2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA) | 978-1-6654-9991-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAICA54878.2022.9844458

Image Recognition based on Multi-scale Feature Fusion Transformer

Zhefeng Zhu
School of Computer Science and Cyber Engineering
Guangzhou University
Guangzhou, China
1808979894@qq.com

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

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

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

Peiyue Li
School of Computer Science and Cyber Engineering
Guangzhou University
Guangzhou, China

Zhenxian Liu
School of Computer Science and Cyber Engineering
Guangzhou University
Guangzhou, China

***Abstract*—Aiming at the problem that image recognition based on transformers has low image recognition rate due to ignoring local information of image blocks, an image recognition framework based on multi-scale Feature Fusion Transformer (FFT) is proposed, where the FFT block is designed to fuse feature information of different scales, and the residual attention module is introduced to emphasize feature channels and feature regions of interest. The FFT framework not only avoids the problem of vision transformer internal structure and local information loss of image feature blocks but also captures richer detailed features, which effectively improves the image recognition rate. A large number of experiments are performed on common image recognition datasets Tiny-ImageNet, CIFAR-10 and CIFAR-100, and the recognition accuracy can reach 57.81%, 82.04% and 56.98%, respectively, which are significantly higher than the mainstream image recognition algorithms.**

Keywords—image recognition, vision transformer, FFT block, multi-scale feature fusion

I. INTRODUCTION

Recently, the transformer model based on self-attention mechanism has shown good performance in computer vision tasks such as image recognition [1,2], image detection [3,4] and image processing [5]. Different from convolutional neural networks, ViT [7] provides a calculation method without image specific inductive bias, and breaks through the limitation of image receptive field by relying on self-attention mechanism. Currently, most of the image recognition frameworks based on transformer [4,5,7] segment the image into a series of image patches, and then use the self-attention mechanism to calculate the relationship between each image patch and the whole image. However, these methods ignore the internal structure of image feature patches and can not make full use of the low resolution local feature information. Therefore, it is difficult for the model to learn the detailed features of the image, resulting in a low recognition rate of the model.

In this paper, an FFT network structure based on multi-scale feature fusion transformer is proposed, which contains three parts: image feature extraction module, residual attention module, and FFT module. The framework first uses convolutional neural network to extract feature representations of images at different scales, then captures the important feature information through residual attention networks, and finally uses FFT block to fuse the feature information at different scales layer by layer, so that the detailed features in the low-level feature maps can be retained.

The main contributions of this paper are summarized as follows:

- An FFT network structure based on multi-scale feature fusion transformer is proposed, which makes up for the deficiency of vision transformer model due to ignoring the internal structure and local information of image feature patches.
- The residual attention network is introduced to emphasize the image feature channels and feature regions of interest to the model, which enhances the robustness of the model and makes the model more efficient and accurate in image recognition tasks.
- Extensive experiments show that the proposed method outperforms the other competing methods significantly on Tiny-ImageNet, CIFAR-10 and CIFAR-100 datasets.

II. RELATED WORK

A. Transformer

Chen et.al proposed iGPT [8], and applied the transformer model to self-supervised pre-training image recognition task. ViT [7] proposed by Dosovitskiy et.al segments the image into a series of patches and uses only the transformer encoder to accomplish the classification task. The design provides ideas for subsequent works, DeiT [2], ViT-FRCNN [9], IPT [6], DETR [4] and SETR [10]. DeiT [2]

further trains and refines ViT [7] more efficiently. IPT [5] uses the transformer to process multiple low-level vision tasks simultaneously in a single model. DETR [4] treats target detection as a direct set prediction problem and uses transformer's encoder-decoder architecture for the detection task. TNT [11] uses internal and external transformers to model the global relationship between image patches [12, 13]. T2T-ViT [14] splices the tokens of the input transformer for many times, using a convolutional sliding window approach to aggregate locally adjacent patches together. Compared with the mainstream CNN models [15-18], these models based on transformers can obtain very competitive accuracy without inductive bias. For example, ViT [7] has 77.9% ImageNet top-1 accuracy while using 86M parameters and 55.5B FLOPs, and DeiT [2] without pre-training has 81.8% ImageNet top-1 accuracy while using 86.4M parameters and 17.6B FLOPs.

