Image Recognition Based on Enhanced-Conformer

Runlin Gong
School of Computer Science and Cyber Engineering
Guangzhou University
Guangzhou, China
gongrunlin111@outlook.com

Ke Qi*
School of Computer Science and Cyber Engineering
Guangzhou University
Guangzhou, China
qikersa@163.com
*Corresponding author

Yicong Zhou
Department of Computer and Information Science
University of Macau
Macau, China
yicongzhou@um.edu.mo

Wenbin Chen
School of Computer Science and Cyber Engineering
Guangzhou University
Guangzhou, China
cwb2011@gzhu.edu.cn

Jingdong Zhang
Aberdeen Institute of Data Science and Artificial Intelligence
South China Normal University
Guangzhou, China
2574615315@qq.com

2022 IEEE 5th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE) | 978-1-6654-7197-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/AUTEEE56487.2022.9994542

***Abstract*—Convolutional Neural Networks (CNNs) has always dominated visual recognition tasks, and it is difficult to link distant information in images due to the size limitation of each convolution filter. Vision Transformer (ViT) can capture features at a distance in an image, but lacks the details of local features. Conformer combines the advantages of both using Convolutional Neural Networks (CNNs) and Attention mechanisms in parallel, but it does not take into account the relationship between different samples. Therefore, we propose a new attention calculation method, Extra-Attention, which can effectively learn intra-sample and inter-sample relationships. In order to combine the advantages of CNNs and Attention, in our work, we proposed a new network Enhanced-Conformer (ENC) based on Conformer, in which the attention mechanism adopts a more efficient computational module Inside And Outside Transformer (IAOT), which contains three parallel Attention: Extra-Attention, Self-Attention, External-Attention. Enhanced-Conformer (ENC) can fuse local features, global features and external features at the same time. We conduct experiments on the commonly used image recognition datasets, the recognition accuracies reach 93.29%, 68.70% and 57.63% on Tiny- ImageNet, CIFAR-10 and CIFAR-100, respectively.**

Keywords—CNNs, Extra-Attention, Transformer, Image recognition

I. Introduction

Convolutional neural networks (CNNs) show their advantages in extracting local features(Figure 1.a), while the Self-Attention mechanism can capture features at a distance(Figure 1.b). Recent works [1]–[5] demonstrate the benefits of using convolution and Transformer in combination. Some works [1] use convolution to extract features at the beginning, reducing the computational effort of the attention mechanism, or incorporating convolution into the backbone of each Transformer [4]. Conformer [6] uses a dual network structure of CNN and Transformer in parallel, which combines Transformer-based global representation with CNN-based local features, preserving the structural advantages of CNNs and Transformer and enhancing the local features and global representation capabilities. Self-Attention updates the features of each location by using the pairing affinity of all positions to calculate the weighting sum of features to capture different samples for a long distance. External-Attention (EA) [7] first calculates the attention by computing the affinity between its own query vector Q and the memory of an externally learnable keyword K, then multiplied by another randomly generated value V to produce a feature map of learnable external features, implicitly considering the correlation between all data samples, but this attention lacks modeling of the relationship between samples internally.

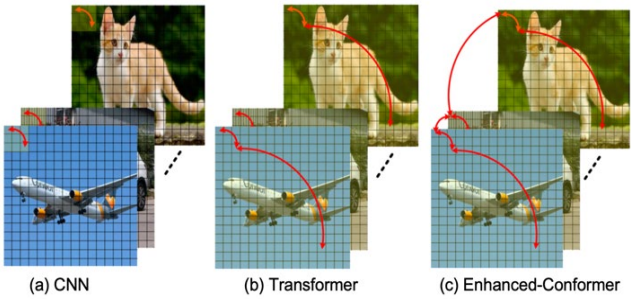

Fig. 1. Areas of interest of CNN, Transformer and Enhanced-Conformer.

Therefore, in this paper, inspired by External-Attention (EA) [7], we propose a better attention mechanism computation method Extra-Attention (EXA), which can efficiently learn not only the relationships within samples, but also between samples. To take advantage of both convolution and attention, we modify the Transformer attention mechanism module based on Conformer and propose a more efficient attention computation module Inside And Outside Transformer (IAOT), which contains three parallel Attentions (Figure 2) : Extra-Attention extracts the relationships within and between samples; Self-Attention enhances the features within samples; External-Attention [7] enhances the relationships between samples.

