2020 19th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Research on classification algorithms for attention mechanism

Zhuoqun Yang[1,2], Tao Zhang[1,2]
[1]Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources
[2]School of Artificial Intelligence and Computer Science
Jiangnan University
E-mail: yzqstudy@163.com, taozhang@jiangnan.edu.cn

Jie Yang
School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University
Shanghai, China
E-mail: jieyang@sjtu.edu.cn

***Abstract*—In this paper, we solve the image classification task by capturing context dependencies based on spatial and channel attention mechanism. Unlike previous research on feature fusion, we propose an attention module based on spatial and channel dimensions. This module derives attention maps respectively from spatial and channel, then for feature refinement we multiply the attention map into the feature map. Meanwhile, the module can be easily embedded into the network structures due to it is lightweight. The channel attention module selectively enhances some feature channels and suppresses certain feature channels by integrating the relationship between each feature channel. By weighting the features of all locations, spatial attention module aggregates location features. Regardless of distance, similar features are interrelated. Our module is evaluated through experiments on the ImageNet-1K and CIFAR-100 datasets.**

Keywords-classification;channel attention;spatial attention

I. Introduction

Image classification is one of the important parts of computer vision, and it is also an important foundation for realizing applications such as object detection [1], face recognition [2] and pose estimation [3]. Therefore, image classification technology has high academic research and scientific application value. Given an input image, the image classification algorithm can correctly determine the category to which the image belongs. However, due to problems such as huge image data and serious image interference, its classification accuracy cannot meet the actual needs, resulting in traditional classifiers not suitable for classification of complex images. Deep learning is an emerging algorithm for machine learning. It has attracted widespread attention from researchers for its significant effect on image feature learning. Compared with the traditional image classification method, it does not need to describe and extract the artificial features of the target image. Instead, it learns the features from the training samples autonomously through neural networks to extract higher-dimensional and more abstract features. Deep learning algorithms solve the problems of artificial feature extraction and classifier selection.

The network has become deeper due to its rich representation capabilities from the LeNet [4] architecture to ResNet [5]. The experimental result of VGGNet [6] show that the same shape blocks can get better classification accuracy through stacking. With the same idea, ResNet stacked similar treatment of the residual blocks that can conduct jump connections to create a deeper network framework. GoogLeNet [7] increases the adaptability of the network to different scales, showing that band width is very essential to improve the transportability of network models. ResNeXt [8] and Xception [9] improves the details of the network model, indicating that the cardinal feedback can not only reduce the overall parameters of the model, but also have a strong ability to represent.

The Attention mechanism is very similar to the logic of humans looking at pictures. When we look at a picture, we do not see the entire content of the picture, but focus on the focus of the picture. The Attention mechanism allows deep learning models to focus on the most important part of the input data. Wang proposed a convolutional neural network with the attention mechanism, which is named the residual attention network. As the layers get deeper, the features perceived by attention from different modules will adaptively change. Finally, it got 4.8% Top-5 error rate on ImageNet. [10] proposed the SENet network model. The model is based on the interdependence between feature channels.

II. Channel and spatial attention module

We first introduce the general framework of the network model, and then introduce the two attention modules separately. Finally, we will describe how to aggregate them together.

After convolution operations, an intermediate feature map $\mathrm{F_{in}} \in \mathbb{R}^{C\times H\times W}$ is obtained. $\mathrm{F_{in}}$ is the input of model. Channel map $\mathrm{M_c} \in \mathbb{R}^{C\times 1\times 1}$ is inferred after the channel attention module and spatial map $M_\mathrm{s} \in \mathbb{R}^{1\times(H\times W)\times(H\times W)}$ is inferred after the spatial attention module as shown in Fig. 1.The entire attention calculation process is summarized as：

$$F_1 = M_c(F_{in}) \otimes F_{in} \tag{1}$$

$$F_{out} = M_s(F_1) \otimes F_1 \tag{2}$$

where $\otimes$ represent element-wise multiplication. F_1 is the output of F_{in} after passing the channel attention module.F_{out} is the final output and the attention value is broadcasted during multiplication.

