applied sciences

MDPI

Article

Attention Block Based on Binary Pooling

Chang Chen * and Huaixiang Zhang

School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China
* Correspondence: 211050109@hdu.edu.cn

Abstract: Image classification has become highly significant in the field of computer vision due to its wide array of applications. In recent years, Convolutional Neural Networks (CNN) have emerged as potent tools for addressing this task. Attention mechanisms offer an effective approach to enhance the accuracy of image classification. Despite Global Average Pooling (GAP) being a crucial component of traditional attention mechanisms, it only computes the average of spatial elements in each channel, failing to capture the complete range of feature information, resulting in fewer and less expressive features. To address this limitation, we propose a novel pooling operation named "Binary Pooling" and integrate it into the attention block. Binary pooling combines both GAP and Global Max Pooling (GMP), obtaining a more comprehensive feature vector by extracting average and maximum values, thereby enriching the diversity of extracted image features. Furthermore, to further enhance the extraction of image features, dilation operations and pointwise convolutions are applied on the channel-wise. The proposed attention block is simple yet highly effective. Upon integration into ResNet18/50 models, it leads to accuracy improvements of 2.02%/0.63% on ImageNet.

Keywords: ResNet; attention; image classification

1. Introduction

Computer vision is widely recognized as a critical component of artificial intelligence, as it enables machines to "see" and understand the physical world. Over the past two decades, computer vision has undergone rapid development, with the emergence of numerous theories and methods, resulting in significant progress in various core issues. Over the years, with the advancement of deep learning technology, image classification has been widely studied and applied successfully in several domains such as medical image analysis, autonomous driving, security monitoring, and garbage classification, among others. Jakub Kufel [1] indicated that artificial intelligence demonstrates promising outcomes in the field of medicine. Particularly in radiation therapy, studies by Liesbeth Vandewinckele [2], Guangqi Li [3], Jakub Kufel [4], and Krithika Rangarajan [5] have all shown that image classification technology can assist doctors in identifying patient conditions. The potential of image classification technology is enormous.

Historically, image classification relied on statistical learning techniques such as Bayesian classifiers and K-Nearest Neighbors for feature extraction and pattern recognition. However, these approaches were limited in handling larger datasets and could not scale to larger tasks. The advent of artificial neural networks, particularly CNNs, revolutionized image classification, providing a flexible and powerful method for training and reasoning with large datasets. CNNs have become the gold standard for image classification and have resulted in significant advancements in image processing and computer vision. The limitations in the structure and learning algorithms of early neural network models hindered their ability to effectively address complex image processing problems. However, in the late 1980s and early 1990s, LeCun [6] first proposed a convolutional neural network model combining convolution operation, pooling operation, and nonlinear activation function, which can effectively process image data. In the 2012 ImageNet large-scale visual recognition challenge, AlexNet [7] achieved a significant result, surpassing traditional

Citation: Chen, C.; Zhang, H. Attention Block Based on Binary Pooling. *Appl. Sci.* **2023**, *13*, 10012. https://doi.org/10.3390/app131810012

Academic Editor: Andrea Prati

Received: 7 August 2023
Revised: 30 August 2023
Accepted: 30 August 2023
Published: 5 September 2023

machine learning methods. Since then, several well-established CNN models, including VGGNet [8], GoogLeNet [9], ResNet [10], and DenNet [11], have shown impressive performance on various image classification tasks. However, these models are computationally expensive and require significant memory resources, making them challenging to deploy on mobile devices. To address this issue, researchers have developed lightweight models such as SqueezedNet [12], MobileNet [13], ShuffleNet [14], and EfficientNet [15], which are designed to be easily deployable on mobile devices.

