Context-Aware Residual Module for Image Classification

Jing Bai[1] Ran Chen[1]

North Minzu University, Yinchuan 750021, China

Abstract

Attention module has achieved great success in numerous vision tasks. However, existing visual attention modules generally consider the features of a single-scale, and cannot make full use of their multi-scale contextual information. Meanwhile, the multi-scale spatial feature representation has demonstrated its outstanding performance in a wide range of applications. However, the multi-scale features are always represented in a layer-wise manner, i.e. it is impossible to know their contextual information at a granular level. Focusing on the above issue, a context-aware residual module for image classification is proposed in this paper. It consists of a novel multi-scale channel attention module MSCAM to learn refined channel weights by considering the visual features of its own scale and its surrounding fields, and a multi-scale spatial aware module MSSAM to further capture more spatial information. Either or both of the two modules can be plugged into any CNNbased backbone image classification architecture with a short residual connection to obtain the context-aware enhanced features. The experiments on public image recognition datasets including CIFAR10, CIFAR100， Tiny-ImageNet and ImageNet consistently demonstrate that our proposed modules significantly outperforms a wide-used state-of-the-art methods, e.g., ResNet and the lightweight networks of MobileNet and SqueezeNet.

Introduction

Convolutional neural networks (CNNs) have been widely used in many vision tasks and made significant advances in these tasks with the state-of-the-art performance. One of the key factors of its success application is the natural ability of learning coarse-to-fine multi-scale features through layers of convolutional operators. However, most of currently CNN architectures only represent multi-scales features in layer-wise manner and treat these features equally, which limits the further improvement of CNNs.

Attention mechanisms make it possible to focus more on specific parts or specific features of the whole feature space as needed, and have play important roles in modern CNNs especially in computer vision tasks. Due to their ability to make fine distinctions of "what features are important" and emphasize "which regions are important", these networks have achieved improved object recognition performance. However, all these methods consider attention mechanisms only on single-scale visual perception fields. Usually, visual patterns occur at multi-scales in natural scenes, i.e. we need to answer what and where is important for a feature map by itself and its surrounding contextual information from different scales. For example, when a task is recognizing a cat, whether or not a circle feature is meaningful relies on it is within a cat face-like region or a cup-like region. Actually, multi-scale information has been widely used in deep learning. Earlier CNNs learn multi-scale features in series by coarse-to-fine layers of convolutional operators[1-5]. And then one kind of networks [13-16] propose to capture multi-scale features in parallel based on multi-branch. Another kind of networks propose to use multi-scale kernel for enlarging receptive fields [17-18]. These different forms of multi-scale representation have achieved outstanding performance in visual recognition, speech recognition et al and demonstrated the powerful ability on recognition tasks. Inspired by above work, in this paper, we propose a generic and flexible multi-scale context-aware residual module, which can be plugged into existing backbone image classification architectures to obtain the context-aware enhanced features.

Network Design

A. MSCAM: multi-scale channel attention module

Fig.1 (a) shows the traditional channel attention module (CAM) achieved by a squeeze-and-excitation block [10]. It can be seen that CAM produces a channel attention map by exploiting the inter-channel relationships of features. Here, it only focuses on single visual perception field. However, visual patterns occur at multi-scales in natural scenes [19], and we need to understand and judge "what" is meaningful for an input image by perceiving from different scales. For instance, we need to rely on the face as context to better tell whether the almond shaped object is an eye or a leaf and whether it is meaningful or not. Accordingly, we propose a novel multiscale channel attention module MSCAM to learn refined channel weights by considering the visual features with multiscale context information.

As Fig.1 (b) shows, given a feature x as input (here, c is the number of channels, and h and w are the height and width of the feature map, respectively), MSCAM constructs its multi-scale channel attention feature by the following steps:

Step 1. A varying-size pyramid pooling operation [20] is introduced so as to abstract different sub-regions, then fuse features under different pyramid scales and obtain its context information. The coarsest level is a h/2×w/2 averaging pooling to generate large-scale surrounding information, while the following pyramid level is a h/4×w/4 averaging pooling to generate its surrounding information, and the last level is a copy of input feature. The multi-stage kernels of pooling size H/2 and H/4 can maintain a reasonable gap in representation, and the above three different levels pyramid pooling operations make it possible to not only preserve original feature but also obtain its context information.