2473-3636/20/$31.00 ©2020 IEEE
DOI 10.1109/DCABES50732.2020.00058

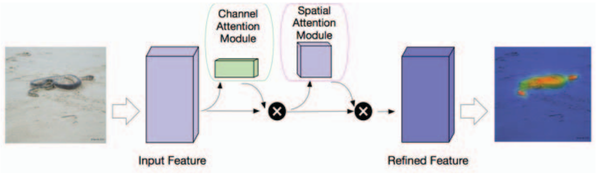

Figure 1. The overview of model

A. Channel attention module

The channel attention module generates attention maps by focusing on the correlation between different channels. Since each channel is considered as a feature detector, this mechanism can make the model pay more attention to the channel features of effective information. Firstly, in order to obtain global features, we perform an extrusion operation on the feature map, then use the global features to learn the relationship between each channel. Finally, we multiply the weight by the original feature map to get the final feature. Next we will introduce the specific operation.

Since the convolution only operates in a local space, it is difficult for the module to obtain enough information to extract the relationship between different channels. This is more serious for the previous layers in the network because the receptive field is relatively small. Therefore, we encode the entire spatial feature on a channel as a global feature, using global average pooling to achieve. The pooling process is computed as:

$$F_{avg} = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} F_{in}\,(i,j) \qquad (3)$$

where $F_{avg} \in \mathbb{R}^{C\times 1\times 1}$ is the result of global average pooling with the input feature map F_{in}.

After obtaining the global description characteristics, we need another operation to capture the relationship between different channels. This operation needs to meet two criteria: first, it must be flexible because it needs to learn the non-linear relationship between various channels; the second point is that the learning relationship is not mutually exclusive, because multi-channel features are allowed here instead of one-hot form. The feature map $\mathrm{M_c}$ is computed as:

$$\mathrm{M_c} = \sigma\left(W_2 ReLU\left(W_1 F_{avg}\right)\right) \qquad (4)$$

where $W_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $W_2 \in \mathbb{R}^{C\times \frac{C}{r}}$, σ denotes Sigmoid function. W_1 and W_2 are fully connected layers. In order to reduce the complexity of the model, a bottleneck structure containing two fully connected layers is used here.W_1 layer plays the role of dimensionality reduction, and the dimensionality reduction factor r is a hyperparameter. Then we use ReLU function to activate it and the final W_2 layer restores the dimensions to the original. The process is shown in Fig. 2.

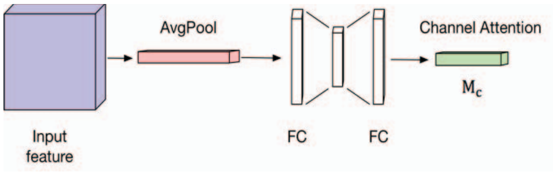

Figure 2. The details of channel attention module

B. Spatial attention module

As shown in Fig. 3, $A \in \mathbb{R}^{C\times H\times W}$ is the input of the spatial module. A generates feature maps B, C and D after convolutional layers, where $\{B,C,D\} \in \mathbb{R}^{C\times H\times W}$. We reshape B and C to $\mathbb{R}^{C\times(H\times W)}$, and $H\times W$ represents pixels in spatial. Then we get $\mathbb{R}^{(H\times W)\times(H\times W)}$ by doing a matrix multiplication between the transpose of B and C. After applying a softmax layer, we get the spatial attention map M_s.

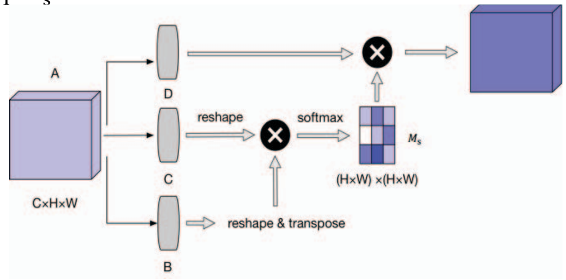

Figure 3. The details of spatial attention module

M_s is computed as:

$$M_{s_{ij}} = \frac{exp(B_i\times C_j)}{\sum_{i=1}^{H\times W} exp(B_i\times C_j)} \qquad (5)$$

where the more similar the feature representation of the two positions, the stronger the correlation between them. And Fig. 4 show the visualization results.

