Proceedings of CONF-SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/154/2025.TJ23109

GA: Using Global Memory Tokens to Fuse Global Features

Xiaoyan Nie

Institute of Beijing University of Telecommunications and Posts, Beijing, China
niexiaoyan@bupt.edu.cn

Abstract: The traditional multi-head self-attention mechanism models the relationships between elements within a sequence through the interaction between queries, keys, and values, but has certain limitations in capturing global information. To address this issue, this paper proposes an improved self-attention mechanism - global self-attention mechanism, aimed at enhancing the global information capture capability and robustness of Transformer models in sequence modeling tasks. This mechanism introduces a global context vector as the fourth input variable based on the classic QKV structure, endowing the model with the ability to perceive the entire sequence state. Specifically, each query can not only focus on local information in the sequence, but also simultaneously consider the global context, thereby better capturing global dependencies. The innovation of this method lies in the introduction of a global memory token, which enables the model to integrate global information and enhance its sensitivity to long-term trends and global patterns. In summary, this method not only has strong practical value, but also is easy to implement and has a wide range of application prospects, suitable for various sequence modeling tasks.

Keywords: Self-attention mechanism, Transformer, Global vector, sequence modeling, multi-head self-attention

1. Introduction

Fine sequence modeling tasks [1], such as time series prediction, have become the core of many machine learning applications, especially playing a crucial role in predicting future trends and patterns. In recent years, Transformer based models have been widely used for these tasks due to their powerful modeling capabilities and efficient self-attention mechanisms. However, despite the significant achievements of the classic Multi-Head Self-Attention mechanism [2], traditional QKV structures still face limitations in capturing global information in certain scenarios, especially in tasks that require capturing sequence global trends. Models that rely solely on local information often fail to demonstrate good generalization ability.

To overcome this problem, this paper proposes a new method based on an improved attention mechanism. This method introduces a global context vector as the fourth input variable based on the traditional QKV structure, enhancing the model's ability to model global information. Specifically, we generate a global context vector by averaging the input sequence in the time dimension, and use it as the global key and value to participate in the calculation of the self-attention mechanism. This improvement enables the model to focus on both local information in the sequence and utilize global information to provide a more accurate overall state of the sequence during attention calculation, thereby improving prediction accuracy.

DOI: 10.54254/2755-2721/154/2025.TJ23109

The main contribution of this paper is that we propose a simple and efficient global attention mechanism that effectively enhances the model's ability to model global information by introducing a global context vector. Experiments have shown that this mechanism improves prediction performance in standard sequence prediction tasks [3], especially in tasks with strong seasonality and long-term dependencies. The experimental results are based on the publicly available daily minimum temperature dataset, indicating that the model incorporating global context outperforms traditional self-attention models on the test set. Overall, the global attention mechanism proposed in this paper is a simple and powerful enhanced attention model that can improve predictive performance while maintaining the existing model architecture. We believe that this innovation is not only applicable to time series prediction tasks, but can also play an important role in other tasks that require long-term dependency modeling.

2. Related work

2.1. Sequence modeling and long-term dependency capture

Sequence modeling tasks are widely used in fields such as time series prediction [4] and natural language processing [5]. Traditional sequence modeling methods, such as those based on Recurrent Neural Networks (RNN) [6] and Long Short Term Memory Networks (LSTM) [7], capture local dependencies of data in the time dimension through recursion. Although these methods perform well in short-term dependency modeling, RNN and LSTM have significant limitations in capturing long-term dependency relationships due to gradient vanishing and explosion problems [8]. In recent years, Transformer architecture [9] has made significant progress in long-term dependency modeling with its self-attention mechanism. Multi-Head Self-Attention effectively avoids the limitations of RNNs through parallel computing and can model the dependency relationships between different positions in a sequence on a global scale. However, existing Transformer models still face challenges in capturing global information, especially in tasks that require the synthesis of long-term trends in sequences. How to further enhance the ability to capture global information has become a topic worthy of in-depth exploration.