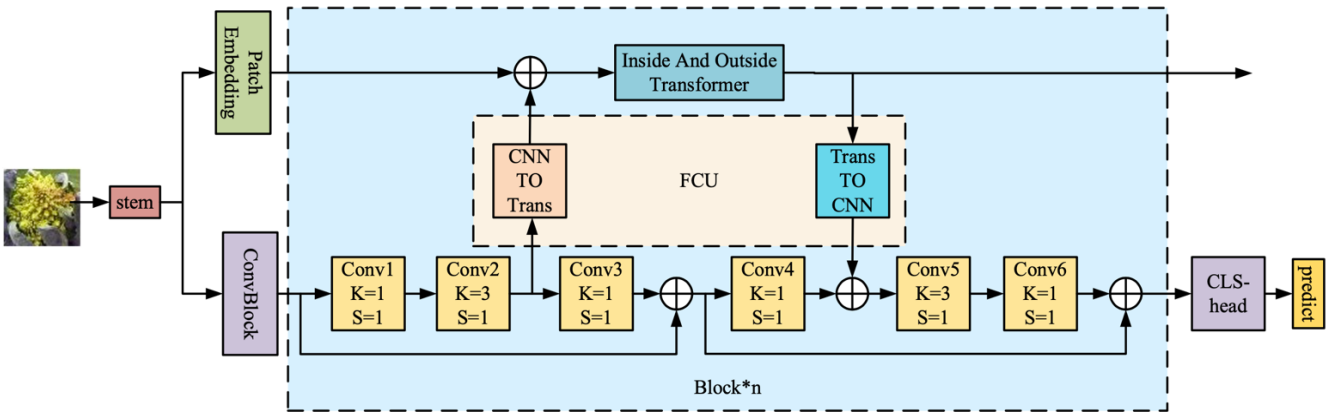

Fig. 2. Enhanced-Conformer.

The main contributions of this work are as follows:

- We propose a better method of Attention calculation Extra-Attention (EXA), which note only learns the relationships within and between samples, but also makes the training more stable.
- We propose a new network structure Enhanced-Conformer based on Conformer, in which we replace the original Transformer with our new Inside And Outside Transformer to fuse global features, local features and external features with each other.
- Extensive experiments show that our Enhanced-Conformer method performs significantly better than other similar methods on CIFAR-100, CIFAR-10 and Tiny- ImageNet datasets.

II. Related Work

A. CNN

Recently, the application of CNNs has greatly advanced the development of computer vision tasks and has been dominating the visual recognition tasks. In 2012, AlexNet [8] network was proposed, which created a new era in the computer vision tasks. In order to solve the problem of gradient and gradient disappearance, and overfitting that come with the increasing depth of the network, ResNet [9] network was proposed. Subsequently, VGGNet [10], GoogLeNet [11], Mo-bileNet [12], RegNet [13], and other representative CNNs have been studied in terms of accuracy, efficiency, and scalability and many useful principles have been designed. CNNs show their powerful advantages in extracting local features, but it is difficult for these networks to relate distant information in a picture, and each convolutional filter is restricted to operate on a small region while long-distance interaction between semantic concepts is crucial.

B. Transformer

Transformer is first widely used in NLP(Natural Language Processing) and successfully sparked special people's attention on its application to vision due to its powerful long-range modeling capabilities. To take the advantage of the self-attention mechanism, many Transformer models have shown perfect performance in many kinds of computer vision tasks such as image recognition [1], [14], object detection [15]–[16], image processing [17], and image segmentation [18]. The core idea of Vision Transformer [19] is to divide the images into non-overlapping fixed size blocks, e.g. 16*16, and then convert them into tokens by linear projection, adding class token before these patch tokens to form the input sequence, and embedding the learnable absolute position into each token before passing these tokens into the encoder. Since class token can learn relevant information from other tokens, on the last layer of the network, we can only use class-token as the final results for classification tasks. Transformer is able to handle long-range feature and complex spatial transformations dependencies. However, it lacks local feature details and proper induction bias, and performs poorly without being pre-trained with large data set.