Step 2. Three squeeze-and-excitation operations [10] are applied to the three scales pyramid pooling outputs so as to calculate channel weights and extract different scales refined features.

Step 3. A combination operation is used to merge different scales refined features by up-sampling and concatenation operations.

Step 4. The output, a multi-scale channel attention feature map x , is obtained by convolving with a standard 1*1 convolution layer.

B. MSSAM: multi-scale spatial aware module

Furthermore, referring to multi-scale residual block MSRB [17], a multi-scale spatial aware module MSSAM is designed by introducing dilated convolutions, so as to effectively detect multi-scale spatial features.

As Fig.2 shown, MSSAM stacks two consecutive contextual combinatorial blocks (CCB) with a sigmoid function. Here, CCB is designed to capture large-scale spatial information by learning contextual combinatorial features of large receptive fields. Obviously, increasing kernel sizes can be used to generate large receptive fields. However, increasing kernel sizes also means increasing the memory computation. To avoid this problem, dilated convolutions are introduced in CCB to enlarge the receptive fields and make the convolution output contain a large range of information. Specifically, in CCB, we introduce two parallel dilated convolution layers with different dilation factors to capture different scale spatial features, then concatenate these multi-scale features to increase the channel sizes, and at the end use 1×1 convolution to make the channels number the same as the input.

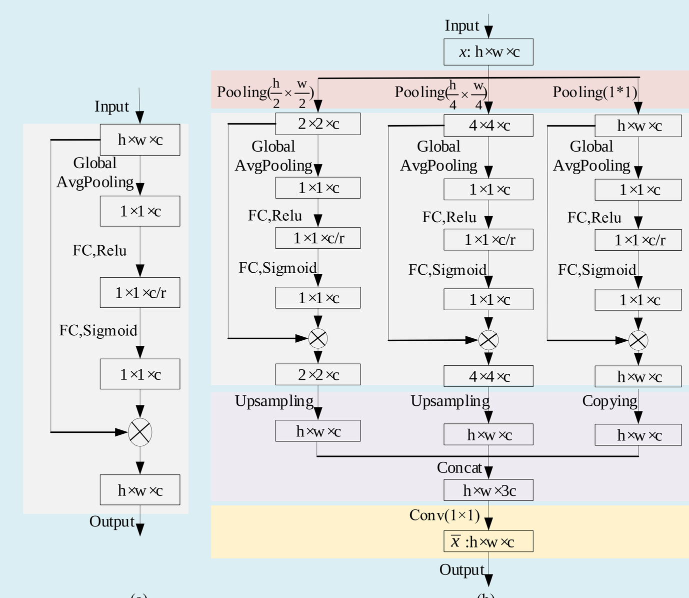

Fig. 1. (a) traditional channel attention module (CAM) and (b) our proposed multi-scale channel attention module