There are a lot of techniques for ameliorating the model performance, which encompass enhancements of the model structure, augmentation of the training data, and adoption of transfer learning approaches. The success achieved by the Vision Transformer [16] has brought attention to the utility of attention mechanisms in facilitating feature extraction. The attention mechanism is a technique for selectively attending to the most informative regions in an image while disregarding irrelevant regions. This approach is inspired by the human visual system's ability to efficiently analyze and comprehend complex scenes through selective attention. The potential of this mechanism has led researchers to explore its application in computer vision systems to enhance their performance. Specifically, the attention mechanism can be viewed as a dynamic selection process that adaptively weights input features according to their importance, which enables the network to effectively focus on the most informative and relevant aspects of the input. The beginning of attention mechanisms in computer vision is the Recurrent Attention Model (RAM) [17], which combined deep neural networks with attention mechanisms to recurrently predict important regions in an image and update the entire network. The initial stage heavily relied on recurrent neural networks (RNNs) for implementing the attention mechanism. Subsequently, the advent of the SENet [18] marked the beginning of a new stage, which introduced a channel attention network. ECANet [19] and Convolutional Block Attention Module (CBAM) [20] were representative works in this phase. Finally, the concept of self-attention mechanism was initially introduced by Transformer [21] in natural language processing and then successfully applied to computer vision by Vision Transformer. Series of models based on Vision Transformers such as Tokens-to-Token Vision Transformer [22], Pyramid Vision Transformer [23], Swin Transformer [24], Convolutional Vision Transformer [25], Vision Outlooker [26], and CoAtNet [27] demonstrate the immense potential of attention mechanisms.

GAP has played an important role in previous attention mechanisms. However, it suffers from significant information loss as it fails to preserve the spatial structure and positional information in the feature maps. Consequently, it cannot capture subtle differences and local structures in different regions of the image. Additionally, since GAP uniformly processes the entire image, it fails to differentiate between important and unimportant regions in the image. This results in the model assigning equal importance to features from all regions, without focusing on the areas that are more relevant to the classification task. Moreover, GAP leads to a single spatially invariant global feature representation lacking diversity. This limitation prevents the model from capturing multiple local feature patterns or information at different scales in the image, thus restricting the expressive capability of the model.

The proposed attention mechanism no longer relies solely on GAP, but instead combines GAP with GMP in a binary pooling approach. GMP is utilized to extract the most salient features from the image or feature maps. By performing max pooling over the entire feature map, only the most important features in each channel are retained, while suppressing less significant features. This helps reduce redundancy and highlight crucial elements in the image, resulting in more discriminative features being extracted. Furthermore, channel attention is incorporated to enhance the extraction of image feature vectors. Channel attention is implemented by applying fully connected layers and point convolutions after the pooling operation. This combination allows the model to better focus on important features in each channel, facilitating improved feature representation.

Our work contributions can be summarized as follows:

- To address the issue of insufficient information in GAP, a binary pooling operation is proposed. This approach effectively enhances the representational capacity of the convolutional network.
- The proposed channel attention mechanism leads to improved image feature representation.
- Experimental results demonstrate that the proposed method outperforms other attention mechanisms, yielding superior performance on ImageNet.

The rest of the paper is structured as follows. Section 2 presents a review of related work. Section 3 illustrates how the model is constructed. In Section 4, comprehensive experiments are conducted on ImageNet to evaluate the effectiveness of the proposed method, which achieves excellent results. Finally, this work is summarized in Section 5.

2. Related Work

2.1. Pooling

Pooling operations are commonly used in convolutional neural networks to make models more robust to variations in feature positions within input images. The most common pooling methods are max pooling and average pooling. However, there are several other pooling methods in convolutional neural networks. GAP [28] helps prevent overfitting and serves as an alternative to fully connected layers, removing the black-box nature of features in fully connected layers and providing each channel with a meaningful category interpretation. Inspired by dropout, which randomly sets some activation functions to zero during model training, mix pooling [29] combines max pooling and average pooling randomly during training. Unlike max pooling, which always selects the maximum element, stochastic pooling [30] randomly selects elements from the feature map based on their probability values, giving it stronger generalization capabilities. Power average pooling [31] combines average pooling and max pooling by using a learnable parameter to determine the relative importance of these two methods. Local importance pooling [32] learns adaptive and discriminative feature maps to aggregate downsampling features while discarding uninformative features, thereby preserving image details and being particularly useful for tasks with exceptionally rich detailed information. Soft pooling [33], based on softmax weighting, aims to minimize information loss during the pooling process while preserving information features and improving the classification performance of convolutional neural networks. In contrast, the proposed binary pooling is a pooling operation specifically tailored for attention, with the aim of enhancing the representation capacity of the fundamental modules throughout the entire network.

