Proceedings of the 3rd International Conference on Mechatronics and Smart Systems
DOI: 10.54254/2755-2721/125/2025.20447

Single Dimension Attention and Mixed Attention

Haoran Qiu[1,a,*]

[1]*Computer Science and Technology Institute, Donghua University, Shanghai, China*
a. 231310231@mail.dhu.edu.cn
**corresponding author*

Abstract: Due to its ability to effectively improve model performance, the attention mechanism has gained widespread adoption across a variety of fields in recent years as neural network models have been developed. Compared to traditional single attention mechanisms, mixed attention offers more diverse prediction dimensions, enabling a more comprehensive data processing approach. This, in turn, significantly improves both prediction accuracy and overall model performance. The main body of this article begins by introducing the single attention mechanism, followed by a detailed explanation of the mixed attention mechanism. It then addresses the current limitations of mixed attention and outlines potential future directions for research. The article seeks to advance the understanding of mixed attention by giving a broad overview of the state of research in this area. In conclusion, the exploration of mixed attention mechanisms holds considerable promise for advancing model performance across a wide range of applications. Future research will likely focus on addressing existing challenges, further unlocking its potential and broadening its scope of use.

Keywords: Attention, Deep Learning, Neural Network, Image Processing, Recommendation Algorithm.

1. Introduction

With the continuous development and iteration of machine learning technology, deep learning has become a hot research field in recent years. Among the various methods of deep learning, neural networks are highly favored and widely used in various methods. Attention is one of the key methods that can effectively improve the performance of the model. The inspiration for the attention mechanism comes from the fact that humans can allocate attention to certain key contents and ignore useless information. By concentrating on important information in various dimensions and removing irrelevant information, the attention mechanism enhances the interference of outside noise on data processing and raises the prediction model's accuracy [1].

The mechanisms of attention have drawn a lot of attention lately and have advanced quickly. However, during the process, the single-head attention mechanism has gradually exposed its limitations, that is, it has limited performance in dealing with complex relationships and diverse features. Therefore, the multi-head attention mechanism came into being. By combining multiple attention heads to expand the representation ability of the model, the model performance is improved and the limitation of a single dimension is broken.

This paper first introduces the single-head attention mechanism, then the hybrid attention mechanism, and finally introduces the existing limitations and future prospects. This paper mainly

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/125/2025.20447

analyzes the four mixed attention mechanisms and introduces the current implementation principles of the mixed attention mechanism and the application of mixed attention in different directions, and aims to summarize the current development of mixed attention.

In conclusion, this review not only highlights the advancements and challenges in the field of mixed attention but also emphasizes its growing importance in enhancing model performance across diverse tasks. This paper attempts to inspire further investigation into optimizing mixed attention for practical applications by providing a foundation for future research by summarizing the state-of-the-art methods.

2. Single-head Attention Mechanisms

2.1. Self-Attention Mechanism

One particular variation of the attention mechanism is the Self-Attention Mechanism. A certain correlation exists between the various sets of vectors of varying sizes that are typically sent to the machine model in the conventional attention mechanism. However, in the actual training process, these internal correlations are often difficult to capture and utilize effectively. Compared with the traditional attention mechanism, the self-attention mechanism relies more on the input data itself and aims to explore the relationship between the elements within the data. It relies less on external information, so it can more efficiently capture the correlation within the input data, improving the expression ability and performance of the model.

2.2. Channel Attention Mechanism

Computer vision makes extensive use of the channel attention mechanism. Assigning weights to various channels by calculating their respective contributions to the input data is its main function. This method enables the model to automatically adjust the weight of the input data according to the importance of each channel to enhance the focus on image features. In contrast to other kinds of attention mechanisms, the channel attention mechanism disregards the spatial position variations of each set of image data in the channel and instead concentrates more on the significance of various channels [2].