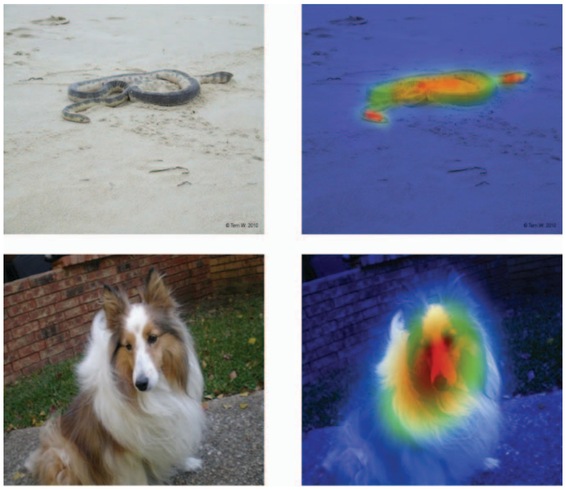

Figure 4. Diagram of the visualization results

It can be seen from Fig. 4 that the model focuses on important information, which saving resources and reducing computing time.

III. EXPERIMENTS

A. Ablation studies about different attention module

The ablation experiment verified the effectiveness of the two attention modules separately. We compared 4 different network models: baseline, baseline with the channel attention module(CAM), baseline with the spatial attention module(SAM), baseline with the channel and spatial attention model(CSM). Top-1 error is shown in Fig. 5.

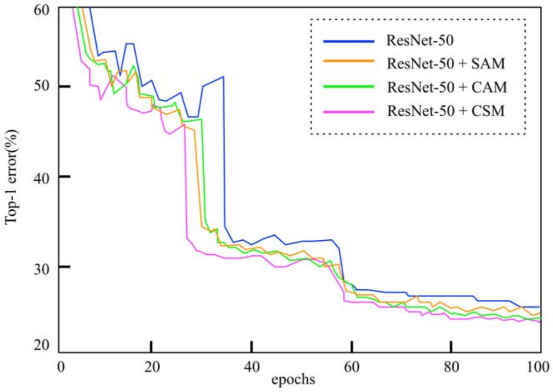

Figure 5. Comparison of different network models

From Fig. 5, we can see that every attention module can make the model performance better, proving the attention module is valid. At the same time, through comparison we find that CAM is more effective than SAM. In addition, the combination of two attention modules can promote further performance. This shows that the fusion of channel features and spatial features can improve the representation ability of the model. Specific results are illustrated in Table 1.

TABLE I. COMPARISON OF DIFFERENT NETWORK MODELS

Architecture	Top-1 error (%)
ResNet-50	24.55
ResNet-50+SAM	23.47
ResNet-50+CAM	23.21
ResNet-50+CSM	22.78

B. Ablation studies about the order of attention module

We want to know if the order of attention modules will affect performance. So in this ablation experiment, we compare the effects of the combination of CAM and SAM: baseline with CAM and SAM, baseline with SAM and CAM. The curve on test set is shown in Fig. 6.

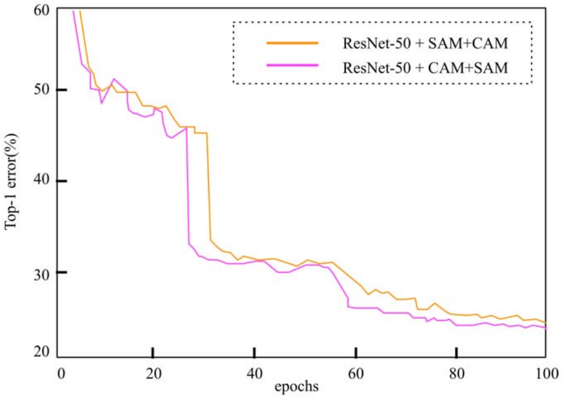

Figure 6. Comparison of different combination method

From Fig. 6, we can see that the CAM-first combination method has achieved higher accuracy, showing that focusing on channel features first can improve the model's representation capabilities.

