Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions

Yichao Liu
Helmholtz-Zentrum Dresden-Rossendorf
Dresden, Germany
y.liu@hzdr.de

Zongru Shao
Helmholtz-Zentrum Dresden-Rossendorf
Dresden, Germany
Center for Advanced Systems Understanding
Görlitz, Germany
z.shao@hzdr.de

Nico Hoffmann
Helmholtz-Zentrum Dresden-Rossendorf
Dresden, Germany
n.hoffmann@hzdr.de

Abstract

A variety of attention mechanisms have been studied to improve the performance of various computer vision tasks. However, the prior methods overlooked the significance of retaining the information on both channel and spatial aspects to enhance the cross-dimension interactions. Therefore, we propose a global attention mechanism that boosts the performance of deep neural networks by reducing information reduction and magnifying the global interactive representations. We introduce 3D-permutation with multilayer-perceptron for channel attention alongside a convolutional spatial attention submodule. The evaluation of the proposed mechanism for the image classification task on CIFAR-100 and ImageNet-1K indicates that our method stably outperforms several recent attention mechanisms with both ResNet and lightweight MobileNet.

1 Introduction

Convolutional neural networks (CNNs) have been widely used in many tasks and applications in the computer vision domain (Girshick et al. [2014], Long et al. [2015], He et al. [2016], Lampert et al. [2009]). Researchers have found that CNNs are performing well in extracting deep visual representations. With technological improvements related to CNNs, image classification on the ImageNet dataset (Deng et al. [2009]) has increased from 63% to 90% accuracy in the past nine years (Krizhevsky et al. [2012], Zhai et al. [2021]). Such an achievement also attributes to the complexity of the ImageNet dataset, which offers exceptional opportunities for related studies. Given the diversity and large scale of real-life scenes it covers, it has been benefitting studies for conventional image classification benchmarking, representation learning, transfer learning, etc. Particularly, it also brings challenges for the attention mechanisms.

The attention mechanisms have been improving performance in multiple applications and attracted research interests in recent years (Niu et al. [2021]). Wang et al. [2017] used an encoder-decoder residual attention module to refine the feature maps to obtain better performance. Hu et al. [2018], Woo et al. [2018], Park et al. [2018] used spatial and channel attention mechanisms separately and achieved a higher accuracy. However, these mechanisms utilize visual representations from limited receptive fields due to information reduction and dimension separation. They lose global spatial-channel interactions in the process. Our research objective is to investigate attention mechanisms

Preprint. Under review.

across the spatial-channel dimensions. We propose a "global" attention mechanism that reserves information to magnify the "global" cross-dimension interactions. Therefore, we name the proposed method Global Attention Mechanism (GAM).

2 Related Works

There have been several studies focusing on performance improvements of attention mechanisms for image classification tasks. Squeeze-and-Excitation Networks (SENet) (Hu et al. [2018]) is the first to use channel attention and channel-wise-feature-fusion to suppress the unimportant channels. However, it is less efficient in suppressing unimportant pixels. The later-on attention mechanisms considered both spatial and channel dimensions. The convolutional block attention module (CBAM) (Woo et al. [2018]) places the channel and spatial attention operation sequentially, while bottleneck attention module (BAM) (Park et al. [2018]) did it in parallel. However, both of them ignore the channel-spatial interactions and lose the cross-dimension information consequently. Considering the significance of the cross-dimension interactions, the triplet attention module (TAM) (Misra et al. [2021]) boosts efficiency by utilizing the attention weights between each pair of the three dimensions – channel, spatial width, and spatial height. However, the attention operations are still applied on two of the dimensions each time instead of all three. To magnify cross-dimension interactions, we propose an attention mechanism that is capable of capturing significant features across all three dimensions.