2.2. Attention Mechanism

Attention can be regarded as a mechanism that strategically allocates computational resources to the most salient components in the input image, ensuring an enhanced focus on the most informative features. SENet [18] introduced a novel squeeze-and-excitation channel attention module that captures the correlation between convolutional feature channels. ECANet [19], an improved version of SENet, replaces fully connected layers with cheaper 1D convolutions. Inspired by SENet, GENet [34] captures remote spatial context information by providing recalibration function in spatial domain, in which the lightweight gather-and-excitation module can be inserted into each residual block like SE. SKNet [35] employs a module composed of Split, Fuse, and Select to adaptively process the output with different parameter weights and receptive fields for different inputs. CBAM [20] and Bottleneck attention module (BAM) [36] focus on the dual attention mechanism, considering both space and channel. Additionally, it should be pointed out that CBAM also uses GAP and GMP but the distinction from our proposed binary pooling is they concatenate the pooling outputs along specified dimensions. Style-based recalibration module (SRM) [37] first extracts style information using style pooling that combines GAP and standard pooling, then allocates attention weights through fully connected layers. Global attention mechanism (GAM) [38] proposes an attention mechanism that can exploit essential features in all

three dimensions (channel, spatial width, and spatial height) to enhance cross-dimensional interactions. In contrast, coordinate attention (CA) [39] captures remote dependencies along one spatial direction while preserving precise location information along another. It encodes feature maps as orientation-aware and position-sensitive attention maps, which are then applied to the input feature maps to enhance the representation of objects of interest. FcaNet [40] designed a unique pooling based on frequency, which can simultaneously pay attention to low-frequency and high-frequency information in images.

3. System Design

In this section, the attention mechanism module put forward in the study is delineated, as depicted in Figure 1. Initially, a binary pooling approach that merges GAP and GMP is introduced. This innovative method empowers the model to adeptly extract image features. Following this, a pair of fully connected layers undertakes the processing of the pooled outcomes. Ultimately, a pointwise convolution is used to amplify the consolidation of channel-related information. Through the employment of this module, a notable augmentation in the model's feature extraction prowess is achieved, consequently leading to heightened accuracy levels.

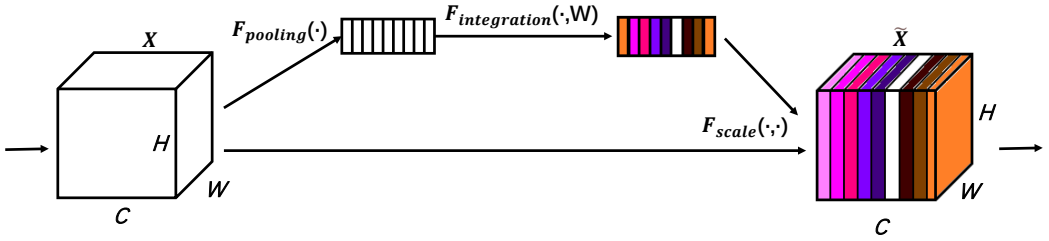

Figure 1. An attention block. C represents the number of channels in the image, H represents the vertical dimension of the image in pixels, W represents the horizontal dimension of the image in pixels, X represents the input, $\tilde{X}$ represents the output, $F_{pooling}(\cdot)$ is binary pooling operation, $F_{integration}(\cdot, W)$ is integration operation and $F_{scale}(\cdot,\cdot)$ is weighting operation.

3.1. Binary Pooling

The idea of binary pooling originates from the principle of "Zhong Yong" in Confucianism, which emphasizes the pursuit of balance and harmony in all ways. The formula for binary pooling is as follow:

$$y_k = F_{pooling}\left(x_{kpq}\right) = 0.5 \times \left(\frac{1}{|\mathcal{R}|} \sum_{(\mathrm{p,q}) \in \mathcal{R}} x_{kpq} + \max_{(p,q) \in \mathcal{R}} x_{kpq} \right) \tag{1}$$

where y_k represents the binary pooling output of the k^{th} feature map, x_{kpq} represents the element located at position (p, q) in the region $\mathcal{R}$ of the k^{th} feature map, $|\mathcal{R}|$ represents the total number of elements in the k^{th} feature map.

Firstly, binary pooling allows the model to capture both average and maximum features within each channel of the feature map. GAP calculates the average, providing a measure of overall importance for each feature, while GMP captures the maximum value, representing the most prominent feature in each channel. Secondly, it also helps in extracting diverse and discriminative features from the input. While GAP emphasizes global features, reducing the impact of noisy or less important features, GMP focuses on capturing the most unique and information-rich features. Therefore, binary pooling helps in capturing both global and local information, resulting in more robust and expressive representations.