C. Attention Mechanism

The advantage of Self-Attention is that it can obtain long-term dependencies by calculating the affinity between features. Its essence is to learn a set of weighted coefficients through the network and strengthen the network's regions of interest in a weighted manner. Swin-Transformer [18] includes overlapping cross-window and non-overlapping local window by sliding window operation, restricting the attention computation to one window, introducing localization of CNN and reducing the computation. External-Attention [7] is based on two learnable, small, external and shared memories and can be easily implemented using only two cascaded linear layers and two normalization layers with linear implicitly and complexity considering the correlation between all pictures in the data set.

Different from the previous work, Enhance-Conformer uses CNN and Inside And Outside Transformer (IAOT) parallel structure. This structure not only combines the structural advantages of Transformer and CNN to maximize the representational power of global representation and local features, but also enables effective fusion of intra-sample and inter- sample relationships (Figure 1.c).

III. Method

We first analyze the structure of Conformer [6] and summarize its advantages and areas for improvement. Secondly, we analyze three Attention computation methods, Self-Attention (SA), External-Attention (EA) and our proposed Extra-Attention (EXA). Finally, we introduce our new network structure, Enhanced-Conformer (ENC), which combines the three attention computation methods.

A. Conformer

Conformer [6] consists of three main components. CNN module, Transformer module and the FCU (Feature Coupling Unit). The CNN branches and Transformer branches are formed into a dual parallel network structure, which aims to combine global representation extracted based on Transformer and local features extracted based on CNN, FCU is used to fuse the local features in CNN branch and the global representation in Transformer branch. Both CNN and Transformer branches have the same number of convolution blocks and Transformer blocks.

Conformer inherits the advantages of the respective structures of Transformer and CNN to maximize the representational power of local features and global representation. However, the Self-Attention in Conformer is focused between different positions in a single sample, and update the features of each location by using the weighted sum of pairs of affinity calculation features in all locations to capture long-term dependencies, but ignores potential associations with other samples.

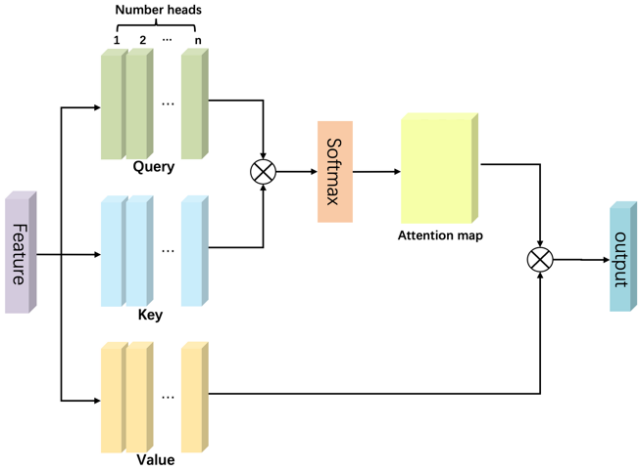

Fig. 3. Self-Attention calculation method.

B. Attention

Attention can be seen as a mechanism to reallocate the resources according to the importance of activation.The key to Transformer is the Attention. Three Attention computation methods are described below.

1) Self-Attention (SA): The core of the Self-Attention is to calculate the affinity to obtain remote dependencies between samples. As shown in Figure 3, the input feature $X \in \mathbb{R}^{n*d}$ are linearly converted to three parts, i.e., the first is queries $Q \in \mathbb{R}^{n*d_k}$, the second is keys $K \in \mathbb{R}^{n*d_k}$ and the last is values $V \in \mathbb{R}^{n*d_v}$, where n is the length of the sequence, the dimensions of both queries and keys are d_k. d_v and d are the dimensions of values and inputs. The scaled dot-product attention is applied on Q,K,V :

$$SA(Q,K,V) \triangleq softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Finally, using the linear layer to generate the output.

2) External-Attention (EA): In order to extract input features,Self-Attention use a linear combination of self-values to achieve it. However, in the Self-Attention, do we really need to compute the similarity between each feature? This would lead to too much redundant computation. In addition, Self-Attention only considers relationships within the same samples but ignores the relationships of other samples, Therefore, this calculation method limits its ability and flexibility. The use of External-Attention [7] mechanism can well integrate intra-sample and inter-sample relationships.