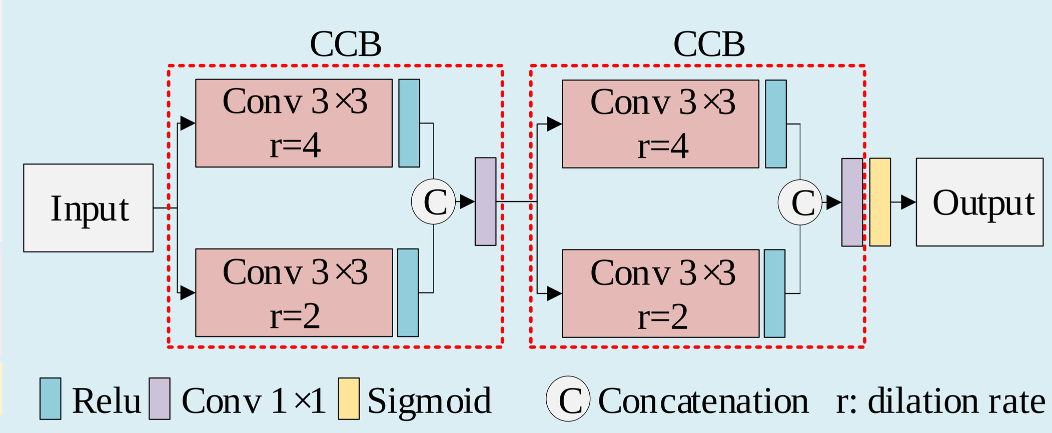

Fig. 2. Illustration of multi-scale spatial aware module (MSSAM).

Experiments

Baselines. We use Residual Networks with depth from 18 to 101 and Xception as the baseline networks in view of their state-of-the-art performance in image classification. Meanwhile, in order to test the adaptability of our module more comprehensively, we also choose some typical lightweight networks including MobileNet and SqueezeNet (SqueezeNet with simple bypass) as the baseline networks..

Datasets. CIFAR10. This dataset is composed of 60,000 colored natural images with 32×32 pixels, evenly divided into 10 classes. There are 5000 images per class in the training set,and 1000 images per class in the test set.

CIFAR100 . This dataset is also composed of 60,000 colored natural images with 32×32 pixels, evenly divided into 100 classes on average. There are 500 images per class in the training set, and 100 images per class in the test set.

Tiny-ImageNet. This dataset is a modified subset of the original ImageNet dataset. It is composed of 110,000 images with 64×64 pixels and 200 classes, and for each class 500 images are for training and 50 images are for testing.

ImageNet. This dataset is a public large dataset, which contains 1.3 million training images, 50,000 validation images, and 100,000 test images with 1,000 classes. Following SENet[10]and CBAM, a 224×224 crop is randomly sampled from the images or its horizontal flip. The images are normalized into [0 1] using mean values and standard deviations. We report the single-crop error rates on the validation set.

It obviously that the number, type and resolution of images in the above four datasets are different, so the classification experiments based on them are still very convincing.

Implementation Details. The proposed modules can be plugged into any CNN architectures. In the experiments, we plug CAM, CBAM and our modules between two conv blocks for baseline networks respectively. According to the total number of images, we use SGD with a mini-batch size of 128 for CIFAR dataset and 256 for Tiny-ImageNet and ImageNet. The learning rate starts from 0.1 and is divided by 10 every 60 epochs for CIFAR100, and every 30 epochs for other datasets considering the number of images in each class.

The network architectures of MSCAM(MSSAM) + ResNet, SqueezeNet and MobileNet are shown as TABLE I, II and III, respectively. Specially, the network Xception contains 3 flows: entry flow， middle flow and exit flow. Considering that all of three flows have their own design philosophy, we only add attention modules at the end of each flow to avoid destroying the network structure.