3.2. Integration

In Contribution 2, we propose a channel attention that enhances the representation of image features through channel-related integration. The proposed channel attention mechanism is carried out by the integration operations. To make fully use the information obtained after pooling, an integration operations is followed with the aim of fully capturing the channel-related dependencies and appropriately enhancing or attenuating them. This objective is accomplished through the following formula:

$$z = F_{integration}(W, y) = \varphi(\lambda(\varphi(f(W, y)))) = \varphi(\lambda(\varphi(\gamma(W_2(W_1 y)))))) \quad (2)$$

where φ refers to the sigmoid function, λ refers to the pointwise convolution, γ refers to the ReLU function, $W_1 \in \mathbb{R}^{Cr \times C}$ and $W_2 \in \mathbb{R}^{C \times Cr}$. To enhance the flexibility of the attention block, two fully connected layers and a pointwise convolution are adopted. Firstly, there is an upscaling layer with parameters W_1 with expansion rate r, a ReLU and then a downsizing layer with parameters W_2 and activated by a sigmoid. Next, a pointwise convolution is applied to aggregate the features, followed by a sigmoid. It is worth mentioning that point convolution can be used to fuse features from different channels. By performing convolution operations on corresponding elements of different channels at the same position, information from different channels can interact and combine with each other, thereby extracting more diverse feature representations.

3.3. Scale

The final output of the attention block is achieved through the following formula:

$$\tilde{X} = F_{scale}(z, \mathbf{x}) = z \cdot \mathbf{x} \quad (3)$$

where $\mathbf{x} \in \mathbb{R}^{H \times W}$ refers to the input feature vector and z refers to the weights. The final operation is to perform matrix multiplication between the weight matrix and $\mathbf{x}$.

3.4. Example

As a plug-and-play attention mechanism module, it is highly flexible. One example based on ResNet is shown in Figure 2. The attention module is the non-identity branch of the residual module, and it operates before the summation with the identity branch. This approach allows the module to be integrated into other advanced backbone networks, enhancing the effectiveness of the model.

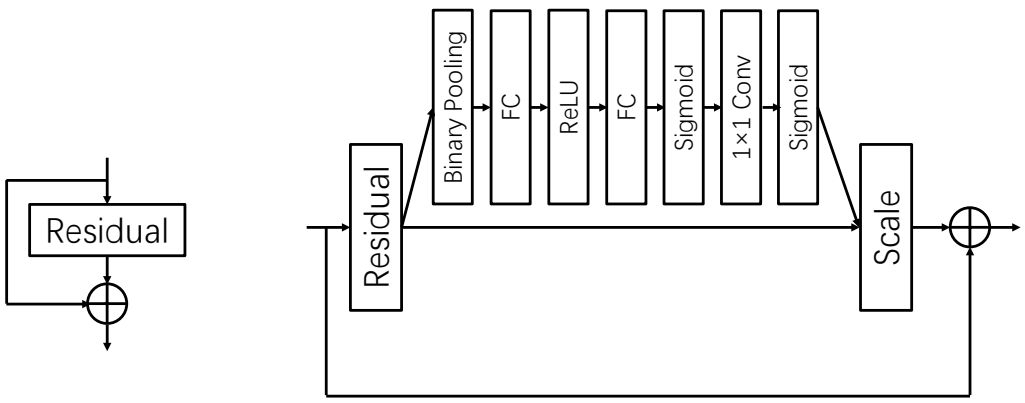

Figure 2. The design of the original residual module (**left**) and the ResNet module with attention mechanism (**right**).

4. Experiments

In this section, the experimental setup is first described and the effectiveness of the proposed method in image classification tasks is studied. Then, the ablation research of the model is introduced. Finally, the data of the model are visualized.

4.1. ImageNet Classification

The evaluation of the proposed method is conducted on the ImageNet [41] classification dataset which consists of 1000 classes. The models are trained on the 1.28 million training images, and evaluated on the 50k test images. For the training set, a 224×224 crop is randomly sampled from an image or its horizontal flip, with the per-pixel mean subtracted. For the test set, images are resized to 256×256 and cropped from the center of the picture to 224×224.

Setup: To evaluate the performance of the proposed attention block on ImageNet, two popular CNN architectures are used as backbone networks: ResNet18 and ResNet50. Regarding the hyperparameters, in ResNet18, the learning rate was initialized as 0.1 and decayed by 1/10 each 30 epochs. An SGD optimizer with a weight decay of 1×10^{-4}, a momentum of 0.9, and a batch size of 256 was used. In ResNet50, the learning rate was initialized as 0.05 and decayed by 1/10 each 30 epochs. An SGD optimizer with a weight decay of 1×10^{-4}, a momentum of 0.9, and a batch size of 128 was used.