B. Attention Mechanism

The essence of attention mechanism [19] is a method that learns a set of weight coefficients through the network and emphasizes the region of interest of the network in a weighted way. Wang et.al proposed Efficient Channel Attention [20], and modeled the correlation between channels by considering each channel and its K nearest neighbor channels. Zhu et.al proposed Spatial Attention [21], and modeled the pixel correlation of different spatial positions on the feature map. Wang et.al combined the channel attention module and spatial attention module to improve the accuracy and robustness of feature extraction, and introduced the residual operation into the attention module [22]. The use of attention module further improves the competitiveness of the model. For example, SENet [23] can be inserted into any network and has a 0.4 to 1.8% reduction in the error rate of ImageNet top-1. CBAM [24] combined the channel attention module and spatial attention module to further reduce the error rate of ImageNet top-1 to 22.96%.

III. Approach

A. Framework

This paper proposes an FFT network structure based on multi-scale feature fusion transformer, as shown in Figure 1. It mainly includes three main parts: a) Image feature extraction network; b) Residual attention module; c) Feature fusion module. Specifically, the image is input into the feature extraction network ResNet [16], which will output four feature layers on four stages. The lower layers are rich in detail information while the higher layers are rich in semantic information, and then the different feature layers are input into the residual attention module to make the network select the feature channels and feature regions of interest, and finally, the feature map is segmented and projected into embedding vectors and input into the FFT Block for feature fusion. The final embedding vector not only contains the high-level semantic information, but also retains the fine feature information in the low-level feature map, which provides richer and effective fusion features for image recognition.

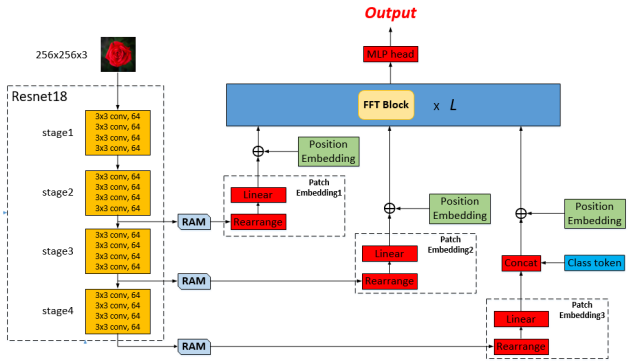


Fig. 1 FFT network framework

B. Image feature extraction

In this paper, ResNet [16] is used as a feature extraction network to extract feature maps of images, removing the adaptive average pooling layer and the fully connected layer from the ResNet network and using only the convolutional layers for feature extraction.

As shown in Figure 2, for the feature extraction network ResNet18, the input 256*256 size image, after the convolution of each stage, the feature maps of 64*64*64, 128*32*32, 256*16*16, 512*8*8 are output in turn, the feature maps in different stages have different scale feature information. In order to avoid the model which is too complex, only the last three feature maps are input to the FFT Block for feature fusion. In addition, in order to improve the robustness of the model, each feature map is input into the RAM module to emphasize the feature channel and the feature region concerned by the network, while keeping the feature dimension unchanged. The structure of Residual Attention Module (RAM) [22] is shown in Figure 2.

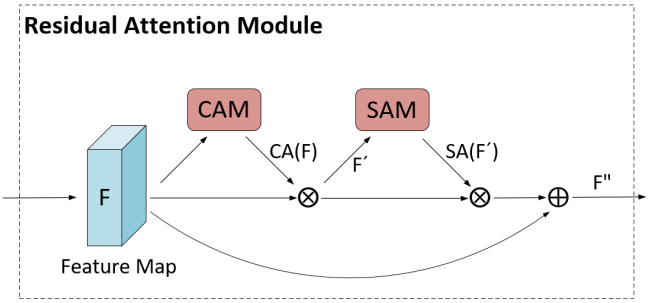


Fig. 2 Residual Attention Module(RAM)

