Lightweight Vision Transformer with Spatial and Channel Enhanced Self-Attention

Jiahao Zheng[*1] Longqi Yang[*2] Yiying Li[2] Ke Yang[2†]
Zhiyuan Wang[2†] Jun Zhou[1†]

[1]University of Electronic Science and Technology of China, Chengdu, China
[2]Defense Innovation Institute, Academy of Military Sciences, Beijing, China

Abstract

Due to the large number of parameters and high computational complexity, Vision Transformer (ViT) is not suitable for deployment on mobile devices. As a result, the design of efficient vision transformer models has become the focus of many studies. In this paper, we introduce a novel technique called Spatial and Channel Enhanced Self-Attention (SCSA) for lightweight vision transformers. Specially, we utilize multi-head self-attention and convolutional attention in parallel to extract global spatial features and local spatial features, respectively. Subsequently, a fusion module based on channel attention effectively combines the extracted features from both global and local contexts. Based on SCSA, we introduce the Spatial and Channel enhanced Attention Transformer (SCAT). On the ImageNet-1k dataset, SCAT achieves a top-1 accuracy of 76.6% with approximately 4.9M parameters and 0.7G FLOPs, outperforming state-of-the-art Vision Transformer architectures when the number of parameters and FLOPs are similar.

1. Introduction

Recently, ViT [3] has achieved remarkable results on major computer vision tasks with the assistance of long-range spatial feature relations captured through Multi-Head Self-Attention (MHSA). However, the secondary complexity of MHSA demands substantial computational resources, leading to efforts to reduce its computational complexity. To reduce computational overhead, PVT [16, 17] uses downsampling of key and value to decrease the complexity of MHSA, while Swin-Transformer [5] reduces complexity by dividing multiple windows and performing MHSA computation within the windows.

However, the performance of these models drops dramatically when reduced to a size and computation suitable for the mobile devices. Therefore, there are many works devoted to designing a lightweight and efficient vision transformers [7, 8, 6, 11, 9, 1, 20, 13, 12]. Some works refer to the perception of the human visual system to study the extraction and fusion of local and global information [7, 8, 6]. MobileViT [7] combines MobileNetv2 [10] with transformer blocks to enhance the global representation capability of the network. EdgeViT [8] proposes a local-global-local block for local and global information aggregation. EdgeNeXt [6] adopts split depth-wise convolution and transposes attention to implicitly increase the receptive field and encode multi-scale features. They both use a serial structure to stack the convolutional and self-attention layers, which model one structure (local or global) at a time and might destroy previous local features when extracting global features, and vice versa. Therefore, we have adopted a parallel structure approach to extract both local and global features simultaneously.

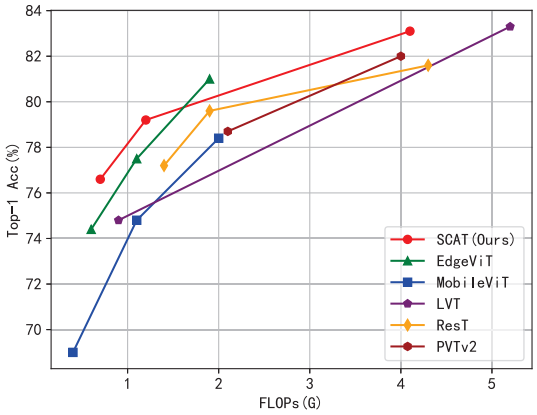

Figure 1: Top-1 accuracy v.s. FLOPs on ImageNet-1k of efficient models.

[*]Equal contributions. [†]Corresponding Authors: Ke Yang, Zhiyuan Wang, and Jun Zhou. This work was supported in part by the National Natural Science Foundation of China under Grants 62006241, and 62206307.