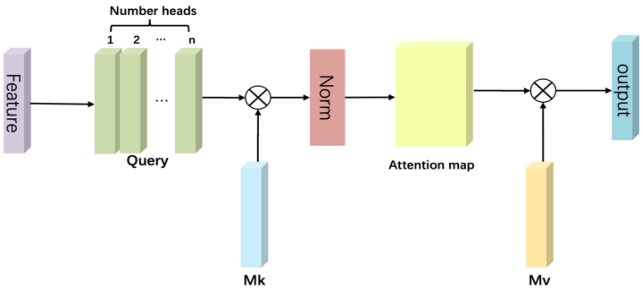

Fig. 4. External-Attention calculation method.

The External-Attention process is shown in Figure 4. First, the attention computes the attention between its own query vector Q and the memory of the external learnable key K . Then, it is multiplied by another external learnable value V to produce a feature map of learnable external features. The calculation method is as follows:

$$A = Norm\left(FM_k^T\right) \tag{2}$$

$$EA\left(Q,M_k,M_v\right) = M_v A. \tag{3}$$

Unlike Self-Attention, which computes the relationship between blocks, External-Attention computes the relationship between that block of pictures and other blocks of pictures. Where F is the input feature, M_k and M_v are input-independent learnable parameters that act as memory for the entire training dataset. A is the similarity inferred from the learning data set-level priori knowledge. Finally, the similarity in A is used to update the input features in M_k and M_v .

3) Extra-Attention (EXA): The Self-Attention can learn the relationships within a single sample, but lacks of intra-sample correlation. The External-Attention mechanism can learn the relationship between samples, but lacks the association within samples. Since M_k and M_v in External-Attention are randomly initialized parameters independent of the input, it learns the parameters with low generalization ability and may also lead to training instability.

Extra-Attention randomly initializes the parameter E_v , which is used to establish the relationship between the samples. Q and K are consistent with the self-attention mechanism and are generated by linear projection of the input features. First caculate the similarity between Q and K to establish the relationship between the interior of the pictures, and then the outputed feature map is obtained by

multiplying it with the extra learnable E_v after softmax, as shown in Figure 5, which is calculated as follows:

$$EXA\left(Q,K,E_v\right)=softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)E_v \tag{4}$$

In this way, Extra-attention not only learns the sample-to-sample relationships, but also models the internal relationships, and the training is more stable.

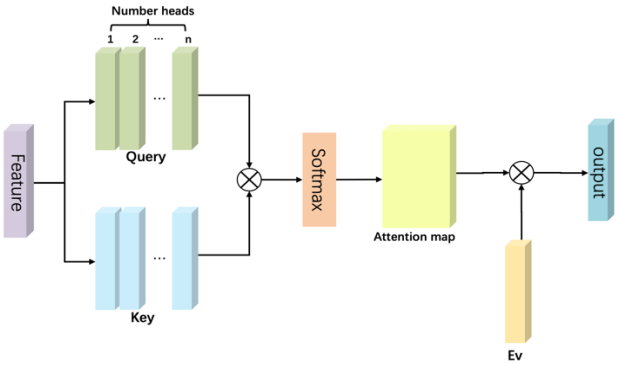

Fig. 5. Extra-Attention calculation method.

C. Enhanced-Conformer

The Conformer uses a dual network structure of CNN and Transformer in parallel to maximize the representational power of local features and global representation. However, the Self-Attention in Conformer focuses on different positions in a single sample, ignoring the potential association with other samples.

The three attention computation methods propose in 3.2 have their own advantages, so how to combine the three attention computation methods becomes a problem for us to think about. After a lot of experiments (refer to 4.4), we decide to use a new Inside and Outside Transformer (IAOT) structure that combines three types of attention in parallel, this structure works best.

Therefore, we propose a new Conformer structure, Enhanced-Conformer, as shown in Figure 2. The convolution part is consistent with the Conformer [6], and a new structure Inside And Outside Transformer (IAOT) is adopted for the Transformer part, as shown in Figure 6. The new structure greatly enhances the representation capability of the network and improves the accuracy of image recognition.