By placing the channel attention module and spatial attention module before and after, after extracting the convolution feature map $F \in R^{C\times H''\times W''}$, the RAM module will first learn the channel attention map $CA(F) \in R^{C\times 1\times 1}$, and then multiply it with the original feature map in a weighted way to obtain the features $F' \in R^{C\times H''\times W''}$ by matrix multiplication. Then input F' into the spatial attention module to calculate the spatial attention map $SA(F') \in R^{1\times H''\times W''}$, multiplication of $SA(F')$ and F' is used to obtain the new feature maps. In addition, the residual operation is introduced to enable better aggregation of the model. The calculation process is as follows:

$$F' = CA(F) \otimes F \tag{1}$$

$$F'' = SA(F') \otimes F' \oplus F \tag{2}$$

where $\otimes$ is element multiplication and $\oplus$ is element addition.

C. FFT Block feature fusion

The fusion of features at different scales occurs in the FFT Block, which is the core of the model, and this module is designed to avoid as much as possible the problem of vision transformer's loss of internal structure and local information of the image feature patches.

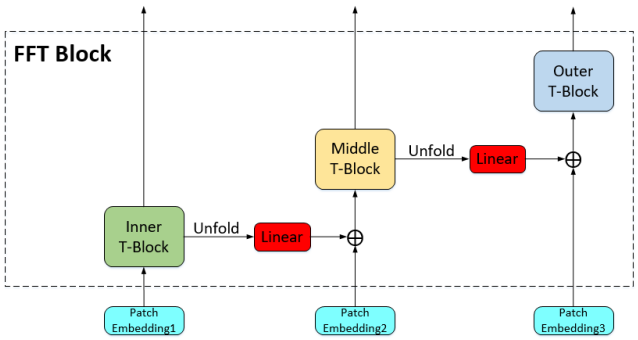


Fig.3 FFT Block structure

As shown in Figure 3, the FFT Block maintains three Transformer-Block at the same time: Inner-T-Block, Middle-T-Block and Outer-T-Block, which are used to process feature embedding of shallow, middle, and deepest feature maps respectively. For the shallow feature map with 128 channels of size 32*32 is uniformly segmented into n patches:

$$x = [x^1, x^2, x^3, \ldots, x^n] \in R^{n \times p \times p \times 128} \tag{3}$$

where$(p \times p)$is the resolution of each patch on the feature map, and then each patch tensor x^i is transformed into an embedding sequence by pixel expansion and projection:

$$X_0 = [X^1, X^2, X^3, \ldots, X^n] \in R^{n \times c} \tag{4}$$

where c is the dimension of each patch embedding vector. For the embedding sequence generated from the shallow feature map, use Inner-T-Block to explore the relationship between patches:

$$X'^i_l = X^i_{l-1} + MSA(LN(X^i_{l-1})) \tag{5}$$

$$X^i_l = X'^i_l + MLP(LN(X'^i_l)) \tag{6}$$

where $l = 1, 2, 3 \ldots\ldots$ and l represents the number of layers.

The feature map with 256 channels of size 16*16 is uniformly segmented into m patches:

$$y = [y^1, y^2, y^3, \ldots, y^m] \in R^{m \times p' \times p' \times 256} \tag{7}$$

The embedded sequence is obtained through pixel expansion and projection:

$$Y_0 = [Y^1, Y^2, Y^3, \ldots, Y^m] \in R^{m \times c'} \tag{8}$$

For feature fusion with the embedded sequence generated by Inner-T-Block, the embedded sequence generated by Inner-T-Block is added to the embedded sequence of this layer after unfold [25] and projection:

$$Y^i_{l-1} = Y^i_{l-1} + Vec(X^i_{l-1})W_{l-1} + b_{l-1} \tag{9}$$

where $Y^i_{l-1} \in R^{c'}$, $Vec(\bullet)$ flattens the embedded sequence locally, W_{l-1} and b_{l-1} are the weights and offsets of linear projection. After adding the embedded sequence, use Middle-T-Block to explore its relationship:

$$Y'^i_l = Y^i_{l-1} + MSA(LN(Y^i_{l-1}) \tag{10}$$

$$Y^i_l = Y'^i_l + MLP(LN(Y'^i_l)) \tag{11}$$

The deep feature map with 512 channels of size 8*8 is uniformly segmented into k patches:

$$z = [z^1, z^2, z^3, \ldots, z^k] \in R^{k \times p'' \times p'' \times 512} \tag{12}$$