2.3. Spatial Attention Mechanism

Processing spatial information is typically accomplished through the spatial attention mechanism. Its core idea is to enhance important information and suppress interference factors by dynamically adjusting the weights of input data at different spatial locations. Currently, the spatial attention mechanism has been widely used in fields such as image processing [3]. The feature of the spatial attention mechanism is that it processes the most significant portion of the data, converts the spatial information in the original data through space, weighs each portion while keeping the characteristic information, and outputs it based on the weight in order to effectively convey the key area's characteristics.

3. Formatting the Text

As single attention can only process data in a single channel, the limitation of the dimension of the data processing channel leads to the fact that the accuracy of the data processing results of single attention is still flawed. In order to solve this problem, mixed attention came into being.

Mixed attention is a mixed attention network that combines multiple attention mechanism channels. Through the joint action of multiple attention mechanisms, the accuracy of the output results can be greatly improved. It can be said that it combines the advantages of various attention mechanisms.

DOI: 10.54254/2755-2721/125/2025.20447

Commonly used mixed attention mechanisms are the Convolutional Block Attention Module (CBAM), Bottleneck Attention Module (BAM), Hybrid Attention transformer (HAT) for image restoration, and Mixed Attention Network (MAN).

3.1. Convolutional Block Attention Module

Channel attention and spatial attention are the two modules that makeup CBAM, which is based on an effective body structure. Data in the two dimensions of space and channel are subjected to attention processing. It is a lightweight attention module [4].

CBAM is an attention module that can be plug-and-play with a wide range of network models. Additionally, CBAM's channel attention module and spatial attention module can save computing power and parameter overhead by being arranged serially. The CBAM workflow is depicted in Figure 1.

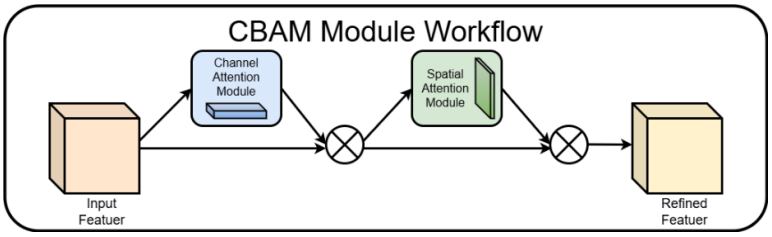

Figure 1: CBAM Module Workflow [4].

The data is initially routed to the channel attention module in CBAM after the data features have been input. Immediately after exiting the channel attention module, the data will enter the spatial attention module. Ultimately, the refined features are obtained after being processed by the spatial attention module.

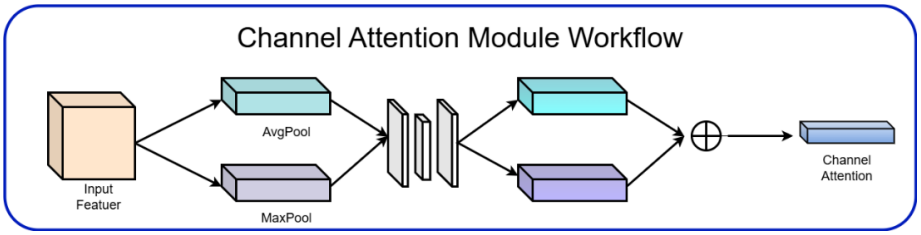

Figure 2: Channel Attention Module Workflow [4].

Figure 2 illustrates how the channel attention module operates. The data processing procedure in the channel attention module involves first performing pooling operations on the image using average pooling and maximum pooling to aggregate spatial information. Next, two spatial context descriptions representing average pooling elements and spatial pooling elements, respectively, are generated. The channel attention map is then produced by feeding the two descriptions into a multi-layer perceptron shared network with one hidden layer. The output feature vector is then combined with element-by-element summation after the shared network has been applied to each descriptor.