3 Global Attention Mechanism (GAM)

Our objective is to design a mechanism that reduces information reduction and magnifies global dimension-interactive features. We adopt the sequential channel-spatial attention mechanism from CBAM and redesign the submodules. The overall process is illustrated in Fig. 1 and formulated in Equation 1 and 2 (Woo et al. [2018]). Given the input feature map $\mathbf{F_1} \in \mathbb{R}^{C\times H\times W}$, the intermediate state $\mathbf{F_2}$ and the output $\mathbf{F_3}$ are defined as:

$$\mathbf{F}_2 = \mathbf{M}_c(\mathbf{F}_1) \otimes \mathbf{F}_1 \tag{1}$$
$$\mathbf{F}_3 = \mathbf{M}_s(\mathbf{F}_2) \otimes \mathbf{F}_2 \tag{2}$$

where $\mathbf{M}_c$ and $\mathbf{M}_s$ are the channel and spatial attention maps, respectively; $\otimes$ denotes element-wise multiplication.

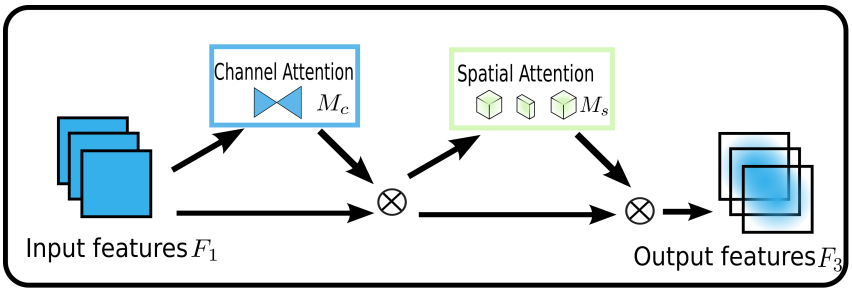

Figure 1: **The overview of GAM**

The **channel attention** submodule uses 3D permutation to retain information across three dimensions. It then magnifies cross-dimension channel-spatial dependencies with a two-layer MLP (multi-layer perceptron). (The MLP is an encoder-decoder structure with a reduction ratio r, same as BAM.) The channel attention submodule is illustrated in Fig. 2.

In the **spatial attention** submodule, to focus on spatial information, we use two convolutional layers for spatial information fusion. We also use the same reduction ratio r from the channel attention submodule, same as BAM. Meanwhile, max-pooling reduces the information and contributes negatively. We remove pooling to further retain the feature maps. As a result, the spatial attention module sometimes increase the number of parameters significantly. To prevent a notable increase of the parameters, we adopt group convolution with channel shuffle (Zhang et al. [2018]) in ResNet50. The spatial attention submodule without group convolution is shown in Fig. 3.

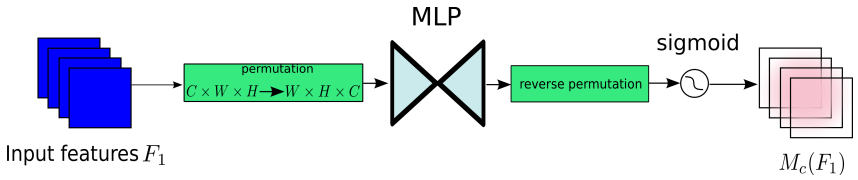

Figure 2: **Channel attention submodule.**

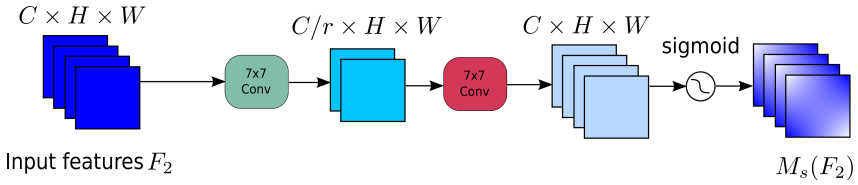

Figure 3: **Spatial attention submodule.**

4 Experiment

In this section, we evaluate GAM on both CIFAR-100 (Krizhevsky et al. [2009]) and ImageNet-1K datasets (Deng et al. [2009]) with classification benchmarking and two ablation studies. We use two datasets to verify method generalization. Note that both datasets are standard for classification. ImageNet-1K has a higher impact on real-life applications.