The embedded sequence is obtained by pixel expansion and projection:

$$Z'_0 = [Z^1, Z^2, Z^3, \ldots, Z^k] \in R^{k \times c''} \tag{13}$$

Different from the first two T-blocks, Outer-T-Block needs to create an additional embedding Z_{class} to store the global relationship between embedding sequences:

$$Z_0 = [Z_{class}, Z^1, Z^2, Z^3, \ldots, Z^k] \in R^{(k+1) \times c''} \tag{14}$$

where Z_{class} is similar to the class token in ViT, they are all initialized with zero, and the dimension remains the same as the embedded sequence in Outer-T-Block. Similarly, the embedded sequence generated by Middle-T-Block is added to the embedded sequence of this layer after unfold and projection:

$$Z^i_{l-1} = Z^i_{l-1} + Vec(Y^i_{l-1})W'_{l-1} + b'_{l-1} \tag{15}$$

After adding the embedded sequence, use Outer-T-Block to explore its relationship:

$$Z'^i_l = Z^i_{l-1} + MSA(LN(Z^i_{l-1}) \tag{16}$$

$$Z^i_l = Z'^i_l + MLP(LN(Z'^i_l)) \tag{17}$$

In summary, the input and output of the FFT Block consists of three embedding vectors data streams, as shown in Figure 3, so the FFT can be expressed as:

$$X_l, Y_l, Z_l = FFT(X_{l-1}, Y_{l-1}, Z_{l-1}) \tag{18}$$

In the FFT model, the relationship between different scale feature embedding is modeled by FFT block. The output of each T-block is also used as the input of the corresponding T-block of the next layer. The FFT block is stacked on L layers to form the FFT network. Finally, the learnable embedding Z_{class} is used as the image representation, and MLP is applied to the image recognition task.

D. Position Encoding

Image spatial information is an important factor in image recognition. Only position coding is added to the image patch in vision transformer, which leads to loss of structural position information in the image patch, and it is difficult for the network to learn the fine feature of the image. Therefore, in this paper, for the patch embedding sequences generated by three different feature layers, position encoding is added for preserving the internal spatial information of the feature blocks separately as follows:

- Inner-T-Block Position Encoding:

$$X_0 \leftarrow X_0 + E_{patch} \tag{19}$$

where $E_{patch} \in R^{n \times c}$, n and c represent the number and dimension of the added position encoding.

- Middle-T-Block Position Encoding:

$$Y_0 \leftarrow Y_0 + E'_{patch} \tag{20}$$

where $E'_{patch} \in R^{m \times c'}$.

- Outer-T-Block Position Encoding:

$$Z_0 \leftarrow Z_0 + E''_{patch} \tag{21}$$

where $E''_{patch} \in R^{(k+1) \times c''}$. In particular, because an additional embedding is added to the embedding sequence of Outer-T-Block, an additional positional encoding is added to the embedding sequence, that is $(k+1)$.

IV. EXPERIMENTS

A. Datasets

Four datasets are used in this paper: Tiny ImageNet, Cifar-10, Cifar-100 and Flowers. Tiny ImageNet is an image classification dataset provided by Stanford University and is a subset of the original ImageNet. Tiny ImageNet contains 200 different classes, each of which includes 500 training images and 50 test images. The resolution of the image is only 64*64 pixels, which makes it more challenging to extract features. Cifar-10 and Cifar-100 contain 10 categories and 100 categories respectively, and both contain 50000 training images and 10000 test images. The Flowers dataset contains 5 flower category images. As Table I shows the introduction of these 4 datasets.

TABLE I. DATASETS INTRODUCTION

Datasets	Train size	Test size	#Classes
Tiny ImageNet	100000	10000	200
Cifar-10	50000	10000	10
Cifar-100	50000	10000	100
Flowers	2204	734	5

B. Experimental Settings

Specific implementation details of the experiment: The computing resources used for model training are two NVIDIA GTX 1070 graphics cards with a memory size of 8G. Experiments were conducted on the Pytorch-based framework using the SGD optimization algorithm to accelerate the convergence of model weights. The initial learning rate is set to 0.001. Using the learning strategy ReduceLROnPlateau provided by pytorch to dynamically select the learning rate, when the loss or accuracy does not change, the learning rate is reduced to one tenth of the original, with the batch size of 4, and the SGD optimizer is set with a momentum of 0.9 to fine tune the weights.