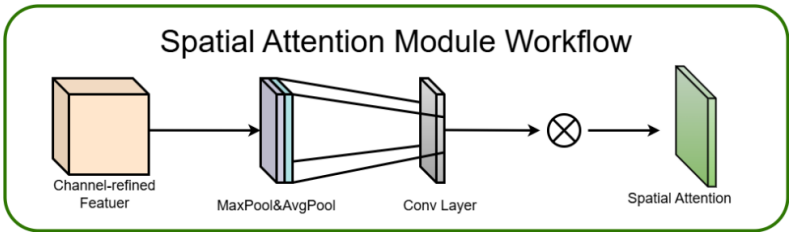

Figure 3: Spatial Attention Module Workflow [4].

DOI: 10.54254/2755-2721/125/2025.20447

Figure 3 illustrates the spatial module's workflow. The spatial relationship between features is utilized in the spatial attention module to create a spatial attention map. In order to augment the channel attention, the spatial attention examines and concentrates on the informative portion of the data. In its internal path, the spatial attention module carries out average pooling and maximum pooling operations, just like channel attention. In contrast to the channel attention module, the spatial attention channel involves stacking the average and maximum pooling layers to create a single pooling layer. Additionally, the data is processed in the order of maximum pooling first, followed by average pooling. Consequently, two efficient two-dimensional mapping feature descriptors can be produced by the spatial attention module by better aggregating the channel information of the feature map. A two-dimensional spatial attention map is then produced by performing a convolution operation on the two feature descriptors.

Apart from conserving parameters and processing power, CBAM ensures simplicity of integration as a plug-and-play module into the existing network architecture, which enhances the accuracy of the inserted model. Convolutional neural networks (CNNs) and the YOLO algorithm are frequently used in conjunction with CBAM in the most recent applications in order to increase prediction accuracy [5, 6, 7].

3.2. Bottleneck Attention Module

BAM is a straightforward and efficient attention module that works with any feedforward neural network. Channel attention and spatial attention are the two separate routes that BAM uses to infer an attention map [8]. BAM's ability to enhance the network's feature expression after being inserted is one of its main advantages. Figure 4 depicts the BAM workflow.

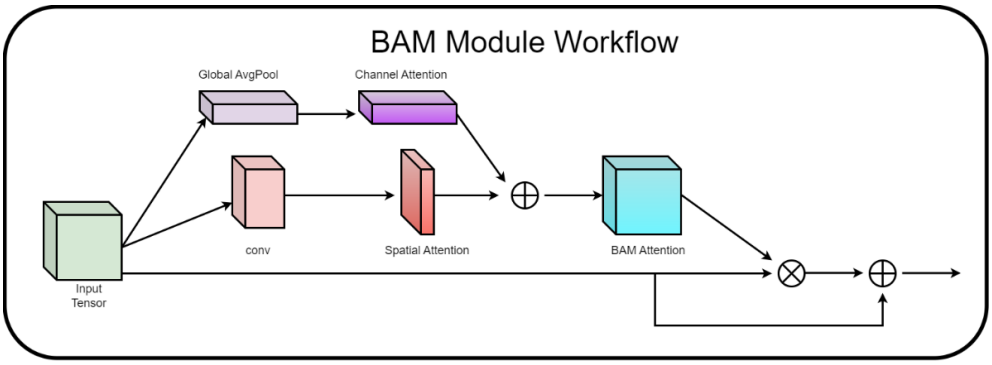

Figure 4: BAM Module Workflow [8].

In a BAM module, a feature tensor is input, and the module processes the data through two independent paths: channel attention and spatial attention. In the channel attention path, the data is globally pooled, and a channel attention path feature result is finally obtained; in the other channel, the spatial attention channel, the data is convolved, and a convolution attention path feature result is finally obtained. Next, the two feature results are summed by the principal element to obtain a BAM attention feature map. In addition, parallel to these two paths, there is also a feature map path, which performs mixed product operations and element-by-element summation operations with the generated attention map to ensure that the final output result still retains the characteristics of the original data. This path can be regarded as a feature detector, and the other two spatial and channel branches focus on "what" and "where". Like CBAM, BAM can also save redundant calculation steps and parameter overhead. Coupled with BAM's simple module characteristics, BAM can be well promoted and applied to various models to process large amounts of data, thereby greatly improving the capacity of network models using BAM.