The proposed method is also compared with other prominent networks with attention, including SENet [18], CBAM [20], SRM [37], FcaNet [40], and ECANet [19]. Evaluation metrics included parameters, floating point operations per second (FLOPs), and top-1/top-5 accuracy. No augmentation techniques such as mixup [42], cutout [43], cutmix [44], etc. or label regularization such as label smoothing [45] are adopted in the implementation. All networks have been trained for 100 epochs on NVIDIA 3090 GPU, Intel i5-12400F CPU, and Pytorch framework.

Comparison of results on ResNet: Figure 3 shows the training curves after inserting the proposed attention blocks into ResNet18 and ResNet50. At the beginning of the training, due to the model's lack of exposure to effective representations in the data, both the training accuracy and the test accuracy are low, and the training and test losses are high. As training progresses, the training accuracy and test accuracy gradually improve, while the training and test losses decrease. This indicates that the model is learning to extract features and patterns from the data and achieving better fit to the training data. Starting from the 60th epoch, the improvement in both training accuracy and test accuracy tends to approach a smaller value, and the training and test losses stabilize. This suggests that the model has learned the general features present in the data and performs well on the test data.

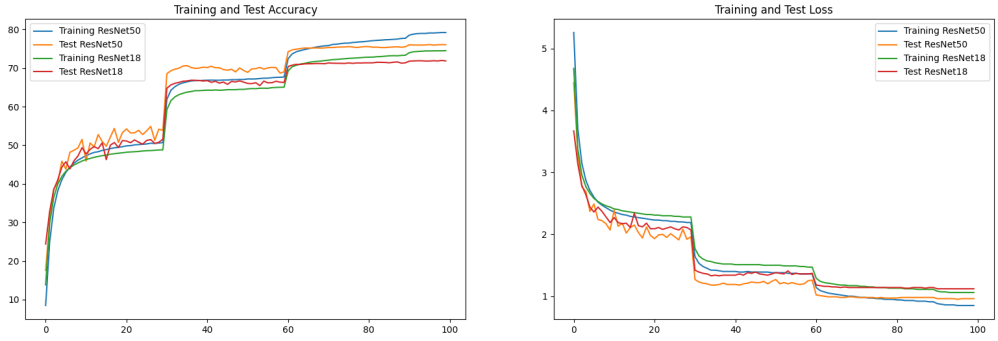

Figure 3. Training process of model on ImageNet. (**Left**) is the TOP-1 accuracy of the model on ImageNet; (**right**) is the CrossEntropy loss of the model on ImageNet.

Table 1 shows the comparison of results on the ResNet18 and ResNet50 backbones. The following observations can be obtained: (1) On the ResNet18 backbone, the proposed method achieves a higher top-1 accuracy compared to all other models, resulting in a 2.02% improvement over the baseline model. (2) However, on the larger ResNet50 backbone, the

proposed method significantly increases the parameter count due to the dimensionality expansion operation in the fully connected layer. Despite the limited increase in computational cost, the proposed method still achieves a higher top-1 accuracy compared to all other models, resulting in a 0.63% improvement over the baseline model.

Taking a comparison between ResNet-18 and ResNet-18 with the proposed attention block, for image inputs of 224 × 224, ResNet-18 requires approximately 1.82 GFLOPs. Our attention block employs binary pooling in the pooling stage, two fully connected layers and a pointwise convolution in the aggregation stage, and finally, an inexpensive scaling operation. Overall, when setting the expansion ratio r to 8, the new ResNet-18 requires approximately 1.83 GFLOPs, which represents a relative increase of 0.55% compared to the original ResNet-18. In exchange for this slight additional computational burden, the new ResNet-18 achieves higher accuracy than ResNet-18. In addition, when the mini-batch size is 256, training a batch on ResNet-18 takes 216 ms, while on the new ResNet-18, it takes 240 ms. We believe this represents a reasonable time overhead, which could potentially be further reduced if pooling and small internal product operations are optimized in PyTorch. Considering its contribution to the model's performance, the minor additional computational cost generated by the attention block is acceptable.

Table 1. Comparison of results on ImageNet test set. All results are the average of five training runs in the same environment.