Model variants of different sizes are obtained by different configurations of the number of module layers and embedding dimensions in the FFT. As shown in Table II, the parameter size of the whole model is determined by the feature extraction network Resnet18, Residual Attention module and FFT module. Due to the limited computing resources, after balancing the accuracy and training time, the comparative experiment is carried out under the FFT-Base model.

TABLE II. DATASETS INTRODUCTION

Model	FFT-Block Layers	Patch dim	Mlp dim	Heads	Params
Base	6	512	1024	16	266M
Tiny	3	256	512	12	43M
Lager	12	512	1024	18	566M

C. Comparative experiment

The experimental results of common classification models such as ViT, ResNet18, and DeiT are compared without taking pre-training, and the training strategy remained consistent, the FFT model outperforms the experimental results on each dataset. As shown in Table III and IV, comparing the accuracy of ViT and ResNet18_vit on each dataset, it shows that transforming the feature patch into embedded sequences to complete the image classification task is much more accurate than inputting the image patch directly into ViT. Comparing the experimental results of Resnet18_vit and Resnet18_linear, it can be obtained that the same image features are processed, and the long-range feature dependence and self-attention mechanisms of ViT is better than simply using the fully connected layer. As shown in Figures 4 and 5, the accuracy of the FFT model is compared to other models on each dataset, and the accuracy of the method in this paper is 1.4%, 6.5%, 3.7%, and 7.8% higher than the top-1 accuracy on the four datasets on the same experimental configuration. The results show that the FFT structure is effective for the fusion of image features at different scales.

TABLE III. COMPARISON OF ACCURACY OF OTHER MODELS(TOP-1)

Approaches	Flowers Top-1	Tiny ImageNet Top-1	Cifar10 Top-1	Cifar100 Top-1
ViT[7]	65.67%	24.79%	66.79%	41.66%
Resnet18 vit	87.74%	51.27%	78.33%	49.13%
Resnet18 linear[16]	87.33%	49.60%	65.82%	38.71%
TNT[11]	66.89%	28.41%	66.86%	38.02%
DeiT[2]	67.98%	28.63%	64.08%	37.12%
CvT[1]	67.30%	29.34%	68.50%	36.23%
Ours	**89.23%**	**57.81%**	**82.04%**	**56.98%**

TABLE IV. COMPARISON OF ACCURACY OF OTHER MODELS(TOP-N)

Approaches	Flowers Top-3	Tiny ImageNet Top-5	Cifar10 Top-5	Cifar100 Top-5
ViT[7]	91.82%	49.77%	97.15%	70.65%
Resnet18 vit	96.87%	64.79%	98.33%	74.15%
Resnet18 linear[16]	90.87%	51.23%	87.46%	52.11%
TNT[11]	93.32%	54.34%	97.47%	68.95%
DeiT[2]	94.14%	55.36%	96.72%	67.89%
CvT[1]	94.28%	56.75%	97.8%	68.08%
Ours	**97.68%**	**75.41%**	**98.82%**	**81.19%**