DOI: 10.54254/2755-2721/125/2025.20447

At present, BAM is also combined with different network models in many aspects to achieve more extensive applications, such as combining with the SNet network for face recognition and facial expression recognition or combining with the CNN network for remote sensing image segmentation[9, 10]. Naturally, as BAM modules have become more widely used in recent years, BAM has also produced other enhanced versions, such as the Transformer-based TransBoNet module, which substitutes self-attention for spatial attention in BAM [11].

3.3. Hybrid Attention Transformer

Channel attention and window-based self-attention are combined in HAT, a hybrid attention transformer. The attribution analysis method Local Attribution Map (LAM) is the only way that the traditional Transformer-based method can use limited input spatial information, which means that the Transformer's potential is not well reflected in the current network. This is why HAT was proposed. To address this issue and improve the extraction of the input image's feature information, HAT was proposed [12]. In order to better aggregate cross-window information, HAT focuses on the feature correlation between neighboring windows because it is closely related to the self-attention mechanism which is based on the window. To improve how adjacent window features interact, HAT adds an overlapping cross-attention module. The same task pre-training approach is used by HAT during the training phase to further investigate the possibility of further model improvement. The network framework of HAT is shown in Figure 5.

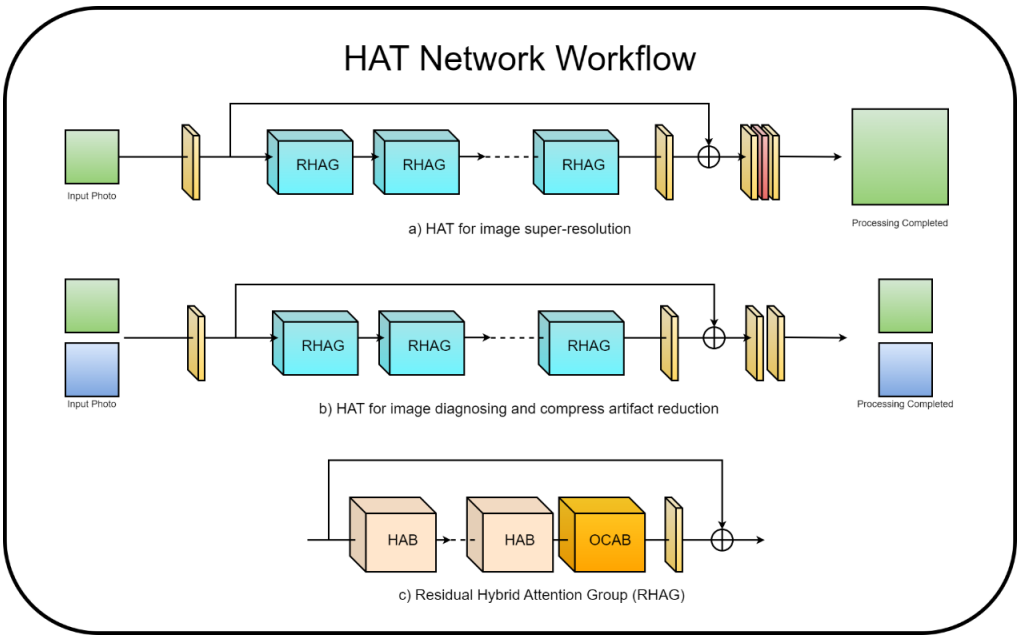

Figure 5: HAT Network Workflow [12].