Model	Backbone	Params (M)	FLOPs (G)	Top-1 Acc (%)	Top-5 Acc (%)
ResNet	ResNet-18	11.69	1.82	70.25	89.38
SENet		11.78	1.82	70.98	90.03
CBAM		11.78	1.82	71.01	89.85
SRM		11.70	1.82	71.23	90.16
ECANet		11.69	1.82	70.60	89.68
FcaNet		11.78	1.82	71.11	90.10
Our		23.53	1.83	72.27	90.62
ResNet	ResNet-50	25.56	4.11	75.91	92.86
SENet		28.07	4.12	76.31	93.33
CBAM		28.07	4.12	76.46	93.49
SRM		25.62	4.11	76.54	93.47
ECANet		25.56	4.12	76.34	93.44
FcaNet		28.07	4.12	76.68	93.54
Our		367.60	4.46	76.94	93.52

4.2. CIFAR-10 and CIFAR-100

Experiments are also conducted on two classic small datasets, CIFAR-10 and CIFAR-100 [46], which consist of sets of 50,000 training and 10,000 test RGB images of size 32 × 32 pixels. These datasets are labeled with 10 and 100 classes, respectively. The attention blocks were integrated into ResNet-18 and ResNet-50 architectures. During training, the images were randomly flipped horizontally and zero-padded with four pixels on each side before undergoing random 32 × 32 cropping. Mean and standard deviation normalization was also applied. The training hyperparameters, such as batch size, initial learning rate, and weight decay, were set following the recommendations in the original paper. The performance on CIFAR-10 and CIFAR-100 is shown in Tables 2 and 3. It can be seen that in each table, the new ResNet outperformed the baseline architectures, indicating that the benefits of the proposed attention block are not limited to the ImageNet dataset.

Table 2. Classification acc (%) on CIFAR-10. All results are the average of five training runs in the same environment.

	Original	Our
ResNet-18	90.25	92.54
ResNet-50	92.96	94.63

Table 3. Classification acc (%) on CIFAR-100. All results are the average of five training runs in the same environment.

	Original	Our
ResNet-18	60.74	63.41
ResNet-50	74.15	76.53

4.3. Ablation Studies

4.3.1. Expansion Rate

The expansion rate r in the fully connected layer is an important hyperparameter that affects the capacity and computational cost of the attention block. To study its impact, experiments with various values of r are performed on ResNet-18. The comparison table in Table 4 suggests that the performance does not monotonically improve with increasing capacity. This is likely due to the channel interdependencies in the attention block that can lead to overfitting on the data set. Specifically, it is found that setting $r = 8$ achieved a good balance between accuracy and complexity. Therefore, this value was used in all experiments.

Table 4. Top-1/top-5 acc (%) on ImageNet test set and params(M) for attention block on ResNet-18 at different expansion ratios r. All results are the average of five training runs in the same environment.

Ratio r	Top-1 Acc (%)	Top-5 Acc (%)	Params (M)
2	72.05	90.68	15.17
4	72.09	90.65	17.96
8	72.27	90.62	23.53
16	72.17	90.62	34.67

4.3.2. Activation Function

The activation function is a nonlinear function in neural networks that introduces nonlinear transformations to increase the expressive power and fitting capacity of the network. The activation function processes the input of a neuron and generates an output signal, which serves as the input for the next layer of neurons. Commonly used activation functions include Sigmoid, ReLU, and Tanh. Different activation functions have different effects on attention block, as shown in the Table 5. Compared to using Sigmoid, using Tanh leads to a slight decrease in accuracy, while using ReLU results in a significant decrease. Therefore, the final selection is to use the Sigmoid activation function.

Table 5. Effect of using different activation function for the attention block in ResNet-18 on ImageNet. All results are the average of five training runs in the same environment.

Function	Top-1 Acc (%)	Top-5 Acc (%)
Sigmoid	72.27	90.62
ReLU	71.08	90.06
Tanh	71.55	90.27

4.4. Data Visualization

It is worth mentioning that our work employs a novel binary pooling operation, as well as fully connected and point convolution layers, to obtain attention weights, which guarantees comprehensive feature vector extraction from images. In order to show this powerful attention mechanism more intuitively, it is necessary to focus on the learned attention weight. By using Grad-CAM++ [47], the attention weight of pre-trained model on ImageNet training set can be seen in Figure 4. It should be noted that the redder the color is, the greater the attention weight is. It can also be concluded that the red is covered on the target, which means that attention mechanism used in this work is really on the target.