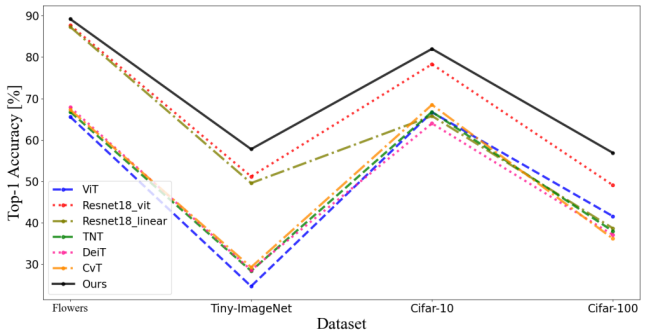


Fig. 4 Top-1 Comparison chart of accuracy

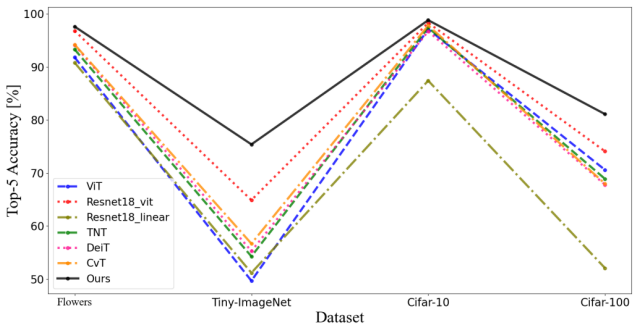


Fig. 5 Top-5 Comparison chart of accuracy

Because the complexity of FFT model is higher than that of Vision Transformer, and there is not pre-training in the model, it is easy to be over-fitting. In order to avoid over-fitting of the model as much as possible, the datasets are enhanced by random horizontal flipping and clipping, and Dropout technology is adopted in the experiment. The training details of FFT model on each dataset are shown in Figure 6.

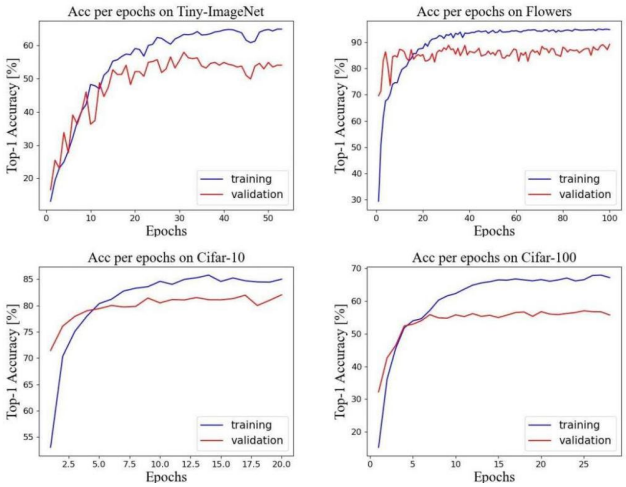


Fig. 6 Training details of FFT model on datasets

D. Ablation Studies

The position coding is very important for image recognition. The FFT module sets the position coding for the embedded sequences of three T-Blocks to save more perfect spatial information. At the same time, the Residual Attention module is also very necessary to distinguish different regions and channels of the image. In the ablation studies, the position coding and Residual Attention module are removed respectively to verify their effects. As shown in Table V, the top-1 accuracy is reduced by about 0.5~2% by removing the position coding and Residual Attention Module respectively.

TABLE V THE INFLUENCE OF REMOVING POSITION CODING AND RAM MODULE

	Flowers	Tiny ImageNet	Cifar10	Cifar100
Ours without ram	88.42%	56.76%	81.25%	55.81%
Ours without position	87.47%	56.91%	81.82%	56.77%
Ours	**89.23%**	**57.81%**	**82.04%**	**56.98%**

E. Visualization

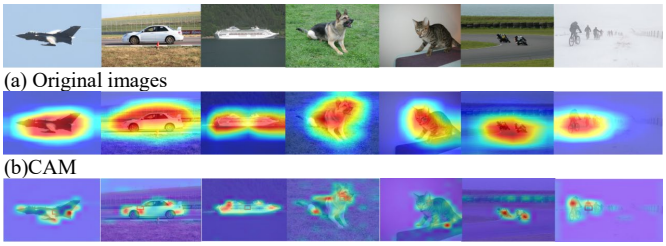
In order to better understand the interaction mode of Query, Key and Value in the self-attention mechanism and the focus area of FFT model on the image. In this paper, the attention map in the FFT model is extracted and visually displayed, as shown in Figure 7, which shows the attention map generated when the red block is used as query. Obviously in the same image, the network pays different attention to different regions of the image when the query is different. As shown in Figure 7(a), the attention of the network is focused on the area of the green leaf in the image when the green leaf is the query, the attention of the network turns to the yellow flower of the image when the flower is used as the Query. As shown in Figure 7(b), when the petal is the Query, the network focuses more on the petal region in the image. When the pistil is used as the Query, the area of interest of the network turns to the pistil. Comparing the class activation maps of the two images in ResNet18, as shown in Figure 8 and 9, it can be seen that the FFT model focuses on more detailed and specific features of the image

during image recognition.