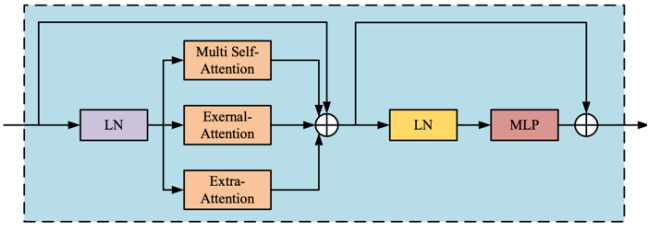

Fig. 6. Inside And Outside Transformer.

The structure of Enhanced-Conformer is shown in Figure 2. It consists of three branches. The convolutional branch is mainly used to extract the local features of the samples. IAOT is used to establish the relationship between the samples and samples and extract the global features of the image, and the FCU is used to fuse the features of this two modules.

First, input the 2D image $X\in\mathbb{R}^{(H*W*C)}$. H, W and C represent the height, width and number of channels of the input image. Using the stem module to extract low-level features, such as border texture information. IAOT can only receive one-dimensional token embedding sequences as input, so before entering the Transformer branch, it needs to go through Patch-Embedding processing to convert 2D images into 1D to get $Z\in\mathbb{R}^{N*\left(P^2*C\right)}$, where P is the size of each patches and $N=HW/P^2$ is the number of patches; before entering the convolution branch it needs to be processed by ConvBlock to reduce the height and width and increase the channel dimension. The calculation is as follows:

$$Z=Patch\_Embedding\left(stem\left(X\right)\right) \tag{5}$$

$$Y=ConvBlock\left(stem\left(X\right)\right) \tag{6}$$

In the convolutional branch, a pyramidal structure is used, where the height and width of the feature map decrease as the depth of the network increases, while the number of channels increases. As shown in Figure 2, there are two bottleneck networks in each layer, and each bottleneck network contains 1×1 down-projection convolution (Conv1 and Conv4), 3×3 spatial convolution (Conv2 and Conv5), 1×1 up-projection convolution (Conv3 and Conv6), and between the input and output of the bottleneck, there is a residual connection. The calculations are as follows:

$$Y_l^{'}=Conv2\left(Conv1\left(Y_{l-1}\right)\right) \tag{7}$$

$$Y_l^{''}=Y_{l-1}+Conv3\left(Y_{l-1}\right) \tag{8}$$

$$Y_l=Y_l^{''}+Conv6(Conv5(Conv4\left(Y_l^{''}\right)+TCC\left(Z_l\right))) \tag{9}$$

where Y_{l-1} denotes the output of the convolutional branch of the previous layer, Z_l denotes the output of the IAOT module of the current layer.

Since the feature dimensions of the convolutional network and IAOT are inconsistent, the dimension of the CNN feature map is $H*W*C$, the shape of the patch embedding is $\left(N+1\right)*E$, where N, 1 and E denote the number of image patches, class token, and embedding dimensions. In order to make the local features, global features, and external features interactively fused with each other, a feature coupling Unit (FCU) is used between the convolutional branch and IAOT, which contains two parts. TCC is used to fuse external and global features to local features by aligning the spatial dimensions using a downsampling module for patch embeddings, and then use a 1×1 convolution to align the number of channels. CTT is

used to fuse local features to global features and external features, first, the local feature map is convolved by 1×1 to align the number of channels of patch embeddings, and then complete the spatial dimension alignment by using the downsampling module.

Figure 6 shows the Transformer branch IAOT, and the module consists of LayerNorm (LN), Self-Attention (SA), External-Attention (EA), Extra-Attention (EXA), and MLP. EXA extracts relationships within and between samples; SA is used to enhance the relationships within samples; EA is used to enhance the relationships between samples. LN are applied before each layer, and there are residual connections in the MLP block and attention layers. The calculation is as follows:

$$Z_l^{'} = LN\left(Z_{l-1} + CCT\left(Y_l^{'}\right)\right) \tag{10}$$

$$Z_l^{''} = Z_l^{'} + EXA\left(Z_l^{'}\right) \oplus EA\left(Z_l^{'}\right) \oplus SA\left(Z_l^{'}\right) \tag{11}$$

$$Z_l = Z_l^{''} + MLP\left(LN\left(Z_l^{''}\right)\right) \tag{12}$$

Where Z_{l-1} represents the output of IAOT in the previous layer, the input of the first layer is the output of Patch Embedding, and CTT is used to fuse local features to extra features and global features. $Y_l^{'}$ - denotes the local features extracted by the l *th* layer CNN branch; EXA denotes the global and extra features extracted by Extra-Attention; EX denotes the external features extracted by External-Attention; SA denotes the internal global features extracted by Self- Attention; $\oplus$ denotes the sum of elements.

