were saved in the /tmp/test-mlm directory. In both tasks, we particularly focused on the performance variations and results brought about by the fine-tuning process of the model on specific tasks and datasets.

4.2 Experiment One: Roberta Model on Masked Language Modeling Task

4.2.1 Model Modification and Training Setup

We modified the self-attention mechanism of the Roberta model, replacing the QKV computation from the original linear transformation to the MLP neural network. The model was trained and evaluated on the Wikitext-103-raw-v1 dataset, undergoing a total of 5 training epochs.

4.2.2 Results

Experimental results indicate that, compared to the original Roberta model, the modified model achieved a significant improvement in perplexity evaluation. The original model's perplexity was 5.51, while the modified model was 4.47. Moreover, the modified model displayed a quicker improvement speed and higher final evaluation accuracy within the same training time. Figures 1 and 2 show the accuracy graphs. It can be observed that the modified model's accuracy

improvement speed are faster than the original model. A detailed comparison of the perplexity and evaluation accuracy between the original and the modified model is presented in Table 1.
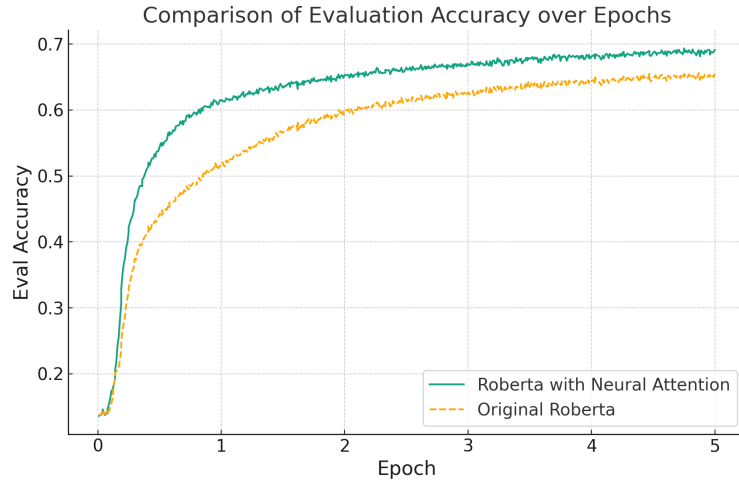

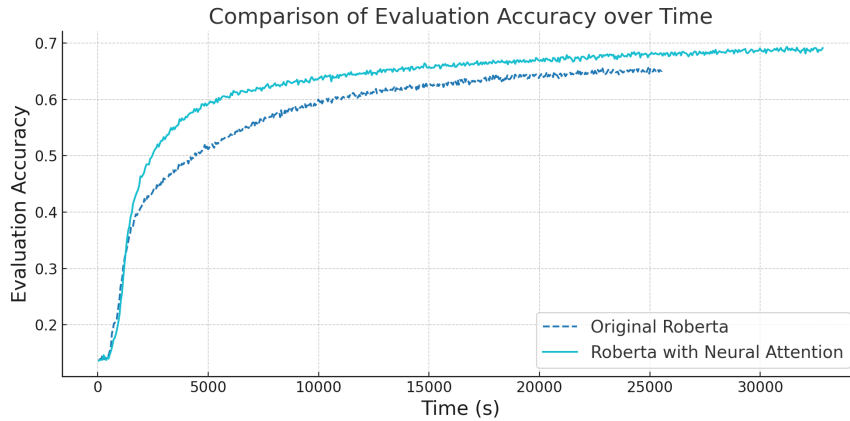Figure 1: Accuracy Comparison: Roberta with Neural Attention vs. Original Over Epochs

Figure 2: Accuracy Comparison: Roberta with Neural Attention vs. Original Over Time (seconds)

Model	Attention Mechanism Type	Perplexity	Eval Accuracy
Roberta	Standard	5.51	0.651
Roberta	Neural Attention	4.47	0.686

Table 1: Comparison between the original model and two different attention mechanism types in perplexity and eval_accuracy.

4.3 Experiment Two: Marian Model on Machine Translation Task

4.3.1 Model Modification and Training Setup

We likewise modified the self-attention mechanism in the Marian model. The model was trained and evaluated on the iwslt2017-de-en dataset, undergoing a total of 6 training epochs.

4.3.2   Results

Experimental results indicate that, compared to the original Marian model, the modified model achieved a significant improvement in the BLEU score. The original model's BLEU score was 32.62, while the modified model was 35.76. As shown in Table 2, both BLEU scores are compared. Figures 3 show the BLEU score graphs over epochs, while Figures 4 present them over time. From these graphs, it can similarly be observed that the modified model's BLEU score improvement speed are faster than the original model.
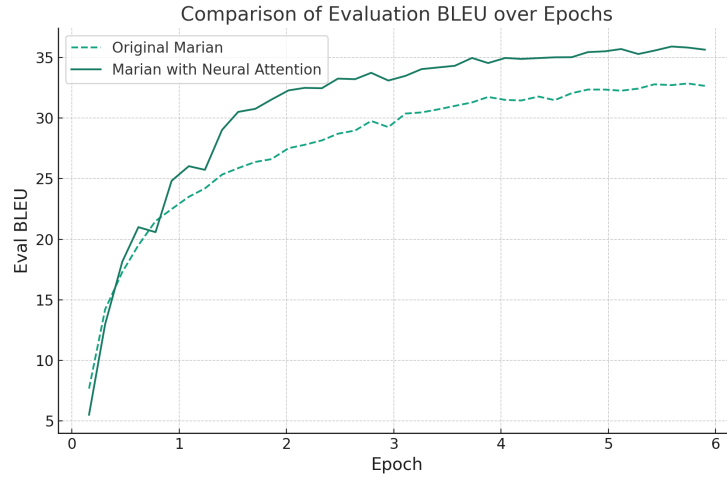

