Self-attention model based on multi-scale convolution

Zhen Lin, Pengfei Xiao, Rong Guan, Zhining You*, Yunming Pu#
School of Computer Technology, Jimei University, Xiamen 361021, Fujian, China

ABSTRACT

In the field of deep learning, convolutional neural networks and transformer architecture have achieved considerable success. This paper combines the advantages of these two architectures to achieve the encoding of multi-scale spatial features in an image. This is done by using convolutional kernels of different sizes for convolutional operations at each stage. This allows the model to obtain markers with rich and diverse features. The self-attention mechanism is then used to further improve the feature representation and introduce residual links. The experimental results demonstrate that the proposed model exhibits robust performance on the CIFAR-100 and CIFAR-10 datasets, with comparable performance and fewer parameters compared to traditional CNN models.

Keywords: Deep learning, convolutional neural networks, self-attention mechanism

1. INTRODUCTION

The integration of Convolutional Neural Networks (CNNs)[1] and Transformer architecture[2] in the field of image recognition is currently a subject of active research.

In China, several advancements in image processing tasks have been proposed. One such advancement is the Asymmetric Cross-Attention Hierarchical Network (ACAHNet)[3], which utilizes an Asymmetric Multiheaded Cross Attention (AMCA) module to enhance inter-feature interactions for remote sensing image change detection. The dual-branch Transformer-CNN (DBTC) model[4] enhances face image super-resolution by extracting multi-scale features with a transformer branch and neighboring pixel change features with a CNN branch. In medical image analysis, a pneumonia recognition framework[5] employs a Multilevel Multi-Head Self-Attention (MMSA) mechanism to focus on task-relevant features. The Hybrid Transformer and CNN Attention Network (HTCAN)[6] is a model that combines transformer and CNN networks for stereo image super-resolution. It enhances single images and fuses stereo information. Furthermore, the Spatial Attention-Guided CNN-Transformer Aggregation Network (SCTANet)[7] employs Hybrid Attention Aggregation (HAA) and an MLP-based upsampling module to efficiently reconstruct facial image details.

In the field of deep learning research conducted abroad, there has been a significant focus on combining CNN and Transformer architecture. The CASCaded Attention DEcoder (CASCADE)[8] is an attention-based decoder for medical image segmentation. It leverages hierarchical visual transformer features with attention gates and convolutional modules. The PolSARFormer[9] model utilizes 3D and 2D CNN with local windowed attention (LWA) for polarimetric synthetic aperture radar (PolSAR) data classification. In the context of weakly supervised semantic segmentation, TransCAM[10] integrates Conformer's attention weights with CNN-generated class activation maps (CAMs) to capture both local and global features. An attention transformer-based architecture[11] has been proposed for MRI brain tumor classification, with the objective of enhancing accuracy by fusing local and global features and introducing an enhanced CNN (iResNet). WetMapFormer[12] is a network designed for the mapping of wetlands. It employs LWA to reduce the cost of the mapping process while improving the generalization of the features extracted.

The objective of this paper is to examine the integration of CNN and Transformer architecture, with a particular focus on the combination of multi-scale convolutions, residual structure, and multi-head self-attention mechanisms. The proposed model in this paper is designated as the Self-Attention Model Based on Multi-Scale Convolution (SAMC).

2. METHODOLOGY

The architectural diagram of the SAMC model is presented in Figure 1.

*youzhn@126.com or cec.jmu.edu.cn; phone 86 13906016909; #yunmingpu@163.com

International Conference on Optics, Electronics, and Communication Engineering (OECE 2024), edited by Yang Yue, Proc. of SPIE Vol. 13395, 133951K · © 2024 SPIE · 0277-786X · Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.3048367