(a) roses (b) daisy

Fig. 7 Examples of attention maps for different Queries

Fig. 8 CAM diagram generated by ResNet18



(a) Original images

(b)CAM

(c)Attention map(ours)

Fig. 9 More sample diagrams

V. Conclusion

In this paper, we propose an FFT framework based on multi-scale feature fusion transformer for image recognition, which combines the channel attention layer and spatial attention layer with residual to emphasize the feature channels and feature regions of interest, and design Inner-T-Block, Middle-T-Block and Outer-T-Block to construct FFT network. The framework mixes the feature representation of different feature layers, retains the structure information of feature patches, which effectively improves the image recognition accuracy. The algorithm is tested on common datasets and compared with mainstream algorithms, which shows good performance. The experimental results verify its effectiveness and feasibility. Next, we plan to optimize the feature extraction process, reduce the complexity of the model, and make the model applicable to more datasets.

Acknowledgment

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

References

[1] Haiping Wu, Bin Xiao, Noel Codella, et al. CvT: Introducing Convolutions to Vision Transformers [J]. arXiv preprint arXiv: 2103.15808, 2021.

[2] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention [C]//International Conference on Machine Learning. PMLR, 2021: 10347-10357.

[3] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers [C]//European Conference on Computer Vision. Springer, Cham, 2020: 213-229.

[4] Xizhou Zhu, Weijie Su, Lewei Lu, et al. Deformable detr: Deformable transformers for end-to-end object detection [J]. arXiv preprint arXiv: 2010.04159, 2020.

[5] Chen H, Wang Y, Guo T, et al. Pre-trained image processing transformer [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 12299-12310.

[6] Kai Han, Yunhe Wang, Hanting Chen, et al. A survey on visual transformer [J]. arXiv preprint arXiv: 2012. 12556, 2020.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv: 2010.11929, 2020.

[8] Mark Chen, Alec Radford, Rewon Child, et al. Generative pretraining from pixels [C]//Proc of the 37th International Conference on Machine Learning. Vienna: ICML Press, 2020.

[9] Josh Beal, Eric Kim, Eric Tzeng, et al. Toward transformer-based object detection [J]. arXiv preprint arXiv: 2012.09958, 2020.

[10] Zheng S, Lu J, Zhao H, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 6881-6890.

[11] Kai Han, An Xiao, Enhua Wu, et al.Transformer in transformer [J]. arXiv preprint arXiv: 2103.00112, 2021.

[12] David G Lowe. Object recognition from local scale-invariant features [C]//Proc of the 7th International Conference on Computer Vision. Kerkyra, Greece: ICCV Press, 1999.

[13] Wieland Brendel, Matthias Bethge. Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet [J]. arXiv preprint arXiv: 1904.00760, 2019.

[14] Li Yuan, Yunpeng Chen, Tao Wang, et al. Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet [J]. arXiv preprint arXiv: 2101.11986, 2021.

[15] Lin M, Chen Q, Yan S. Network in network [J]. arXiv preprint arXiv:1312.4400, 2013.

[16] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[17] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks [C]//International Conference on Machine Learning. PMLR, 2019: 6105-6114.

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

[19] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Advances in neural information processing systems. 2017: 5998-6008.

[20] Qilong Wang, Banggu Wu, Pengfei Zhu, et al. Eca-net: Efficient Channel Attention for Deep Convolutional Neural Networks [J]. arXiv preprint arXiv: 1910.03151, 2019.

[21] Zhu X, Cheng D, Zhang Z, et al. An empirical study of spatial attention mechanisms in deep networks [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 6688-6697.

[22] Wang F, Jiang M, Qian C, et al. Residual attention network for image classification [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 3156-3164.

[23] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[24] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module [C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[25] Adam Paszke, Sam Gross, Francisco Massa, et al. Pytorch: An imperative style, high-performance deep learning library [C]//Proc of the 33th Conference and Workshop on Neural Information Processing Systems. Vancouver, Canada: NeurIPS Press, 2019.