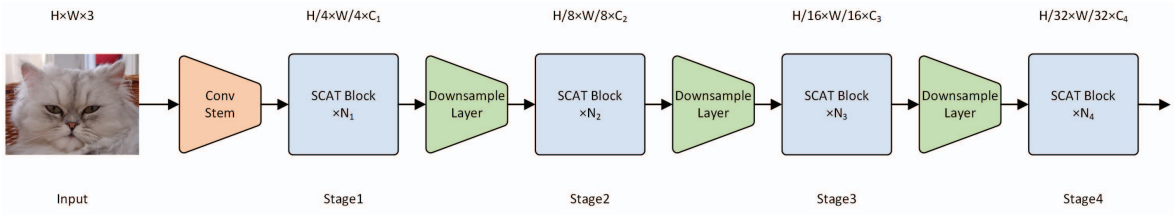

Figure 2: Architecture of our SCAT.

Before ViT [3] was proposed, there have been many attention-related works [4, 18, 15, 21]. SENet [4] introduces a channel attention module to highlight the important channels. It first compresses the feature map in the spatial dimension and then learns the importance of each channel in the channel dimension. The spatial attention module in CBAM [18] performs different pooling operations on the feature map in the channel dimension and then mixes the weights obtained from pooling to learn the importance of spatial locations. These works can be summarized as using the feature map to generate weights that act back on the feature map itself.

Based on the above analysis, we introduced the Spatial and Channel enhanced Self-Attention Block (SCSA). To be specific, we utilize MHSA to capture global long-range spatial features and employ convolutional attention to model local spatial features in parallel. Moreover, a channel attention-based fusion module is applied on top of the parallel global and local spatial attention block to learn their relationship and enhance the fusion of local and global features after concatenation. Furthermore, we propose a Convolutional Tokens Reduction (CTR) block to decrease the computational costs of MHSA by reducing the token length. Based on SCSA and CTR, following the common principles of lightweight transformer architecture design [8, 16], we propose the Spatial and Channel enhanced Attention Transformer (SCAT). Our main contributions are summarized as follows:

- We propose a Spatial and Channel enhanced Self-Attention (SCSA) mechanism that employs a two-branch architecture to efficiently extract local and global features and balances local and global features using channel attention.
- Our SCAT-XXS achieves a top-1 accuracy of 76.6% on ImageNet-1K with only 4.8M parameters and 0.7G FLOPs.

2. Method

2.1. Overview

The architecture of Spatial and Channel enhanced Attention Transformer (SCAT) is shown in Figure 2. We follow the same pyramid architecture as [16, 17], decreasing the resolution of the feature maps while increasing the number of channels of the feature maps during the forward propagation. First, we use the convolutional stem proposed in [19] to generate feature maps with a resolution of $H/4 \times W/4$, the convolutional stem consists of four 3×3 convolutions and one 1×1 convolution, where the stride of the first two 3×3 convolutions is 2 and the remaining is 1. Then we follow the 4-stage architecture adopted in [8, 6], where each stage consists of n SCAT blocks. Except for the first stage, the resolution of the feature map is reduced using non-overlapping large-step convolution before the other stages.

As shown in Figure 3a, the SCAT block is mainly composed of three parts: Conditional Position Encoding (CPE), Spatial and Channel enhanced Self-Attention (SCSA), and Feed-Forward Network (FFN). Our SCAT block can be formulated as:

$$\begin{aligned} X &= \boldsymbol{CPE}(X_{in}) + X_{in}, \\ Y &= \boldsymbol{SCSA}(\boldsymbol{Norm}(X)) + X, \\ X_{out} &= \boldsymbol{FFN}(\boldsymbol{Norm}(Y)) + Y. \end{aligned} \tag{1}$$

At first, the input tensor $X \in \mathcal{R}^{H \times W \times C}$ is embedded with the position information of tokens through CPE, which uses DWConv. Then SCSA extracts the fused and enhanced multi-scale features from both local and global branches, and finally the features are redistributed among channels by a classical feed-forward neural network.

2.2. Spatial and Channel enhanced Self-Attention

As shown in Figure 3b, Spatial and Channel enhanced Self-Attention (SCSA) consists of three parts: local branch, global branch and fusion module. The local branch extracts and reinforces local features, the global branch learns the global representation, and the fusion module further learns and fuses local and global features.