The HAT network is mainly used in two aspects: image super-resolution improvement and image denoising. The basic logic of both methods is to process the image information in parallel through several Residual Hybrid Attention Groups (RHAGs) after transformation, and then sum the data processed by the RHAG layer with the data processed once, and then perform another processing operation. However, in this operation before generating the result, the super-resolution function has one more pixel shuffling operation sandwiched in the processing layer than the image denoising function. One could argue that the dominant RHAG group is the central component of the HAT network, while an RHAB group is made up of an Overlapping Cross Attention Block (OCAB) and multiple Hybrid Attention Blocks (HABs). In the HAB module, the channel attention mechanism and

DOI: 10.54254/2755-2721/125/2025.20447

a standard Swin Transformer structure are combined, and in the OCAB module, an overlapping attention window module is introduced to achieve a more direct interaction between adjacent window features and reduce the generation of pseudo shadows that will cause blockage.

3.4. Mixed Attention Network

MAN is different from the three previously mentioned mixed attention mechanisms or modules applied in the field of traditional computer vision. The goal of this network is to improve the performance of data-scarce domains by addressing the problem of data sparsity that occurs in recommendation algorithms and training the model with data from multiple domains [13].

Both domain-specific and cross-domain information can be extracted using the MAN network's local and global attention modules. Its operating mechanism is shown in Figure 6.

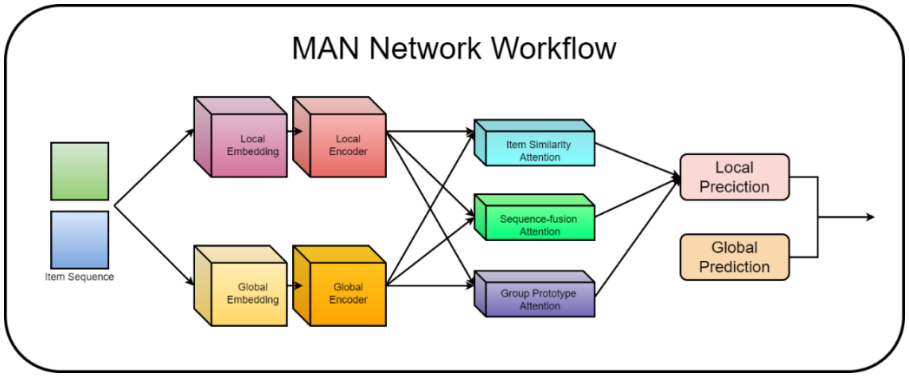

Figure 6: MAN Network Workflow [13].

MAN is composed of three primary layers that are local-global prediction, mixed attention, and local-global encoding. The first module of the MAN network, the local and global encoding layer, generates local and global embeddings for each set of data and extracts local and global sequences using local and global encoders. The data will then move on to the second module, the mixed attention layer, which comprises all prototype attention modules, sequence fusion attention, and project similarity attention. After passing through these three attention modules, the data is integrated with the target project features in the prediction layer to obtain a local prediction. The target project features are also fused with the global sequence in the prediction layer to generate a global prediction, and finally, the local and global predictions are integrated to obtain the output result. In the project similarity attention module, the local sequence and the global sequence are element-by-element spherical and operated to obtain a new set of sequences, which are combined with the target features and processed by the attention mechanism to obtain a set of predicted sequences. In the sequence fusion attention mechanism, the local sequence and the global sequence are fused after being processed by the cross attention mechanism to form another set of predicted sequences. In the whole prototype attention module, the local sequence is split into two operations: aggregation and pooling. In the split path, the sequence will reorganize the elements and finally perform a mixed product operation with the result obtained by the pooling path to obtain the operation result.

4. Existing Limitations and Future Prospects

4.1. Existing Limitations

Mixed attention mechanisms need to be implemented through different series or parallel paths to increase the accuracy of predictions. By assisting the machine in concentrating on the characteristics of the input data during training and learning, the attention mechanism can reduce parameter overhead

DOI: 10.54254/2755-2721/125/2025.20447

compared to conventional techniques. However, the current mixed attention operation will still increase the number of calculation steps accordingly, making this method more complex than the single attention network. This, in turn, raises the computational cost of training and reasoning the model.