In the last layer of this network, we only take the output of the convolution result to get the final result after CLS-head processing.

IV. Experiment

A. Dataset

In our experiment, we used three data sets, Tiny ImageNet , CIFAR-10 and CIFAR-100. The first data set Tiny ImageNet is a part of ImageNet, it contains 200 categories, and the resolution of images is 64*64, The last two data sets CIFAR-10 and CIFAR-100 contain 10 and 100 categories respectively, and the image resolution is 32*32. Each category contains 6,000 pictures, including 5,000 training sets and 1000 verification sets. The three datasets are presented in Table I.

TABLE I. Datasets

Datasets	Train size	Test size	Classes	Image size
Tiny ImageNet	100000	100000	200	64*64
CIFAR-100	50000	10000	100	32*32
CIFAR-10	50000	10000	10	32*32

B. Setup

Experimental implementation details: the computational resource used for modality training is an NVIDIA GTX 3080 graphics card with the memory size of 10G. Based on the Pytorch-based framework, our model is trained with 300 Epochs, using the AdamW optimizer to accelerate the convergence of the model weights, with a batch size of 56, and decay according to the cosine schedule, considering the limited computing resources, the input image resolution is uniformly set to 224*224. The detailed configuration of the experiment is shown in Table II.

TABLE II. Experimental Parameters Settings, LR Means Learning Rate

Epochs	Optmizer	Batch size	LR	LR scheduler	Weight decay	Warmup Epochs	Lable Smothing	Drop Patch
300	AdamW	56	0.001	cosine	0.05	5	0.1	0.1

By stacking different layers of convolutional blocks and IOAT blocks in Enhanced-Conformer, modifying the embedding dimension, and setting the size of different patch sizes, we can obtain different size models, and refer to Table III for detailed settings. due to limited computational resources, after weighing the training time and model accuracy, we used the Base model for subsequent experiments.

TABLE III. Model Size

Model	Patch size	Patch dim	Number heads	MLP ratio	Depth	Macs	Model size
Tiny	16	384	6	1	6	2.22GMac	9.49M
Base	16	384	6	4	12	6.37GMac	29.18M
Large	32	576	9	4	12	20.74GMac	64.76M

C. Result

All the following models are trained from scratch without any pre-training or fine-tuning. While maintaining the same training configuration as much as possible, we compared common model results such as Vit, CvT, Conformer and ResNet18. As shown in Table IV, Enhanced-Conformer outperformed the other models in all three datasets, Tiny ImageNet, CIFAR-100, and CIFAR-10. By comparing the results of ResNet18-Vit and Vit, we can find that ResNet18-Vit with the combination of convolutional net-work and attention mechanism is significantly better than Vit with pure attention mechanism. Although Transformer tends to have a larger model capacity, it has weaker generalization ability due to the lack of correct induction bias, and it requires a large amount of data set pre-training and fine-tuning to achieve better results. The main difference between ResNet18-Vit and Conformer is that the former connects the convolutional network and Transformer in tandem, using convolution to extract local features first, and then using the attention mechanism to establish global associations, while the latter connects convolution and Transformer in parallel. The CNN and the Transformer process the feature maps in parallel, and it is clear from the experimental data that the parallel structure is significantly better than the series structure.

TABLE IV. COMPARED WITH OTHER MODELS

Approaches	Tiny ImageNet		CIFAR-100		CIFAR-10	
	Top-1(%)	Top-5(%)	Top-1(%)	Top-5(%)	Top-1(%)	Top-5(%)
ViT[17]	24.79	49.77	41.66	70.65	66.79	97.15
TNT[20]	28.41	54.34	38.02	68.95	66.86	97.47
Deit[15]	28.63	55.36	37.12	67.89	64.08	96.72
CvT[1]	29.34	56.75	36.23	68.08	68.50	97.80
Resnet18_linear[9]	49.60	51.23	38.71	52.11	65.82	87.46
Resnet18_vit	51.27	64.79	49.13	74.15	78.33	98.33
Conformer[6]	52.08	69.39	59.26	83.80	91.86	99.66
Enhanced-Conformer(Ours)	**57.63**	**80.65**	**70.46**	**91.54**	**93.29**	**99.80**