2.2.1 Global Branch

Inspired by PVT [16, 17], we use MHSA with resolution reduction of key and value, which can significantly reduce the computational complexity while still retaining the global receptive field. We propose the Convolutional Tokens Reduction (CTR) module to scale down the resolution of the fea-

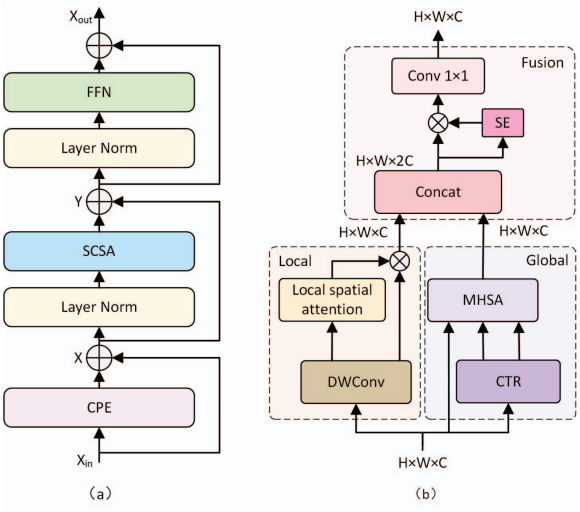

Figure 3: (a) Model architecture of our SCAT block. The SCAT block consists of Conditional Position Encoding (CPE), Spatial and Channel enhanced Self-Attention (SCSA) and Feed-Forward Network (FFN). (b) SCSA consists of three parts: local branch, global branch and fusion module.

ture map. CTR leverages a DWConv with a kernel size of $2k \times 2k$ and a stride of k, where the k is the reduction rate. The process can be formulated as:

$$\begin{aligned} X' &= \boldsymbol{CTR}(X), \\ Q &= W^Q X, \\ K', V' &= W^K X', W^V X', \\ X_{global} &= \boldsymbol{MHSA}(Q, K', V'), \end{aligned} \tag{2}$$

where the $X' \in \mathcal{R}^{\frac{H}{k} \times \frac{W}{k} \times C}$ is the resolution reduced feature map, k is reduction rate. W^Q, W^K, W^V are linear projection parameters.

2.2.2 Local Branch

Inspired by CBAM [18], we employ depth-wise convolution and local spatial attention to extract local features in local branch. Convolution with inductive bias can effectively extract local features, we further introduce local spatial attention to strengthen local features in spatial dimension. The details of local branching can be formulated as follows:

$$\begin{aligned} Q' &= \boldsymbol{DWConv}(Q), \\ W_{spatial} &= \sigma(\boldsymbol{Conv}([\boldsymbol{AvgPool}(Q'), \boldsymbol{MaxPool}(Q')]), \\ X_{local} &= Q' \odot W_{spatial}, \end{aligned} \tag{3}$$

where the σ denotes the sigmoid function and Conv represents a convolution operation with the kernel size of 7×7, the $\odot$ donates element-wise multiplication.

2.2.3 Fusion Module

In the fusion module, we concatenate the local and global features; then, we employ the channel attention to further learn the relationship between local and global features in the channel dimension. We use the SE module in [4] as a channel attention operation. We follow SENet and set the reduction rate in the SE module to 4. The fusion module can be formulated as follows:

$$\begin{aligned} W_{chanal} &= \boldsymbol{SE}([X_{local}, X_{global}]), \\ Y &= \boldsymbol{FC}([X_{local}, X_{global}] \odot W_{chanal}), \end{aligned} \tag{4}$$

where the $\odot$ donates element-wise multiplication. $[\cdot]$ is a concat operation.

3. Experiments

3.1. Data Set

We conduct the experiment on the ImageNet-1K dataset. ImageNet-1K [2] provides 1.28 million training images and $50,000$ validation images from 1000 categories. We report top-1 accuracy on the validation set for all experiments.

3.2. Implementation Details

We follow the training strategy in DeiT [14]. We use the AdamW optimizer to train the network, setting the batch size, initial learning rate, weight decay and momentum to 1024, 0.01, 0.05, and 0.9. Different from DeiT, we use a linear warm-up of 20 epochs. The maximum rates of increasing stochastic depth are set to $0.05/0.05/0.15$ for SCAT-XXS/XS/S. We used the same data augmentation in Swin-Transformer [5], including RandAugment, Mixup, CutMix, and Random Erasing.

