Context-Aware Residual Module for Image Classification

Jing Bai, Ran Chen

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 CNN-based 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 outperform a wide variety of 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, we need to answer what and where is important for a feature map by itself and its surrounding context 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 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 enlarge receptive fields (17-18). These different forms of multi-scale representation have achieved outstanding performance in visual recognition, speech recognition 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

As Fig. 2(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 exploring the inter-channel relationships of features. Here, we only focus 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 one eye or a leaf and whether it is meaningful or not. Accordingly, we propose a novel multi-scale channel attention module MSCAM to learn refined channel weights by considering the visual features with multi-scale context information. As Fig. 3(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 $2 \times 2 \times c$ averaging pooling to generate large-scale surrounding information, while the following pyramid level is a $4 \times 4 \times c$ averaging pooling to generate its surrounding information, and the last level is a copy of input feature. The multi-stage kernels of pooling size $2 \times 2$ and $4 \times 4$ can maintain a reasonable gap in representation, and the above three different levels pyramid pooling to generate 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 CBLC to enlarge the receptive fields and make the convolution output contain a large range of information. Specifically, in CBLC, 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 $1 \times 1$ convolution to make the channels number the same as the input.

Conclusion

The experiments on the state-of-the-art networks and lightweight networks demonstrate the effectiveness of the proposed modules. Currently, we only plug 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 datasets and produces better results than a wide range of 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.

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).

Fig. 3. Schematic diagram of the attention module.

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

Fig. 5. Dual CAM visualization results of our two networks based on ResNet50 on the test set of Tiny-ImageNet dataset.