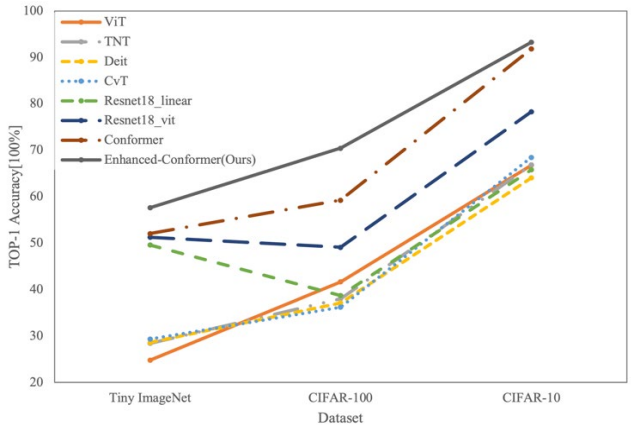

Fig. 7. Top-1 Comparison chart of accuracy.

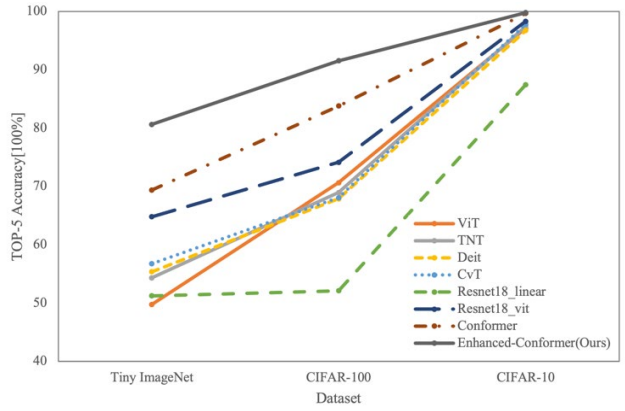

Fig. 8. Top-5 Comparison chart of accuracy.

As shown in Figures 7 and 8, compared with Conformer, our proposed Enhanced-Conformer improves Top1 accuracy by 5.55%, 11.2% and 1.43% in the three datasets, respectively, and the results show that we replace the original Transformer of Conformer with the new Inside And Outside Transformer (IAOT) for the effectiveness of modeling intra-sample.

D. Ablation Study

To verify the effectiveness of the IAOT module, we designed ablation experiments, as shown in table V. The three attention mechanism calculation methods listed in 3.2 have their own advantages. Self-Attention (SA) is good at extracting intra-sample features. External-Attention (EA) is good at extracting sample-to-sample features. And Extra-Attention (EXA) can extract not only intra-sample features but also inter-sample features. The question of how to combine the three attention computation methods become our consideration. For this purpose, we design five different structures, including a single- line structure by directly replacing the original SA with EA or EXA, a two-line structure with SA in parallel with EA and EXA respectively, and finally a three-line structure with SA, EXA, and EA in parallel. It is found that the best results are achieved when the three Attentions are connected in parallel to form the new Inside And Outside Transformer (IAOT) structure.

TABLE V. ABLATION EXPERIMENTAL DATA

Approaches	Tiny ImageNet		CIFAR-100		CIFAR-10	
	Top-1(%)	Top-5(%)	Top-1(%)	Top-5(%)	Top-1(%)	Top-5(%)
EA	56.25	79.87	65.20	89.29	91.06	99.73
EXA	53.14	77.53	67.93	90.27	89.02	99.64
SA+EXA	56.27	78.93	69.05	90.38	92.21	99.78
SA+EA	55.74	78.63	69.64	91.29	91.84	99.73
SA+EXA+EA	**57.63**	**80.65**	**70.46**	**91.54**	**93.29**	**99.80**

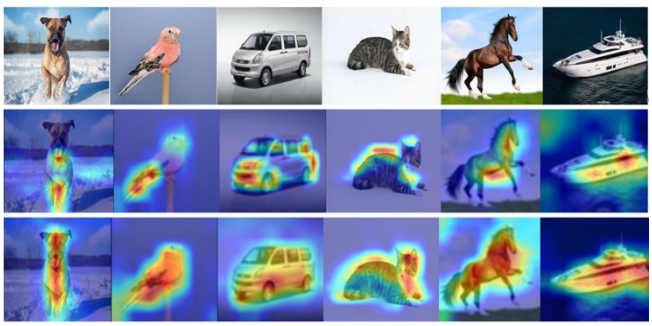

Fig. 9. Attention area.