In table 1, we present the specific parameter details of the three variants of SCAT. In order to save FLOPs, we used small convolutional kernels to capture low-level features in the early stages and large convolutional kernels to capture high-level features in the later stages.

3.3. Ablation Study

3.3.1 Local Spatial Attention

To verify the role of local spatial attention for local feature extraction and enhancement, we evaluated the performance of SCAT without local spatial attention. As shown in Table 2, the local spatial attention module [18] improved the accuracy of SCAT by 0.23% with almost no additional parameters and FLOPs. The results show that local spatial attention plays an important role in enhancing local features.

3.3.2 Convolutional Tokens Reduction

To comprehensively assess the CTR performance, we conduct a comparative analysis with three downsampling meth-

Model	Channals	Blocks	Heads	Kernel size	FLOPs(G)	Param(M)
SCAT-XXS	[32,80,160,256]	[2,2,5,2]	[2,5,10,16]	[3,5,7,9]	0.7	4.9
SCAT-XS	[48,96,192,384]	[2,2,5,2]	[3,6,12,24]	[3,5,7,9]	1.2	8.7
SCAT-S	[64,128,256,512]	[4,4,12,4]	[2,4,8,16]	[3,5,7,9]	4.1	31.0

Table 1: Configuration of three SCAT variants.

Model	Params (M)	FLOPs (G)	Top-1 (%)
SCAT w/o lsa	4.9	0.71	76.38
SCAT w/ lsa	4.9	0.71	76.61

Table 2: Ablation study of local spatial attention.

Model	Params (M)	FLOPs (G)	Top-1 (%)
sampling	4.8	0.70	76.08
mean pooling	4.8	0.70	76.54
conv w/o overlap	4.8	0.70	76.27
CTR	4.9	0.71	76.61

Table 3: Ablation study of CTR.

Model	Params (M)	FLOPs (G)	Top-1 (%)
SCAT-G w/o SE	4.3	0.70	75.68
SCAT-G w/ SE	4.4	0.70	75.95
SCAT w/o SE	4.3	0.70	75.82
SCAT w/ SE	4.9	0.71	76.61

Table 4: Ablation study of channel attention.

ods: sampling, mean pooling, and non-overlapping large-step convolution. As shown in Table 3, with minor differences in parameters and FLOPs, our CTR method outperforms the other three downsampling methods. This result suggests that our method might better preserve the integrity of information when downsampling tokens.

3.3.3 Fusion Module

To evaluate the effectiveness of the proposed fusion module, we carried out experiments on both full SCAT and SCAT without the local branch. The results of the experiment are shown in Table 4, SCAT-G denotes SCAT with only a global branch. In order to maintain SCAT-G and SCAT at the same FLOPs size thus reflecting the role of channel attention on feature fusion, we adjusted the model depth from $[2, 2, 5, 2]$ to $[2, 2, 6, 2]$. In SCAT-G, the addition of the SE module [4] only increased the accuracy by 0.27%, while in SCAT, the SE module increased the accuracy of the model by 0.79%. The experiments indicate that the SE module plays a great role in learning the relationship between local and global features.

3.4. Compare with State-Of-The-Art

We compare our SCAT against many state-of-the-art models in Table 5. The comparison results show that our

Model	Input	Params (M)	FLOPs (G)	Top-1 (%)
MobileViT-XXS [7]	256^2	1.3	0.4	69.0
EdgeViT-XXS [8]	224^2	4.1	0.6	74.4
LVT [20]	224^2	5.5	0.9	74.8
EdgeNeXt-XS [6]	256^2	2.3	0.5	75.0
PVT-T [16]	224^2	13.2	1.6	75.1
ViT-C [19]	224^2	4.6	1.1	75.3
SCAT-XXS	224^2	4.9	0.7	76.6
ResT-lite [22]	224^2	10.5	1.4	77.2
EdgeViT-XS [8]	224^2	6.7	1.1	77.5
MobileViT-S [7]	256^2	5.6	2.0	78.4
PVTv2-B1 [17]	224^2	13.1	2.1	78.7
EdgeNext-S [6]	224^2	5.6	1.0	78.8
SCAT-XS	224^2	8.7	1.2	79.2
Swin-T [5]	224^2	29.0	4.5	81.3
ResT-Base [22]	224^2	30.3	4.3	81.6
PVTv2-B2 [17]	224^2	25.4	4.0	82.0
SCAT-S	224^2	31.0	4.1	83.1

Table 5: Comparison with the state-of-the-art on ImageNet-1k classification.