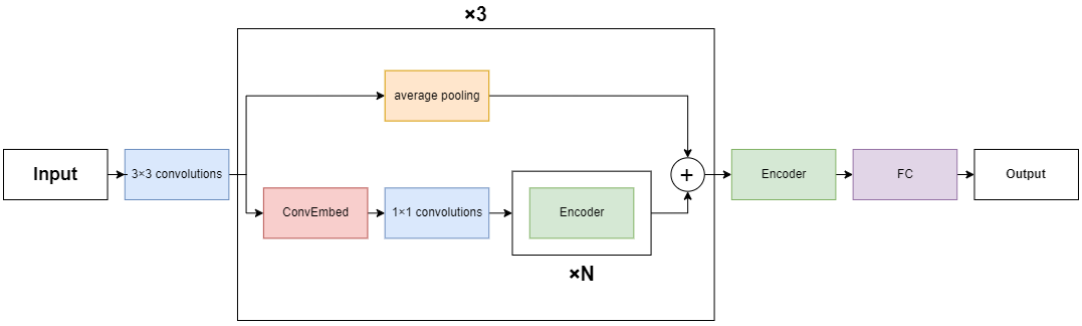

Figure 1. SAMC structure diagram.

2.1 Positional embedding

The utilization of positional embedding is of paramount importance for the deployment of the transformer in image classification scenarios where there is no inherent sequence order. The ViT model[13] accepts an RGB image with dimensions H×W×3, divides it into P×P patches, and then reorganizes it into a sequence for the transformer to process. The length *N* of the sequence is calculated in accordance with equation (1).

$$N = \mathrm{HW}/\mathrm{P}^2 \tag{1}$$

However, the ViT model may be subject to two potential limitations. Firstly, the model may not be able to effectively capture low-level features. Secondly, it may result in the parameters becoming too large. These two factors may impact the performance of the ViT model in tasks such as image classification.

In this paper, we draw on GoogLeNet's Inception module[14]. Firstly, a multi-size convolutional kernel is used to extract low-level features from the preprocessed image. Downsampling is performed after feature extraction using average pooling to reduce the computational burden for subsequent processing. Furthermore, due to the use of convolutional operations, the process of positional coding can be eliminated. The convolutional operation enables the network to learn the local features of the input data through the properties of local connectivity and weight sharing. Furthermore, the positions of these features in space are naturally encoded by the sliding process of the convolutional kernel. Therefore, the convolution operation here both extracts diverse low-level features and implicitly encodes positional information[15].

2.2 Convolution projection

The self-attention mechanism is a component of the Transformer model that allows the model to compute the attentional weight of each element in a sequence relative to all other elements. In natural language processing (NLP), the self-attention mechanism achieves its functionality by generating three vectors of query, key and value through a linear projection. However, this approach may lead to an increase in computation and the number of parameters.

To address this problem, CvT[15] proposes the use of convolutional projection instead of linear projection. Convolutional projection takes advantage of the spatial structure of the image by capturing local image features and encoding them as Q, K, and V. Additionally, the parameter sharing property of the convolutional layer reduces the model parameters and improves computational efficiency. Moreover, the parameter sharing property of the convolutional layer reduces the number of model parameters, improves computational efficiency, and preserves sensitivity to local image features.

In convolutional projection, multiple convolutional kernels are applied independently to each channel of the input feature map to produce a set of feature maps. Each convolution kernel is responsible for extracting one spatial feature, and this design helps the model to extract features at multiple scales, resulting in richer Q, K, and V representations. The formula for the convolutional projection is given by equation (2).

$$x_i^{q/k/v} = Flatten\big(Conv2d(Reshape2D(x_i), s)\big) \tag{2}$$

The notation x_i represents the token prior to convolutional projection, while $x_i^{q/k/v}$ denotes the input token of the Q/K/V matrix of the *i*th layer. The term Conv2d denotes depth-separable convolution, while s refers to the step size of the convolution.

The output of the convolutional projection can be used directly as input to the transformer layer, providing a flexible

feature representation for the model. The convolutional projection allows the model to better simulate the relationships between different regions in an image, improving performance in tasks such as image classification.

2.3 Residual structure

In this study, we used residual structure[16]. The residual structure formula is given in equation (3).

$$y = \text{Encoder}\big(\text{ConvEmbed}(x)\big) + \text{avg\_pool}(x) \tag{3}$$

The input image data, represented by the variable x, is processed through the ConvEmbed layer, which performs convolutional embedding operations to achieve positional coding and extract low-level features from the image. The output of ConvEmbed is then fed to the Encoder layer for further processing and encoding, which generates higher-level feature representations.