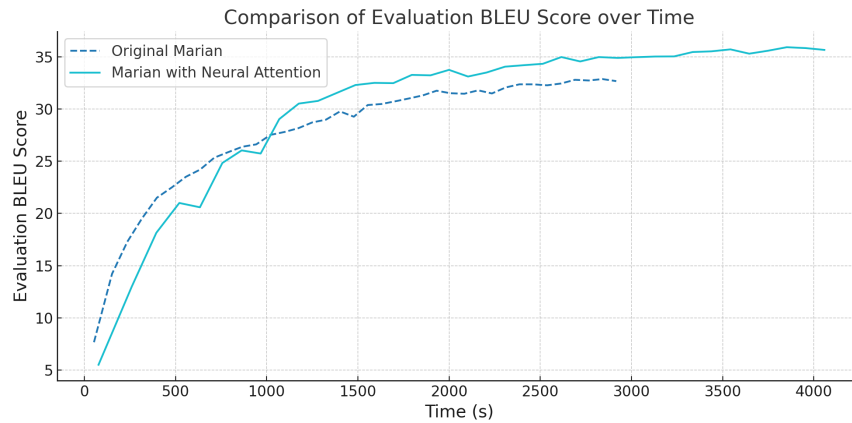Figure 3: BLEU Comparison: Marian with Neural Attention vs. Original Over Epochs

Figure 4: BLEU Comparison: Marian with Neural Attention vs. Original Over Time (seconds)

Model	Attention Mechanism Type	DE-EN BLEU
Marian	Standard	32.63
Marian	Neural Attention	35.76

Table 2: Comparison between the original model and two different attention mechanism types in the DE-EN BLEU score.

4.4 Ablation Study

In order to validate the pivotal role of the ReLU activation function in the enhancement of model performance, we conducted a comparison of methods for calculating keys (Key) and values (Value) in the Marian model on the iwslt2017-de-en dataset.

Dual Linear Projection (DLP): A two-step linear mapping was utilized to compute the keys and values. Initially, the input is mapped to a twice-as-large intermediate representation through a linear layer, then it is mapped back to the original dimension through a second linear layer. Formally, given an input $x \in \mathbb{R}^d$, the computation of keys and values is as follows:

$$k, v = W_2 \cdot \text{LayerNorm}(W_1 \cdot x + b_1) + b_2$$

Neural Attention: On top of DLP, we introduced a ReLU non-linear activation function between the two linear mappings. Therefore, the computation of keys and values becomes:

$$k, v = W_2 \cdot \text{ReLU}(\text{LayerNorm}(W_1 \cdot x + b_1)) + b_2$$

The results of the ablation study using these two mechanisms, along with the standard self-attention for comparison, are presented in Table 3.

Model	Attention Mechanism Type	DE-EN BLEU
Marian	Standard Self-Attention	32.63
Marian	Dual Linear Projection	32.64
Marian	Neural Attention	35.76

Table 3: BLEU score results on the iwslt2017-de-en dataset.

Through the results presented in Table 3, we can observe that the non-linear layer (ReLU) plays a crucial role in the Neural Attention mechanism. The DLP method, which removes the ReLU activation function, did not demonstrate significant performance improvement, while the Neural Attention method markedly increased the BLEU score. This validates our initial intention and the necessity of introducing non-linear transformations.

5 Conclusion

This research delves deeply into the effectiveness and potential of employing Multi-Layer Perceptrons (MLP) for Query, Key, and Value (QKV) computation in the self-attention mechanism. Through a series of experiments and analyses on different models and tasks, we observed significant improvements in model performance in various aspects by introducing non-linear activation functions and complex key and value generation strategies.

In the masked language modeling task, we found that substituting the traditional linear transformation with an MLP neural network significantly improved the Roberta model in terms of Perplexity and other evaluation metrics. Similarly, in the machine translation task, the Marian model also showed a significant enhancement in the BLEU score after introducing a new attention mechanism computation method.

It is noteworthy that, due to computational resource limitations, we were unable to train a large model from scratch to attempt to surpass the existing state-of-the-art (SOTA) models. Our experiments mainly focused on existing pre-trained models and explored the feasibility and effects of fine-tuning models by modifying the self-attention mechanism.

Although we have made some positive strides in this research, there is still ample space for exploration and optimization in using an MLP neural network as the QKV computation method in the self-attention mechanism. Future research directions could include exploring different network architectures, activation functions, and the potential applicability of this method on other NLP tasks and models.

In summary, this research offers a novel approach and perspective, showcasing the potential value of introducing non-linearity and complex computations into the self-attention mechanism, and lays the foundation for further research and exploration. We hope these findings provide insights for researchers and developers in the field of natural language processing and further propel the development and innovation of attention mechanisms.

References

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.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. Marian: Fast neural machine translation in c++. *arXiv preprint arXiv:1804.00344*, 2018.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.6`.

Jörg Tiedemann and Santhosh Thottingal. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal, 2020.