SCAT consistently outperforms SOTA vision transformer architectures when the parameters and FLOPs are close. our SCAT-XXS achieves 76.6% Top1-accuracy with only 4.9M parameters and 0.7G FLOPs. SCAT-XS achieves a better trade-off between FLOPs and top-1 accuracy than MobileViT and EdgeViT.

Furthermore, we evaluate the scaling capacity of our SCAT model by introducing a scaled-up SCAT-S, which incorporates 31M parameters and 4.1G FLOPs. As shown in the third part of Table 5, our SCAT-S model still demonstrates excellent competitiveness, outperforming Swin-T [5] and PVTv2-B2 [17] at similar parameters and FLOPs.

4. Conclusion

In this paper, we proposed SCAT, an efficient vision transformer. The core of our network is Spatial and Channel enhanced Self-Attention, which combines local spatial attention, global spatial attention, and channel attention. Local spatial attention and global spatial attention extract and reinforce local and global features, respectively. The channel attention further learns the relationship between local and global features. The experimental results demonstrate the efficiency of the SCAT model in the image classification task. In the future, we plan to evaluate our SCAT model on more vision tasks, such as object detection and image segmentation.

References

[1] Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. Mobile-former: Bridging mobilenet and transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5270–5279, 2022. 1

[2] 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. 3

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2

[4] 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. 2, 3, 4

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1, 3, 4

[6] Muhammad Maaz, Abdelrahman Shaker, Hisham Cholakkal, Salman Khan, Syed Waqas Zamir, Rao Muhammad Anwer, and Fahad Shahbaz Khan. Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications. In *European Conference on Computer Vision*, pages 3–20. Springer, 2022. 1, 2, 4

[7] Sachin Mehta and Mohammad Rastegari. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021. 1, 4

[8] Junting Pan, Adrian Bulat, Fuwen Tan, Xiatian Zhu, Lukasz Dudziak, Hongsheng Li, Georgios Tzimiropoulos, and Brais Martinez. Edgevits: Competing light-weight cnns on mobile devices with vision transformers. In *European Conference on Computer Vision*, pages 294–311. Springer, 2022. 1, 2, 4

[9] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 367–376, 2021. 1

[10] 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. 1

[11] Chenyang Si, Weihao Yu, Pan Zhou, Yichen Zhou, Xinchao Wang, and Shuicheng Yan. Inception transformer. *Advances in Neural Information Processing Systems*, 35:23495–23509, 2022. 1

[12] Zhao Song, Ke Yang, Naiyang Guan, Junjie Zhu, Peng Qiao, and Qingyong Hu. Pvp: Pre-trained visual parameter-efficient tuning. *arXiv preprint arXiv:2304.13639*, 2023. 1

[13] Zhao Song, Ke Yang, Naiyang Guan, Junjie Zhu, Peng Qiao, and Qingyong Hu. Vppt: Visual pre-trained prompt tuning framework for few-shot image classification. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1

[14] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 3

[15] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11534–11542, 2020. 2

[16] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 1, 2, 4

[17] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 1, 2, 4

[18] 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. 2, 3

[19] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *Advances in neural information processing systems*, 34:30392–30400, 2021. 2, 4

[20] Chenglin Yang, Yilin Wang, Jianming Zhang, He Zhang, Zijun Wei, Zhe Lin, and Alan Yuille. Lite vision transformer with enhanced self-attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11998–12008, 2022. 1, 4

[21] Ke Yang, Dongsheng Li, and Yong Dou. Towards precise end-to-end weakly supervised object detection network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8372–8381, 2019. 2

[22] Qinglong Zhang and Yu-Bin Yang. Rest: An efficient transformer for visual recognition. *Advances in neural information processing systems*, 34:15475–15485, 2021. 4