In order to achieve effective residual joining, the embodiment further comprises the step of performing an average pooling operation on the original image x. The purpose of this operation is to compress the size of the image and reduce the amount of computation while preserving critical global information. As a result, the output of avg_pool(x) is summed with the output of Encoder(ConvEmbed(x)) to form a residual link. This ensures that the model makes full use of the information in the original image during training, thereby improving the model's performance.

3. RESULTS

The SAMC model was trained and tested on the CIFAR-100 and CIFAR-10 datasets, respectively. The number of parameters and accuracy of the observed model are presented in Table 1.

Table 1. Accuracy and parameter table for SAMC model on CIFAR-10 and CIFAR-100.

Dataset	Parameter(M)	Top 1	Top 5
CIFAR-10	1.21	93.42%	99.80%
CIFAR-100	1.22	73.47%	92.22%

3.1 Application on CIFAR-100

After the SAMC model was trained on the CIFAR-100 dataset, the model parameters were increased from the original 1.22 M to 3.6 M, resulting in an improvement in accuracy from 73.47% to 74.74%. This was achieved by utilizing Optuna search parameters. The final parameters and accuracy of each network are presented in Table 2.

Table 2. Accuracy and parameter table for each network on CIFAR-100.

Net	Parameter(M)	Accuracy
SAMC	1.22	73.47%
VGG19	39.33	71.65%
MobileNetV2	2.37	66.12%
ShuffleNet	1.01	69.58%
ResNet34	21.33	76.83%
GoogLeNet	6.4	76.55%

Figure 2 illustrates the top-1 and top-5 accuracies of several networks on the CIFAR-100 dataset.

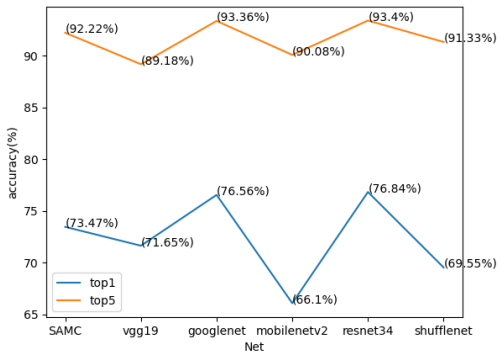

Figure 2. Top 1 and top 5 accuracy by network on CIFAR-100.

Figure 2 and Table 2 demonstrate that SAMC has a smaller number of parameters under the same conditions, with only approximately 1.22 M, and an accuracy of 73.47%.

3.2 Application on CIFAR-10

A series of convolutional neural networks and the SAMC model were trained and tested on the CIFAR-10 dataset under identical experimental conditions.

Figure 3 illustrates the Top1 and Top5 accuracies of the SAMC model and multiple convolutional neural networks on the CIFAR-10 dataset.

The parameters and maximum accuracy of each network on CIFAR-10 are presented in Figure 4.

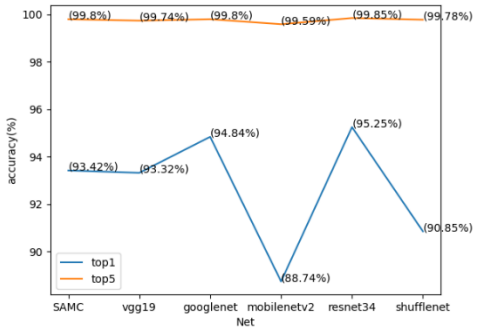

Figure 3. Top 1 and top 5 accuracy by network on CIFAR-10.

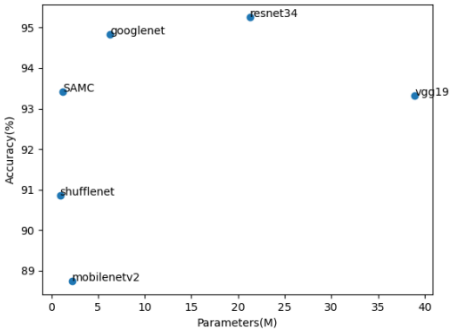

Figure 4. Parameters and accuracy of each network on CIFAR-10.