C. Image Classification on ImageNet-1K

We perform extensive image classification experiments on the ImageNet-1K dataset. We adopt ResNet and WideResNet as the baseline model. Results of the experiment on the ImageNet-1K dataset are illustrated in Table 2.

TABLE II. RESULTS OF THE EXPERIMENT ON THE IMAGENET-1K DATASET

Architecture	Top-1 error (%)	Top-5 error (%)
ResNet-18	29.62	10.56
ResNet-18+SE	29.43	10.24
ResNet-18+CSM	29.29	10.12
ResNet-34	26.71	8.62
ResNet-34+SE	26.16	8.37
ResNet-34+CSM	26.03	8.28
ResNet-50	24.55	7.52
ResNet-50+SE	23.21	6.74
ResNet-50+CSM	22.78	6.57
WideResNet18(widen=1.5)	26.86	8.91
WideResNet18(widen=1.5) +SE	26.23	8.51
WideResNet18(widen=1.5) +CSM	26.14	8.49

The results show that networks with CSM performs better than other models, indicating that CSM can be well generalized to various network models. SENet won the ILSVRC2017 classification task championship. CSM is better than SENet indicates the effectiveness of the module. CSM fuses channel features with spatial features for better representation capabilities.

D. Image Classification on CIFAR-100

The CIFIR-100 dataset contains 100 classification categories and we perform image classification experiments on the model. ResNet and WideResNet are adopted as baseline. Results of the experiment on the CIFIR-100 dataset are illustrated in Table 3.

TABLE III. RESULTS OF THE EXPERIMENT ON THE CIFIR-100 DATASET

Architecture	Accuracy (%)
ResNet-18	91.7
ResNet-18+CSM	93.1
ResNet-34	92.4
ResNet-34+CSM	93.8
ResNet-50	92.9
ResNet-50+CSM	94.3
WideResNet18(widen=1.5)	92.8
WideResNet18(widen=1.5) +CSM	94.2

The experimental results prove that the module can improve classification accuracy.

IV. CONCLUSION

Unlike previous research on feature fusion, we propose an attention module based on spatial and channel dimensions. This module derives attention maps respectively from spatial and channel, then for feature refinement we multiply the attention map into the feature map. The channel attention module selectively enhances some feature channels and suppresses certain feature channels by integrating the relationship between each feature channel. By weighting the features of all locations, spatial attention module aggregates location features. To verify the effectiveness of the module, we perform extensive image classification experiments on the ImageNet-1K and CIFIR-100 datasets. The results show that baseline with the attention has higher classification accuracy and better fitting effect.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (NO. 61702226); the 111 Project (B12018); the Natural Science Foundation of Jiangsu Province (NO. BK20170200); Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Land and Resources (NO. KF-2018-03-065); the Fundamental Research Funds for the Central Universities (NO. JUSRP11854, NO. JUSRP11851).

REFERENCES

[1] DIBA A, SHARMA V, PAZANDEH A, et al. Weakly super- vised cascaded convolutional networks[C]//IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: ACM Press, 2017: 5131-5139.

[2] HU G, YANG Y X, YI D, et al. When face recognition meets with deep learning: an evaluation of convolutional neural net- works for face recognition[C]//International Conference on Computer Vision, December 11-18, 2015, Santiago, Chile. Pis- cataway: IEEE Press, 2015: 142-150.

[3] CAO Z, SIMON T, WEI S, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]//IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. EprintArxiv, 2017: 1302-1310.

[4] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE 86(11) (1998) 2278–2324

[5] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition[C]. In: Proc. of Computer Vision and Pattern Recognition (CVPR). (2016)

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

[7] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions[C]. In: Proc. of Computer Vision and Pattern Recognition (CVPR). (2015)

[8] Xie,S.,Girshick,R.,Dollár,P.,Tu,Z.,He,K.: Aggregated residual transformations for deep neural networks[J]. arXiv preprint arXiv:1611.05431 (2016)

[9] Chollet, F.: Xception: Deep learning with depthwise separable convolutions[J]. arXiv preprint arXiv:1610.02357 (2016)

[10] HU J, SHEN L, SUN G. Squeeze-and-excitation net- works[C]//IEEE Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, New York, USA. Piscataway: IEEE Press, 2018: 7132-7141