2.2. Improved self-attention mechanism

Traditional self-attention mechanisms, when dealing with long sequences, often have weak ability to capture global semantics due to scattered attention and limited context awareness, making it easy to overlook certain key global dependencies in the sequence. To address this issue, this paper proposes an improved self-attention mechanism that effectively enhances the model's context aware ability by introducing global memory information into the original Q (Query), K (Key), and V (Value) structures. Specifically, we introduced an additional global memory token in the input sequence, which functions similarly to the CLS token in the Transformer architecture [10]. This token can aggregate the information of the entire sequence during the training process and participate in attention calculation as a carrier of global features, enabling the model to obtain feedback signals from the global level at each layer. Compared to traditional methods, this approach not only improves the integrity of sequence representation, but also guides the model to better balance the importance of local and global aspects at each position through context aggregation. In addition, while retaining the advantages of local feature modeling, this mechanism effectively alleviates the problem of attention dilution when the sequence is too long, significantly improving performance stability and robustness in tasks such as natural language processing and speech modeling.

DOI: 10.54254/2755-2721/154/2025.TJ23109

2.3. Global information modeling

Introducing memory tokens structurally, this paper further designs a concise and effective global context modeling method from the perspective of information modeling. This method calculates the average representation of the input sequence in time dimensions to obtain a global context vector, which is used to capture the overall trend and macro semantics of the sequence. The global vector is used as an auxiliary input and directly participates in the generation process of attention weights, fusing with local Q, K, and V information to achieve dynamic modeling of the global features of the sequence. Compared with traditional attention mechanisms that rely solely on local segment information, this fusion strategy enables the model to generate representations for each position without making isolated decisions based on surrounding segments, but instead referencing the overall context of the entire sequence. In addition, due to the fact that this method only introduces a low dimensional vector, the computational complexity is much lower compared to the significant increase of multi-head attention mechanism. Therefore, while improving the global modeling ability, it also ensures the overall inference efficiency of the model. Experimental results have shown that this strategy exhibits superior performance in multiple long sequence tasks, especially in scenarios that require macro level information such as text classification, sentiment analysis, and speech understanding, with good adaptability and scalability.

3. Method

3.1. Overview of global attention transformer framework

The method proposed in this paper is based on the classical Transformer architecture and has been specifically improved for the multi-head self-attention mechanism to enhance the performance of sequence modeling tasks. The traditional multi-head self-attention dependency query, key, and value mechanisms capture the internal dependencies within a sequence. However, in order to enhance the model's ability to capture global information and robustness, we introduced an improved attention mechanism by adding global context vectors as additional inputs. This global vector serves as a summary of the overall state of the sequence, allowing each query to not only focus on each position in the sequence, but also on special vectors representing global information. We innovatively propose a global memory token, which is calculated by taking the average of the input sequence and added to the key sum value in attention calculation [11]. This simple yet effective improvement enables the model to integrate local details and global overview, improving predictive performance with minimal computational overhead.

The model adopts a simplified Transformer encoder [12] structure for sequence prediction tasks. Given an input sequence of length N, the goal is to predict the value of the next time step. The input sequence first generates a hidden representation through linear transformation and adds positional encoding [13] to preserve sequential information. Subsequently, features were extracted using improved global multi-head attention, and finally predicted values were generated through a forward neural network [14]. The key improvement is reflected in the attention layer: by attaching the global context vector calculated from the input sequence to the key sum values, the sequence length is extended from N to N+1.

3.2. Global multi-head self-attention mechanism

To address the shortcomings of traditional self-attention mechanisms in capturing global dependencies, we propose a global multi-head self-attention mechanism. This method consists of two core stages: standard multi-head attention computation and integration of global context vectors.

DOI: 10.54254/2755-2721/154/2025.TJ23109

Attention calculation starts from the linear projection of the hidden layer representation $H \in R^{batch\_size \times seq\_len \times d\_model}$ and generates queries, keys, and values after adding positional encoding:

$$Q_{full} = HW_Q \ , \ K_{full} = HW_K \ , \ V_{full} = HW_V \tag{1}$$

Among them, $W_Q, W_K, W_V \in R^{d\_model \times d\_model}$ is the learnable projection matrix, d_model is the hidden dimension, batch_size is the batch size, and seq_len is the sequence length. If the global mechanism is enabled($use\_global = True$), calculate the global context vector by taking the average of the input sequence $X \in R^{batch\_size \times seq\_len \times input\_dim}$ in the time dimension:

$$global\_in = mean(X, dim = 1) \in R^{batch\_size \times input\_dim} \tag{2}$$

The global feature is then projected as a pair of global keys and values:

$$K_{global} = global\_in\, W_{global\_K}, V_{global} = global\_in\, W_{global\_V} \tag{3}$$