The results of each network tested on the CIFAR-10 dataset show that all have high accuracy. MobileNetV2 and ShuffleNet have slightly lower accuracy. In contrast, SAMC, VGG19, GoogLeNet and ResNet34 have relatively close accuracy, with SAMC having fewer parameters than the other networks.

4. CONCLUSION

Finding lightweight deep learning models can be beneficial in applications. Our study shows that by optimizing the network structure, model parameters can be significantly reduced while maintaining classification accuracy.

Looking ahead, we see several promising directions for future research. Further investigation of a more efficient network structure is needed. Applying the proposed model to a wider range of tasks and datasets will facilitate a deeper understanding of its generalizability. In addition, it is necessary to explore more effective attention mechanisms in order to reduce the computational requirements and further reduce the overall weight of the model.

In conclusion, it is anticipated that our work and the efforts of the research community will continue to drive progress in more efficient and effective models for widespread applications.

ACKNOWLEDGEMENTS

This work was financially supported by the Fujian science technology foundation under grant No. 2023R0048.

REFERENCES

[1] Krizhevsky, A., Sutskever, I. and Hinton, G. E., "ImageNet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems 60, 84-90 (2012).
[2] Vaswani, A., Shazeer, N., Parmar, N., et al., "Attention is all you need," Advances in Neural Information Processing Systems 30, 6000-6010 (2017).
[3] Zhang, X., Cheng, S., Wang, L., et al., "Asymmetric cross-attention hierarchical network based on CNN and transformer for bitemporal remote sensing images change detection," IEEE Transactions on Geoscience and Remote Sensing 61, 1-15 (2023).
[4] Shi, J., Wang, Y., Yu, Z., et al., "Exploiting multi-scale parallel self-attention and local variation via dual-branch transformer-CNN structure for face super-resolution," IEEE Transactions on Multimedia 26, 2608-2620 (2023).
[5] Chen, S., Ren, S., Wang, G., et al., "Interpretable CNN-multilevel attention transformer for rapid recognition of pneumonia from chest x-ray images," IEEE Journal of Biomedical and Health Informatics 28, 753-764 (2023).
[6] Cheng, M., Ma, H., Ma, Q., et al., "Hybrid transformer and CNN attention network for stereo image super-resolution," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 1702-1711 (2023).
[7] Bao, Q., Liu, Y., Gang, B., et al., "SCTANet: A spatial attention-guided CNN-transformer aggregation network for deep face image super-resolution," IEEE Transactions on Multimedia 25, 8554-8565 (2023).
[8] Rahman, M. M. and Marculescu, R., "Medical image segmentation via cascaded attention decoding," Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 6222-6231 (2023).
[9] Jamali, A., Roy, S. K., Bhattacharya, A., et al., "Local window attention transformer for polarimetric SAR image classification," IEEE Geoscience and Remote Sensing Letters 20, 1-5 (2023).
[10] Li, R., Mai, Z., Zhang, Z., et al., "Transcam: Transformer attention-based cam refinement for weakly supervised semantic segmentation," Journal of Visual Communication and Image Representation 92, 103800 (2023).
[11] Tabatabaei, S., Rezaee, K. and Zhu, M., "Attention transformer mechanism and fusion-based deep learning architecture for MRI brain tumor classification system," Biomedical Signal Processing and Control 86, 105119 (2023).
[12] Jamali, A., Roy, S. K. and Ghamisi, P., "WetMapFormer: A unified deep CNN and vision transformer for complex wetland map**," International Journal of Applied Earth Observation and Geoinformation 120, 103333 (2023).
[13] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv 2010.11929, (2020).
[14] Szegedy, C., Liu, W., Jia, Y., et al., "Going deeper with convolutions," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1-9 (2015).
[15] Wu, H., Xiao, B., Codella, N., et al., "CvT: Introducing convolutions to vision transformers," Proceedings of the IEEE/CVF International Conference on Computer Vision, (2021).
[16] He, K., Zhang, X., Ren, S., et al., "Deep residual learning for image recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778 (2016).