TABLE V. ACCURACY RATES (%) AND PARAMTER QUANTITY (M) COMPARISONS OF STATE-OF-THE-ART NETWORKS ON VARIOUS DATASETS

Network	CIFAR10 Top-1 err. (%)	CIFAR10 params. (M)	CIFAR100 Top-1 err. (%)	CIFAR100 params. (M)	Tiny-ImageNet Top-1 err. (%)	Tiny-ImageNet Top-5 err. (%)	Tiny-ImageNet params. (M)
ResNet18	7.36	**11.17**	25.07	**11.22**	44.87	21.10	**11.28**
ResNet18+CAM[10]	7.32	11.22	24.42	11.26	44.41	20.45	11.32
ResNet18+CBAM[11]	7.32	11.22	24.84	11.27	44.10	20.50	11.32
ResNet18+MSCAM	7.07	12.35	23.55	12.40	43.72	20.61	12.45
ResNet18+MSC&SAM	**7.02**	26.63	**23.41**	26.68	**43.62**	**20.29**	26.73
ResNet34	7.36	**21.28**	23.90	**21.33**	44.13	20.40	**21.39**
ResNet34+CAM[10]	7.09	21.33	23.08	21.37	44.41	21.12	21.42
ResNet34+CBAM[11]	**6.60**	21.33	23.20	21.37	44.14	20.95	21.42
ResNet34+MSCAM	6.77	22.46	22.62	22.51	43.86	20.33	22.56
ResNet34+MSC&SAM	6.92	36.74	**22.59**	36.79	**43.53**	**20.33**	36.84
ResNet50	6.86	**23.52**	23.13	**23.71**	46.17	22.95	**23.92**
ResNet50+CAM[10]	6.61	24.22	22.23	24.41	47.83	23.65	24.61
ResNet50+CBAM[11]	7.35	24.22	21.88	24.41	46.45	22.92	24.61
ResNet50+MSCAM	**6.44**	42.34	21.46	42.52	43.68	20.60	42.73
ResNet50+MSC&SAM	6.57	270.75	**21.33**	270.9	**42.58**	**19.33**	271.1
ResNet101	7.78	**42.51**	21.90	**42.70**	43.41	20.40	**42.91**
ResNet101+CAM[10]	6.97	43.21	22.44	43.40	43.83	20.28	43.60
ResNet101+CBAM[11]	7.41	43.21	21.67	43.40	44.29	21.17	43.60
ResNet101+MSCAM	**6.95**	61.33	**21.33**	61.51	43.81	20.28	61.72
ResNet101+MSC&SAM	6.99	289.75	21.83	290.0	**42.36**	**19.41**	290.1
Xception	8.47	**20.83**	23.82	**21.01**	46.86	21.27	**21.22**
Xception+CAM[10]	**7.78**	21.49	23.64	21.67	40.28	18.45	21.88
Xception+CBAM[11]	7.94	21.49	23.64	21.67	40.72	18.65	21.88
Xception+MSCAM	8.14	38.57	**23.49**	38.75	39.96	18.06	38.96
Xception+MSC&SAM	8.15	254.02	23.64	254.2	**39.74**	**17.59**	254.4

TABLE VI. ACCURACY RATES (%) COMPARISONS OF LIGHTWEIGHT NETWORKS ON VARIOUS DATASETS

Network	CIFAR10 Top-1 err. (%)	CIFAR10 params. (M)	CIFAR100 Top-1 err. (%)	CIFAR100 params. (M)	Tiny-ImageNet Top-1 err. (%)	Tiny-ImageNet Top-5 err. (%)	Tiny-ImageNet params. (M)
SqueezeNet	11.66	**0.73**	31.09	**0.78**	46.44	21.43	**0.83**
SqueezeNet+CAM[10]	11.40	0.86	29.61	0.91	45.33	20.52	0.91
SqueezeNet+CBAM[11]	11.81	0.86	31.55	0.91	46.87	21.25	0.91
SqueezeNet+MSCAM	**10.01**	1.84	**29.19**	1.89	**43.54**	**19.19**	1.94
MobileNetV1	13.63	**3.22**	34.57	3.32	42.31	19.66	**3.43**
MobileNetV1+CAM[10]	9.94	3.40	34.73	3.49	43.78	20.20	3.59
MobileNetV1+CBAM[11]	12.53	3.40	33.89	3.49	43.27	19.84	3.59
MobileNetV1+MSCAM	**9.90**	7.93	**32.91**	8.02	**42.11**	**19.03**	8.13