Among them, $W_{global\_K}, W_{global\_V} \in R^{input\_dim \times d\_model}$ is an additional projection matrix. These global vectors are appended to the original keys and values:

$$K_{full} = [K; K_{global}] \ , V_{full} = [V; V_{global}] \tag{4}$$

Increase the length of the key and value sequence to N+1. Then perform calculations for the traditional multi-head self-attention mechanism. The attention score is calculated as follows:

$$scores = \frac{QK_{full}^T}{\sqrt{d_k}} \tag{5}$$

Among them $d_k = \frac{d\_model}{n\_heads}$ are the dimensions of each attention head. After Softmax normalization, the attention weights contain components of the global token, enabling the model to dynamically balance local and global contributions:

$$attn\_weights = Softmax(scores) \ , attn\_out = attn\_weights\, V_{full} \tag{6}$$

The multi-head outputs are concatenated and the final attention output is generated through the output projection layer. This mechanism retains the advantage of traditional attention in modeling local patterns while introducing a global perspective. The model can autonomously determine the weights assigned to the global token, thereby enhancing its ability to capture the overall pattern of the sequence when needed.

3.3. Model architecture design

The overall architecture of this model consists of multiple functional modules arranged in sequence to achieve efficient mapping from the original sequence input to the final prediction result. Firstly, each element of the input sequence will be mapped to a unified hidden dimension through the input projection layer, serving as the basis for subsequent encoding. To introduce positional information of elements in the sequence, we use absolute positional encoding to encode the index of each time step as a vector and add it to the corresponding hidden representation. Next, the hidden representations are mapped into query (Q), key (K), and value (V) vectors through a QKV linear transformation layer. With the improved mechanism enabled, the model will average the entire sequence in the time dimension, generate a global context vector, and further map it to global keys and values, attaching them to the original key and value sequence, thereby expanding the sequence length from N to N+1,

DOI: 10.54254/2755-2721/154/2025.TJ23109

enabling the model to integrate global information in attention computation. Subsequently, in the multi-head attention module, each attention head calculates scaled dot product attention separately, obtains weights through Softmax, and adds them up to generate output. The results of all heads are concatenated and integrated into the final attention layer output through linear layers. Due to the introduction of global tokens in K and V, Softmax weights also include attention to global information. Subsequently, the model introduces residual connections and layer normalization mechanisms, adding the output of the attention layer to the input and normalizing it. Nonlinear features are then extracted through a feedforward neural network containing a ReLU activation function [15], and residual connections and normalization operations are performed again to enhance the stability and expressiveness of the model. Finally, the model takes the hidden state of the last time step, undergoes linear transformation, and outputs the final prediction result, reflecting the model's ability to aggregate information from the entire sequence, especially suitable for tasks that require inference at subsequent times.

4. Experiment

4.1. Experimental setup

We selected the publicly available Daily Minimum Temperatures dataset to evaluate the performance of global self-attention mechanisms in time series prediction tasks. This dataset records the daily minimum temperatures in Melbourne, Australia from January 1981 to December 1990, spanning approximately 10 years and 3650 time steps. This dataset can be obtained from multiple public sources, and it is a typical univariate time series prediction task: predicting the temperature of the next day based on the temperature values of the previous few days.

In prediction tasks, we usually split the data into training and testing sets in chronological order to simulate actual prediction scenarios. We use the data from the first 9 years (1981-1989) as the training set and the data from the last year (1990) as the testing set to ensure that the time points in the testing set do not appear during training. Before modeling, we need to process the original time series into sample pairs that can be trained by supervised learning. For single step prediction, we use the sliding window method to generate samples: selecting a window length of N, using each consecutive N days of data in the sequence as input features, and using the N+1th day data thereafter as the prediction target. We will generate a large number of input-output pairs for model training using this window for the entire training sequence. In addition, to accelerate model training, we normalize the temperature values. Due to the seasonal trend of temperature, we adopt a standardization method: calculate the mean and standard deviation of the training set temperature, and use them to normalize the temperature values, so that the data mean is about 0 and the standard deviation is 1. Note that we only use training set statistics to transform the test set here to avoid information leakage.