E. Visualization

In order to better reflect the focus of the model, The CAM method is used to draw the attention area, as shown in Figure 9. The first row represents the original picture, the second row is the area of interest to Conformer, and the last row is the area of interest to our model Enhanced-Conformer. which can be compared to find that our model Enhanced-Conformer focuses on more details and more useful areas, implying the effect of Enhanced Conformer feature extraction is better. So the recognition accuracy is higher.

V. Conclusion

First, we propose a new attentional computation method, Extra-Attention, which can model the relationships between intra-samples and inter-sapmles. To combine the advantages of convolution for extracting local features. We propose a new network structure Enhanced-Conformer based on Conformer, in which we replace the original Transformer with our new Inside And Outside Transformer, it based on three kinds of Attention (Self-Attention, External-Attention, Extra-Attention) in parallel. The Enhanced-Conformer is able to fuse local features, global features and external features with each other, which effectively improves the accuracy of image recognition. Next, we will further optimize our Attention algorithm. We expect that the three Attention mechanisms can be merged and extended to other more complex computer vision tasks.

Acknowledgment

This work is supported by the Science and Technology Projects in Guangzhou (202102010412) and Yangcheng Scholars Research Project of Guangzhou (202032832), and the innovation training program for college students of Guangzhou University (202111078028,s202011078043), and National Natural Science Foundation of China (u1936116) .

References

[1] H. Zhang, W. Hu, and X. Wang, "Parc-net: Position aware circular convolution with merits from convnets and transformer,"in European Conference on Computer Vision. Springer, 2022, pp. 61.

[2] H. Zhang, W. Hu, and X. Wang, "Edgeformer: Improving light-weight convnets by learning from vision transformers," arXiv e-prints, 2022.

[3] M. Maaz, A. Shaker, H. Cholakkal, S. Khan, S. W. Zamir, R. M. Anwer, and F. S. Khan, "Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications," arXiv preprint arXiv:2206.10589, 2022.

[4] Y. Chen, X. Dai, D. Chen, M. Liu, X. Dong, L. Yuan, and Z. Liu, "Mobile-former: Bridging mobilenet and transformer," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5270–5279.

[5] Y. Liang, C. Ge, Z. Tong, Y. Song, J. Wang, and P. Xie, "Not all patches are what you need: Expediting vision transformers via token reorganizations," arXiv preprint arXiv:2202.07800, 2022.

[6] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, and Q. Ye, "Conformer: Local features coupling global representations for visual recognition," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 367–376.

[7] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. Hu, "Beyond self-attention: External attention using two linear layers for visual tasks," arXiv preprint arXiv:2105.02358, 2021.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural informa- tion processing systems, vol. 25, 2012.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[11] C.Szegedy,W.Liu,Y.Jia,P.Sermanet,S.Reed,D.Anguelov,D.Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[12] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[13] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dolla´r, "Designing network design spaces," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10 428–10 436.

[14] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Je´gou, "Training data-efficient image transformers & distillation through attention," in International Conference on Machine Learning. PMLR, 2021, pp. 10 347–10 357.

[15] D. Bolya, C.-Y. Fu, X. Dai, P. Zhang, and J. Hoffman, "Hydra attention: Efficient attention with many heads," arXiv preprint arXiv:2209.07484, 2022.

[16] J. Li, X. Xia, W. Li, H. Li, X. Wang, X. Xiao, R. Wang, M. Zheng, and X. Pan, "Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios," arXiv preprint arXiv:2207.05501, 2022.

[17] R. Yang, H. Ma, J. Wu, Y. Tang, X. Xiao, M. Zheng, and X. Li, "Scalablevit: Rethinking the context-oriented generalization of vision transformer," arXiv preprint arXiv:2203.10790, 2022.

[18] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7262–7272.

[19] L. Ke, M. Danelljan, X. Li, Y.-W. Tai, C.-K. Tang, and F. Yu, "Mask transfiner for high-quality instance segmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4412–4421.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.