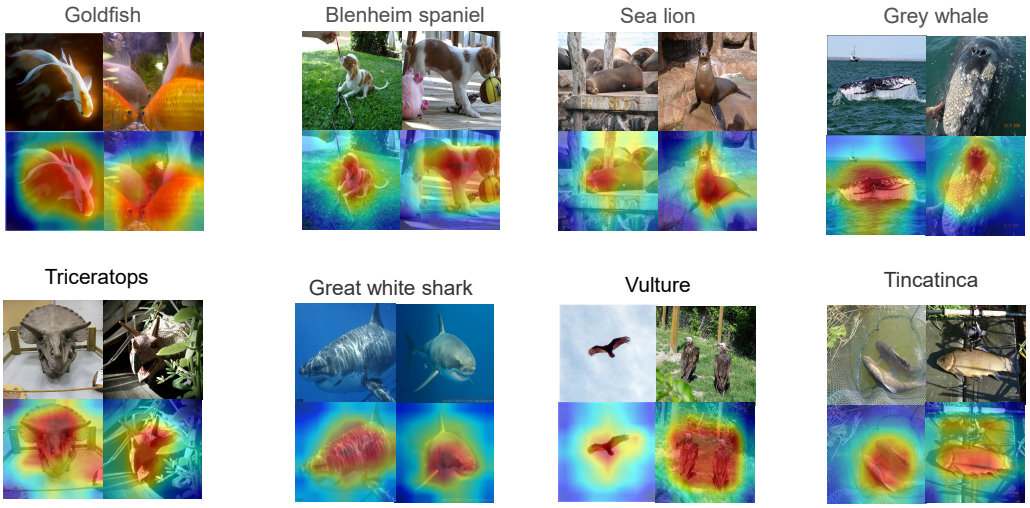

Figure 4. Attention weight distribution map. The redder the color, the higher the attention weight. These pictures well prove that using proposed attention block can bring good accuracy to the model.

5. Conclusions

This paper proposes a novel attention block for image classification. Compared to previous research on attention, our work utilizes more advanced pooling operations. Additionally, to enhance the extraction capability of feature vectors, fully connected and point convolution layers are adopted to aggregate feature maps maximally. Experiments on the ImageNet demonstrate the effectiveness and superior performance of the proposed model. The performance of the model is not limited to ImageNet alone. It is believed to exhibit promising outcomes in medical applications such as disease diagnosis, organ segmentation, and skin anomaly detection. Due to limitations in acquiring datasets, the efficacy of this model in medical imaging will be our focus in future research. However, there is still room for further research. For example, to avoid information loss and noise amplification, the dimensionality of operation is roughly chosen, which significantly increased the number of parameters in the attention block, although the increased computational complexity is acceptable. There are considerable optimization techniques in the fully connected layers. Furthermore, our proposed attention block is channel-based, but incorporating spatial attention may yield even better results. Exploring spatial attention mechanisms is thus an area for future research.

Author Contributions: Conceptualization, C.C.; methodology, C.C.; software, C.C.; validation, C.C.; formal analysis, C.C.; investigation, C.C.; resources, H.Z.; data curation, C.C.; writing—original draft preparation, C.C.; writing—review and editing, C.C. and H.Z.; visualization, C.C.; supervision, H.Z.; project administration, C.C.; funding acquisition, H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grant U1809206.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Datasets used in this work: ImageNet https://image-net.org/ and CIFAR10/100 http://www.cs.toronto.edu/~kriz/cifar.html (accessed on 20 August 2023).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolution Neural Network
GAP	Global Average Pooling
GMP	Global Max Pooling
RAM	Recurrent Attention Model
RNN	Recurrent Neural Network
CBAM	Convolutional Block Attention Module
BAM	Bottleneck attention module
SRM	Style-based recalibration module
GAM	Global attention mechanism
CA	Coordinate attention