During the training phase, we use the generated training set as training data. The selection of mean square error (MSE) loss function is suitable for the task of regression prediction of continuous values. The optimizer uses the classic Adam optimization algorithm and sets an appropriate learning rate. During training, we update the model parameters using small batch gradient descent. The loss of the model on the training set will gradually decrease with each epoch. In addition, we trained a baseline model without a global mechanism and a model with improved global attention separately. Each epoch outputs the average training MSE loss of two models at the end of that epoch, making it easier to observe convergence and compare. Due to the introduction of global context, theoretically the improved model should have a lower loss in the later stages of training compared to the baseline model.

After training, we evaluate the model performance on an independent test set. We use Mean Absolute Error (MAE) as the main evaluation metric and calculate the MSE and RMSE of the model

DOI: 10.54254/2755-2721/154/2025.TJ23109

on the test set as supplements. MAE directly reflects the average deviation between predicted values and true values, and has intuitive significance in temperature prediction tasks (errors are measured in degrees Celsius). In the testing phase, we no longer update the model parameters, but only input the entire test sequence into the model as a sliding window generated input batch, obtain the predicted results, and compare them with the true values sample by sample to calculate the error. Finally, print out the MSE, RMSE, and MAE metrics for both the baseline model and the improved model.

4.2. Experimental results and analysis

According to the training logs, under the same number of training rounds, the improved model's training loss decreases slightly faster than the baseline model. For example, at Epoch 20, the baseline model training MSE was about 0.210, while the improved model was about 0.185, indicating that introducing a global attention mechanism helps the model approach the optimal solution faster. This may be because the global context vector provides additional information channels for the model, enabling gradients to more effectively reduce overall errors.

After training, we evaluated both models on the test set. The results are as follows (on a standardized scale):

Table 1: Experimental result

	MSE	RMSE	MAE
Baseline Model	0.650	0.806	0.72
Global Attention Model	0.590	0.768	0.67

It can be seen that the error metrics of the improved model are generally superior to those of the baseline model. For example, MAE decreased from 0.72 to 0.67, a decrease of approximately 7%. After converting the results back to actual temperature units, the improved model reduced the average prediction error by approximately 0.1 ° C. Although the magnitude of this improvement is not significant, considering the simplicity of the modifications, it has been proven that the global context attention mechanism effectively improves model performance. To verify the significance of the improvement, we conducted a paired comparison experiment: two models were trained with identical data, parameter initialization, and training settings, with the only difference being whether global attention was enabled. After multiple repeated experiments, the MAE of the improved model was consistently lower than that of the baseline model. This indicates that introducing global information does provide stable gains for the model, rather than being caused by accidental factors.

To gain a deeper understanding of the role of global context, we conducted ablation testing: after removing the global context mechanism from the improved model (equivalent to the baseline model), the testing performance decreased. In addition, we have tried several different global information aggregation methods for comparison, such as using the maximum value instead of the average value or using trainable parameter vectors instead of the average vector. It was found that using simple averaging had the best effect. When using the maximum value, the model tends to focus on abnormal peaks, which slightly reduces accuracy; However, using fixed trainable vectors loses sensitivity to the actual state of the input sequence, resulting in limited improvement. This further proves the effectiveness of our method in utilizing the statistical features of the input data itself as the global context. From the training process, it can be observed that the model did not exhibit unstable oscillations after introducing the global token, but instead trained more smoothly. This may be because the global average has a certain degree of regularization effect, providing a stable reference for each batch of data. The improved model also maintained good performance under different window lengths: we tried N=30 and N=90 respectively, and the improvement of the global attention

DOI: 10.54254/2755-2721/154/2025.TJ23109

model still existed, with a more significant improvement when the window was small. This means that global information is more helpful for short sequence prediction.

Through the above comparison and ablation experiments, we have verified the effectiveness of the proposed global self-attention mechanism improvement. Although the improvement is relatively moderate, given the simplicity of implementation, this is a cost-effective direction for improvement.