4.2. Future Prospects

Currently, mixed attention networks are being used in a growing number of different fields and have produced positive outcomes in related fields. More sophisticated iterations of fundamental mixed attention modules or mixed attention network models have emerged concurrently with the rapid advancement of this technology. The optimal solution can be achieved by inserting various mixed attention modules into various network models and achieving different effects through various module combinations and permutations. With technology's constant advancement and modification, more improvements in a variety of fields will result from the mixed attention mechanism becoming more precise and effective.

5. Conclusion

By integrating various channels and features in various dimensions, mixed attention enhances the model's overall performance and prediction accuracy and precision. Mixed attention solves the prediction accuracy bottleneck problem faced by traditional single-head attention due to the single dimension. This article first introduces the single-head attention mechanism, and then summarizes some existing mixed attention methods, focusing on CBAM, BAM, HAT and MAN. At present, the introduction of mixed attention in this article still has limitations in the number and depth of methods. Subsequent work will focus on these two points, and conduct a more comprehensive and in-depth review of the mixed attention mechanism. It is also expected that a more efficient and accurate mixed attention mechanism will be proposed in the future. For the reference list. If there are two or more appendices they should be called appendix A, appendix B, etc. Numbered equations should be in the form (A.1), (A.2), etc., figures should appear as figure A1, figure B1, etc. and tables as table A1, table B1, etc.

References

[1] *Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. Neurocomputing, 452, 48-62.*

[2] *Yang, X., Luo, Y., Li, M., Yang, Z., Sun, C., & Li, W. (2021). Recognizing pests in field-based images by combining spatial and channel attention mechanism. IEEE Access, 9, 162448-162458.*

[3] *Chen, C., Gong, D., Wang, H., Li, Z., & Wong, K. Y. K. (2020). Learning spatial attention for face super-resolution. IEEE Transactions on Image Processing, 30, 1219-1231.*

[4] *Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV) (pp. 3-19).*

[5] *Zhang, S., Jing, Y., & Liang, Y. (2024). EACVP: An ESM-2 LM Framework Combined CNN and CBAM Attention to Predict Anti-coronavirus Peptides. Current Medicinal Chemistry.*

[6] *Mekruksavanich, S., Jantawong, P., Phaphan, W., & Jitpattanakul, A. (2024, January). Hybrid Attention with CNN-BiLSTM and CBAM for Efficient Wearable Activity Recognition. In 2024 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON) (pp. 572-576). IEEE.*

[7] *Yan, J., Zeng, Y., Lin, J., Pei, Z., Fan, J., Fang, C., & Cai, Y. (2024). Enhanced object detection in pediatric bronchoscopy images using YOLO-based algorithms with CBAM attention mechanism. Heliyon, 10(12).*

[8] *Park, J. (2018). Bam: Bottleneck attention module. arXiv preprint arXiv:1807.06514.*

[9] *Sundaram, S. M., & Narayanan, R. (2023). Human Face and Facial Expression Recognition Using Deep Learning and SNet Architecture Integrated with BottleNeck Attention Module. Traitement du Signal, 40(2).*

DOI: 10.54254/2755-2721/125/2025.20447

[10] *Shou, Y., Ai, W., Meng, T., & Yin, N. (2023). Graph information bottleneck for remote sensing segmentation. arXiv preprint arXiv:2312.02545.*

[11] *Song, X., Li, H., Liang, L., Shi, W., Xie, G., Lu, X., & Hei, X. (2024). TransBoNet: Learning camera localization with Transformer Bottleneck and Attention. Pattern Recognition, 146, 109975.*

[12] *Chen, X., Wang, X., Zhang, W., Kong, X., Qiao, Y., Zhou, J., & Dong, C. (2023). Hat: Hybrid attention transformer for image restoration. arXiv preprint arXiv:2309.05239.*

[13] *Lin, G., Gao, C., Zheng, Y., Chang, J., Niu, Y., Song, Y., ... & Wang, M. (2024, March). Mixed Attention Network for Cross-domain Sequential Recommendation. In Proceedings of the 17th ACM International Conference on Web Search and Data Mining (pp. 405-413).*