References

1. Kufel, J.; Bargieł-Łączek, K.; Kocot, S.; Koźlik, M.; Bartnikowska, W.; Janik, M.; Czogalik, Ł.; Dudek, P.; Magiera, M.; Lis, A.; et al. What Is Machine Learning, Artificial Neural Networks and Deep Learning?—Examples of Practical Applications in Medicine. *Diagnostics* **2023**, *13*, 2582. [CrossRef] [PubMed]
2. Vandewinckele, L.; Claessens, M.; Dinkla, A.; Brouwer, C.; Crijns, W.; Verellen, D.; van Elmpt, W. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. *Radiother. Oncol.* **2020**, *153*, 55–66. [CrossRef] [PubMed]
3. Li, G.; Wu, X.; Ma, X. Artificial intelligence in radiotherapy. *Semin. Cancer Biol.* **2022**, *86 Pt 2*, 160–171. [CrossRef] [PubMed]
4. Kufel, J.; Bargieł, K.; Koźlik, M.; Czogalik, Ł.; Dudek, P.; Jaworski, A.; Cebula, M.; Gruszczyńska, K. Application of artificial intelligence in diagnosing COVID-19 disease symptoms on chest X-rays: A systematic review. *Int. J. Med. Sci.* **2022**, *19*, 1743. [CrossRef]
5. Rangarajan, K.; Muku, S.; Garg, A.K.; Gabra, P.; Shankar, S.H.; Nischal, N.; Soni, K.D.; Bhalla, A.S.; Mohan, A.; Tiwari, P.; et al. Artificial Intelligence–assisted chest X-ray assessment scheme for COVID-19. *Eur. Radiol.* **2021**, *31*, 6039–6048. [CrossRef]
6. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
7. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
8. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
9. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
10. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
11. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
12. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. *arXiv* **2016**, arXiv:1602.07360.
13. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
14. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
15. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning (PMLR), Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
16. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
17. Mnih, V.; Heess, N.; Graves, A. Recurrent models of visual attention. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Cambridge, MA, USA, 8–13 December 2014; Volume 27.

18. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
19. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
20. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
22. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.H.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 558–567.
23. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 568–578.
24. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
25. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 22–31.
26. Yuan, L.; Hou, Q.; Jiang, Z.; Feng, J.; Yan, S. Volo: Vision outlooker for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 6575–6586. [CrossRef] [PubMed]
27. Dai, Z.; Liu, H.; Le, Q.V.; Tan, M. Coatnet: Marrying convolution and attention for all data sizes. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 3965–3977.
28. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv* **2013**, arXiv:1312.4400.
29. Yu, D.; Wang, H.; Chen, P.; Wei, Z. Mixed pooling for convolutional neural networks. In Proceedings of the Rough Sets and Knowledge Technology: 9th International Conference (RSKT 2014), Shanghai, China, 24–26 October 2014; Proceedings 9; Springer: Berlin/Heidelberg, Germany, 2014; pp. 364–375.
30. Zeiler, M.D.; Fergus, R. Stochastic pooling for regularization of deep convolutional neural networks. *arXiv* **2013**, arXiv:1301.3557.
31. Estrach, J.B.; Szlam, A.; LeCun, Y. Signal recovery from pooling representations. In Proceedings of the International Conference on Machine Learning (PMLR), Beijing, China, 21–26 June 2014; pp. 307–315.
32. Gao, Z.; Wang, L.; Wu, G. Lip: Local importance-based pooling. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 15–20 June 2019; pp. 3355–3364.
33. Stergiou, A.; Poppe, R.; Kalliatakis, G. Refining activation downsampling with SoftPool. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10357–10366.
34. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Vedaldi, A. Gather-excite: Exploiting feature context in convolutional neural networks. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, QC, Canada, 3–8 December 2018; Volume 31.
35. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 510–519.
36. Park, J.; Woo, S.; Lee, J.Y.; Kweon, I.S. Bam: Bottleneck attention module. *arXiv* **2018**, arXiv:1807.06514.
37. Lee, H.; Kim, H.E.; Nam, H. Srm: A style-based recalibration module for convolutional neural networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 15–20 June 2019; pp. 1854–1862.
38. Liu, Y.; Shao, Z.; Hoffmann, N. Global attention mechanism: Retain information to enhance channel-spatial interactions. *arXiv* **2021**, arXiv:2112.05561.
39. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Montreal, BC, Canada, 11–17 October 2021, pp. 13713–13722.
40. Qin, Z.; Zhang, P.; Wu, F.; Li, X. Fcanet: Frequency channel attention networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 783–792.
41. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
42. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.
43. DeVries, T.; Taylor, G.W. Improved regularization of convolutional neural networks with cutout. *arXiv* **2017**, arXiv:1708.04552.
44. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 15–20 June 2019; pp. 6023–6032.

45. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
46. Krizhevsky, A.; Hinton, G. Learning Multiple Layers of Features from Tiny Images. Technical Report. 2009. Available online: http://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf (accessed on 6 August 2023).
47. Chattopadhyay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V. Grad-CAM++: Improved visual explanations for deep convolutional networks, arXiv. *arXiv* **2018**, arXiv:1710.11063.