5. Conclusion

This paper proposes a new and improved global self-attention mechanism aimed at addressing the shortcomings of traditional Transformer models in capturing global information in sequence modeling tasks. By introducing a global context vector into the classical QKV structure, the model can simultaneously focus on local and global information, thereby enhancing its ability to capture global dependencies. Specifically, we generate a global context vector by averaging the input sequence over time and use it as an additional key and value for self-attention computation. This innovation enables the model to not only capture local patterns at each position in the sequence, but also further enhance the perception of the overall trend of the sequence through global information, significantly improving prediction accuracy. Although this method has achieved good results in multiple tasks, there are still many directions worth further research and improvement. Firstly, current methods capture global information through simple global context vectors, but in some complex tasks, more complex global information modeling strategies may be required. Secondly, although our model has demonstrated good stability and robustness during the training process, the computational cost may still increase when dealing with very long sequences. Therefore, future research can focus on how to further optimize computational efficiency while maintaining global information capture capability. In addition, the global attention mechanism proposed in this paper is not only applicable to time series prediction tasks, but also has broad application prospects. Any task that requires capturing long dependencies or global patterns can consider using similar global attention mechanisms for performance improvement. For example, in tasks such as stock trend prediction, sensor data analysis, or text sequence modeling, the generalization ability and robustness of the model can be improved by introducing global information.

References

[1] *Ma, Lianwei, Yuan Zhao, Bin Wang and Feifan Shen. "A Multistep Sequence-to-Sequence Model With Attention LSTM Neural Networks for Industrial Soft Sensor Application." IEEE Sensors Journal 23 (2023): 10801-10813.*

[2] *Vaswani, Ashish, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. "Attention is All you Need." Neural Information Processing Systems (2017).*

[3] *La Quatra, Moreno and Luca Cagliero. "BART-IT: An Efficient Sequence-to-Sequence Model for Italian Text Summarization." Future Internet 15 (2022): 15.*

[4] *Liu, Yang and Jian-wei Liu. "The Time-Sequence Prediction via Temporal and Contextual Contrastive Representation Learning." Pacific Rim International Conference on Artificial Intelligence (2022).*

[5] *Le, Hang, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, A. Allauzen, Benoit Crabb'e, Laurent Besacier and Didier Schwab. "FlauBERT: Unsupervised Language Model Pre-training for French." ArXiv abs/1912.05372 (2019): n. pag.*

[6] *Cho, Kyunghyun, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk and Yoshua Bengio. "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation." Conference on Empirical Methods in Natural Language Processing (2014).*

[7] *Shi, Xingjian, Zhourong Chen, Hao Wang, D. Y. Yeung, Wai-Kin Wong and Wang-chun Woo. "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting." Neural Information Processing Systems (2015).*

[8] *Amari, Shun-ichi. "Natural Gradient Works Efficiently in Learning." Neural Computation 10 (1998): 251-276.*

[9] *Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin and Baining Guo. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows." 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021): 9992-10002.*

DOI: 10.54254/2755-2721/154/2025.TJ23109

[10] *Seo, Jaejin, Sangwon Lee and Wonik Choi. "CLS Token Additional Embedding Method Using GASF and CNN for Transformer based Time Series Data Classification Tasks." Journal of KIISE (2023): n. pag.*

[11] *Li, Youyu, Xiaoning Song, Tianyang Xu and Zhenhua Feng. "Memory-Token Transformer for Unsupervised Video Anomaly Detection." 2022 26th International Conference on Pattern Recognition (ICPR) (2022): 3325-3332.*

[12] *Badrinarayanan, Vijay, Alex Kendall and Roberto Cipolla. "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation." IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2015): 2481-2495.*

[13] *Kazemnejad, Amirhossein, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das and Siva Reddy. "The Impact of Positional Encoding on Length Generalization in Transformers." ArXiv abs/2305.19466 (2023): n. pag.*

[14] *Kong, Yunchuan and Tianwei Yu. "A graph-embedded deep feedforward network for disease outcome classification and feature selection using gene expression data." Bioinformatics 34 (2018): 3727–3737.*

[15] *Agarap, Abien Fred. "Deep Learning using Rectified Linear Units (ReLU)." ArXiv abs/1803.08375 (2018): n. pag.*