4.1 Classification on CIFAR-100 and ImageNet datasets

We evaluate GAM with both ResNet (He et al. [2016]) and MobileNet V2 (Sandler et al. [2018]) as (a) they are standard architectures for image classification (b) they represent the regular and the lightweight networks respectively. We compare GAM against SE, BAM, CBAM, TAM, and Attention Branch Network (ABN) (Fukui et al. [2019]). We re-implement the networks & mechanisms and evaluate them under the same conditions. All models are trained on four Nvidia Tesla V100 GPUs.

For CIFAR-100, we evaluate GAM with and without group convolution (gc). We train all networks for 200 epochs with a starting learning rate of 0.1. Then, we drop the learning rate at the epochs of 60, 120, and 160. The results are shown in Table 1. It shows that GAM outperforms SE, BAM, and CBAM.

Table 1: Classification results on Cifar100

Architecture	Parameters	FLOPs	Top-1 Error (%)	Top-5 Error (%)
ResNet 50	23.71M	1.3G	22.74	6.37
ResNet 50 + SE	26.22M	1.31G	20.29	5.18
ResNet 50 + BAM	24.06M	1.33G	19.97	5.03
ResNet 50 + CBAM	26.24M	1.31G	19.44	4.66
ResNet 50 + GAM	149.47M	8.02G	**18.67**	**4.54**
ResNet 50 + GAM (gc*)	57.05M	3.08G	18.99	4.87

⋆ gc stands for group convolution (we set its hyper-parameter as 4).

For ImageNet-1K, we pre-process the images to 224×224 (He et al. [2016]). We include both ResNet18 and ResNet50 (He et al. [2016]) to verify method generalization on different network depths. For ResNet50, we include a comparison with group convolution to prevent a notable increase of the parameters. We set the starting learning rate as 0.1 and drop it for every 30 epochs. We use 90 training epochs in total. In the spatial attention submodule, we switch the first stride of the first block from 1 to 2 in order to match the size of the features. Other settings are preserved from CBAM for a fair comparison, including the use of max-pooling in the spatial attention submodule.

MobileNet V2 is one of the most efficient lightweight models for image classification. We use the same setup of ResNet for MobileNet V2 except using an initial learning rate of 0.045 and a weight decay of 4×10^{-5}.

The evaluation on the ImageNet-1K is shown in Table 2. It shows that GAM stably enhances the performance across different neural architectures. Especially, for ResNet18, GAM outperforms ABN with fewer parameters and better efficiency.

Table 2: Classification results on ImageNet-1K

Architecture	Parameters	FLOPs	Top-1 Error (%)	Top-5 Error (%)
ResNet 18	11.69M	1.82G	30.91	11.12
ResNet 18 + SE	11.78M	1.82G	30.07	10.59
ResNet 18 + BAM	11.71M	1.82G	30.18	10.77
ResNet 18 + CBAM	11.78M	1.82G	29.89	10.53
ResNet 18 + TAM	11.69M	1.83G	30.0	10.64
ResNet 18 + ABN	21.61M	3.76G	29.4	10.34
ResNet 18 + GAM	16.04M	2.45G	**29.34**	**10.23**
ResNet 50	25.56M	4.11G	24.81	7.69
ResNet 50 + SE	28.07M	4.12G	23.56	6.82
ResNet 50 + BAM	25.92M	4.19G	24.0	7.01
ResNet 50 + CBAM	28.09M	4.12G	23.1	6.57
ResNet 50 + TAM	25.56M	4.16G	23.29	6.7
ResNet 50 + ABN	43.58M	7.64G	23.43	6.92
ResNet 50 + GAM	151.32M	24.66G	**22.78**	**6.43**
ResNet 50 + GAM (gc)	58.9M	9.56G	23.01	6.52
MobileNet V2	3.51M	0.31G	30.52	11.20
MobileNet V2 + SE	3.53M	0.32G	29.77	10.65
MobileNet V2 + BAM	3.54M	0.32G	29.91	10.80
MobileNet V2 + CBAM	3.54M	0.32G	29.74	10.66
MobileNet V2 + GAM	4.93M	0.47G	**29.31**	**10.43**