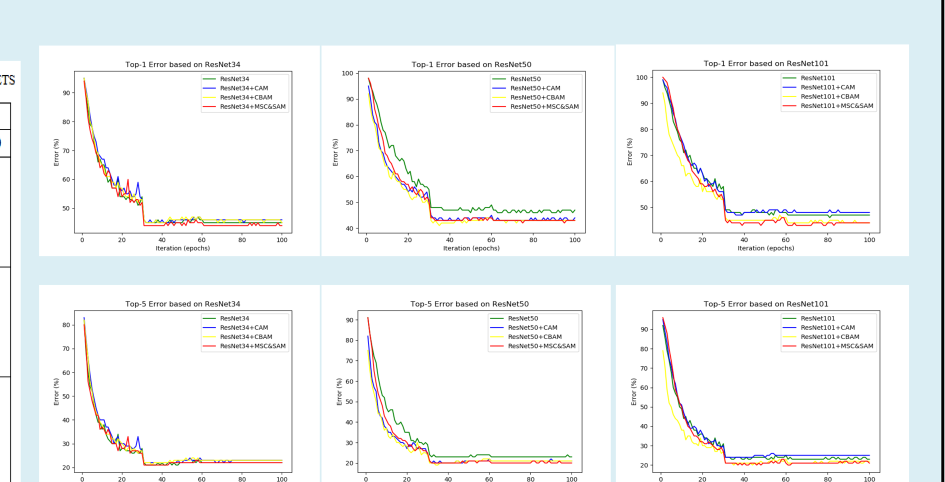

Fig. 4. Error curves during training process on Tiny-ImageNet dataset

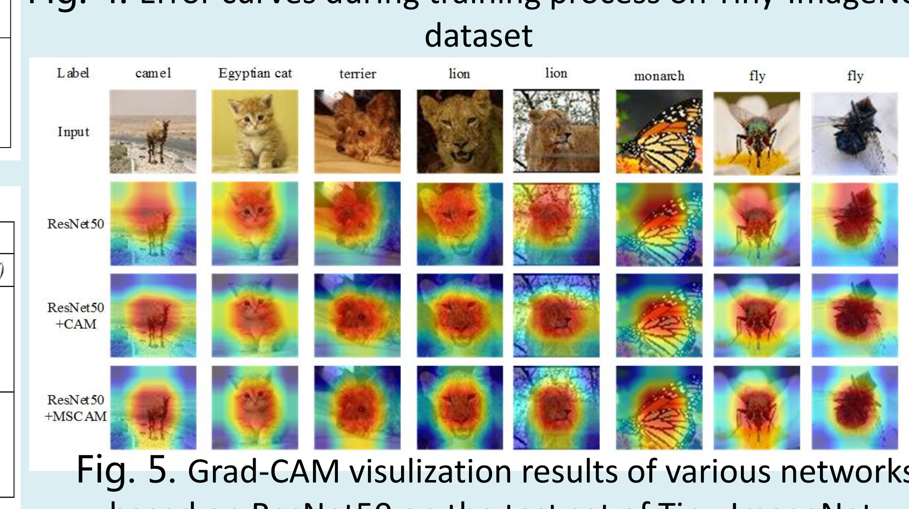

Fig. 5. Grad-CAM visulization results of various networks based on ResNet50 on the test set of Tiny-ImageNet.

Conclusion

The experiments on the state-of-the-art networks and lightweight networks demonstrate the effectiveness of the proposed modules. Currently, we only apply our proposed modules in classification tasks, while future work we will use them for exploring other vision tasks, such as segmentation, detection and so on. In addition, we also will focus on setting up more lightweight MSCAM for lightweight networks.

Our Contributions

(1) A novel Multi-Scale Channel Attention Module MSCAM is proposed to learn refined channel weights by considering the visual features of their own scale and their surrounding fields.

(2) A Multi-Scale Spatial Aware Module MSSAM is designed to further capture its multi-scale contextual information at a granular level, which can be combined with MSCAM and then plugged into any CNN-based backbone image classification architecture (alone or in combination) with a short residual connection to obtain the context-aware enhanced features.

(3) The proposed module is evaluated on a number of public datasets and achieves better results than a wide-used state-of-the-art methods, including ResNet, Xception and the lightweight networks of MobileNet and SqueezeNet. For example, both the accuracy rate and the parameter quantity of the ResNet50+MSCAM are superior to ResNet101.