4.2 Ablation studies

We conduct two ablation studies on ImageNet-1K with ResNet18. We first evaluate the contributions of spatial and channel attention separately. Then, we compare GAM against CBAM with and without max-pooling.

To better understand the contribution of spatial and channel attention separately, we conduct the ablation study by turning one on and the other off. For example, ch indicates the spatial attention is switched off and the channel attention is on. sp indicates the channel attention is turned off and the spatial attention is on. The results are shown in Table 3. We could observe a boost of performance on both of the on-off experiments. It indicates that both spatial and channel attentions are contributing to the performance gain. Note that their combination improves the performance with a further step.

Table 3: Ablation studies on ImageNet

Architecture	Parameters	FLOPs	Top-1 Error (%)	Top-5 Error (%)
ResNet 18†	11.69M	1.82G	30.91	11.12
ResNet 18 + GAM (sp*)	15.95M	2.45G	29.61	10.41
ResNet 18 + GAM (ch*)	11.78M	1.83G	30.25	10.97
ResNet 18 + GAM (ch+sp)†	16.04M	2.45G	**29.34**	**10.23**

⋆ sp stands for spatial attention only. ch stands for channel attention only.
† same as Table 2.

It is possible for max-pooling to contribute negatively in spatial attention depends on the neural architecture (e.g., ResNets). Therefore, we conduct another ablation study that compares GAM

against CBAM with and without max-pooling for ResNet18. The results are shown in Table 4. It is observed that our method outperforms CBAM under both conditions.

Table 4: Ablation studies on ImageNet

Architecture	Parameters	FLOPs	Top-1 Error (%)	Top-5 Error (%)
ResNet 18 + CBAM†	11.78M	1.82G	29.89	10.53
ResNet 18 + GAM†	16.04M	2.46G	29.34	10.23
ResNet 18 + CBAM (wmp*)	11.78M	1.83G	29.44	10.24
ResNet 18 + GAM (wmp*)	16.05M	2.47G	**28.57**	**9.83**

⋆ wmp stands for without max pooling.
† same as Table 2.

5 Conclusion

In this work, we proposed GAM to magnify salient cross-dimension receptive regions. Our experimental results indicate that GAM stably improves the performance for CNNs with different architectures and depths.

CIFAR-100 and ImageNet-1K are benchmarked in our evaluation as proof of concept. They represent a scaling up with the number of classes and images. Therefore, our experiments imply that GAM is prone to data scaling capability and robustness. We consider the full ImageNet dataset serves better for applications in production. It is expensive for large-model training, especially the up-to-date top-tier solutions. Our evaluation with ResNet and MobileNet proves its feasibility on model scaling as well. We aim to investigate detailed scaling capability of GAM as the next step.

GAM obtains performance gain with an increase in the number of network parameters. In the future, we plan to investigate technologies that reduce the number of parameters for large networks, e.g., ResNet50, ResNet101, etc. Meanwhile, we also plan to explore other cross-dimension attention mechanisms that utilize parameter-reduction techniques.

References

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958. IEEE, 2009.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *arXiv preprint arXiv:2106.04560*, 2021.

Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021.

Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018.

Diganta Misra, Trikay Nalamada, Ajay Uppili Arasanipalai, and Qibin Hou. Rotate to attend: Convolutional triplet attention module. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3139–3148, 2021.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10705–10714, 2019.