Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

Zhang, Liqiang

July 4, 2018

Abstract

Now, existing convolutional neural networks (CNNs) have a requirement for the size of input image (e.g., 224×224). This requirement may influnce the recognition accuracy for the images or sub-images of an arbitrary size. In this paper, the author propose a networks with a more principled pooling strategy, "spatial pyramid pooling", to eliminate the above requirement. This new network stucture they propose is called SPP-net, which can generate a fixed-lenth representation regardless of image size. Because this network can remove the fixed-size limitation, it has a excelent performance in all CNN-based image classification methods in general. SPP-net also achieves state-of-the-art accuracy on the datasets of ImageNet 2012, Pascal VOC 2007, and Caltech101.

The power of SPP-net is more significant in object detection. SPP-net can compute the feature maps from the entire image only once, and then pool features in arbitrary regions (sub-images) to generate fixed-length representations for training the detectors. This method avoids repeatedly computing the convolutional features. In processing test images, this method computes convolutional features 30-170× faster than the recent leading method R-CNN, while achieving better or comparable accuracy on Pascal VOC 2007.

1. Itroduction

Deep-networks-based approaches have recently been substantially improving upon the state of the art in image classification [4], object detection [2], many other recognition tasks [6], and even non-recognition tasks.

However, there is a technical issue in the training and testing of the CNNs: the prevalent CNNs require a fixed input image size (e.g., 224×224), which limits both the aspect ratio and the scale of the input image. When applied to images of arbitrary sizes, current methods mostly fit the input image to the fixed size, either via cropping [4] or via warping, as shown in Fig. 1 (top). But the cropped region may not contain the entire object, while the warped content may result in unwanted geometric distortion. Recognition

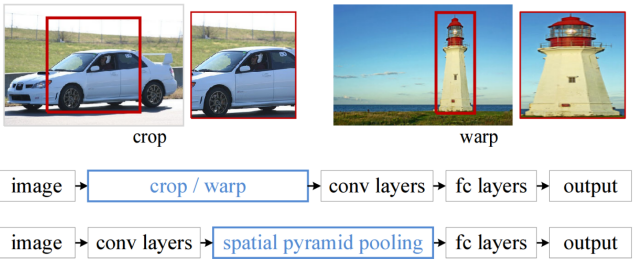

Figure 1. Top: cropping or warping to fit a fixed size. Middle: a conventional deep convolutional network structure. Bottom: the spatial pyramid pooling network structure.

accuracy can be compromised due to the content loss or distortion. Besides, a pre-defined scale (e.g., 224) may not be suitable when object scales vary. Fixing the input size overlooks the issues involving scales.

The reason why CNNs require a fixed input size is that a CNN mainly consists of two parts: convolutional layers, and fully-connected layers that follow. The convolutional layers operate in a sliding-window manner and output feature maps which represent the spatial arrangement of the activations (Fig. 2). In fact, convolutional layers do not require a fixed image size and can generate feature maps of any sizes. On the other hand, the fully-connected layers need to have fixed-size/length input by their definition. Hence, the fixed-size constraint comes only from the fully-connected layers, which exist at a deeper stage of the network.

In this paper, the atuhor introduce a *spatial pyramid pooling* (SPP) layer to remove the fixed-size constraint of the network. Specifically, this team add an SPP layer on top of the last convolutional layer. The SPP layer pools the features and generates fixed-length outputs, which are then fed into the fully-connected layers (or other classifiers). In other words, the author perform some information "aggregation" at a deeper stage of the network hierachy to avoid the need for cropping or warping at the beginning. Fig. 1 (bottom) shows the change of the network architecture by introducing the SPP layer. The author call the new network

structure SPP-net.

2. SPP-net for Image Classification

The author trained this network on the 1000-category training set of ImageNet 2012. the details follow the practices of previous work [1]. The images are resized so that the smaller dimension is 256, and a 224×224 crop is picked from the center or the four corners from the entire image. The data are augmented by horizontal flipping and color altering. Dropout is used on the two fully-connected layers. The learning rate starts from 0.01, and is divided by 10 (twice) when the error plateaus.

Table 1 (e2)(e3) show the results using single-size training and (e4) shows the our result using multi-size training. The training sizes are 224 and 180, while the testing size is still 224. (e4) still use the 10 cropped views for prediction. The top-1 error drops to 34.60%. Note the networks in (e3) and (e4) have exactly the same structure and the same method for testing. So the gain is solely because of the multi-size training.

References

[1] D. Forsyth. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 32(9):1627–1645, 2010. 2

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1

[3] A. G. Howard. Some improvements on deep convolutional neural network based image classification. *arXiv preprint arXiv:1312.5402*, 2013. 3

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012. 1, 3

[5] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. Lecun. OverFeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013. 3

[6] Y. Taigman, M. Yang, Marc, and L. Wolf. DeepFace: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 1

[7] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *arXiv preprint arXiv:1311.2901*, 2014. 3

Figure 2. Visualization of the feature maps. (a) Two images in Pascal VOC 2007. (b) The feature maps of some conv 5 filters. The arrows indicate the strongest responses and their corresponding positions in the images. (c) The ImageNet images that have the strongest responses of the corresponding filters. The green rectangles mark the receptive fields of the strongest responses.

method	test scale	test views	top-1 val	top-5 va
(a) Krizhevsky *et al.* [4]	1	10	40.7	18.2
(b1) Overfeat (fast) [5]	1	-	39.01	16.97
(b2) Overfeat (fast) [5]	6	-	38.12	16.27
(b3) Overfeat (fast) [5]	4	-	35.74	14.18
(c1) Howard (base) [3]	3	162	37.0	15.8
(c1) Howard (high-res) [3]	3	162	36.8	16.2
(d1) Zeiler & Fergus (ZF) (fast) [7]	1	10	38.4	16.5
(d2) Zeiler & Fergus (ZF) (fast) [7]	1	10	37.5	16.0
(e1) The impl of ZF (fast)	1	10	35.99	14.76
(e2) PP-net$_4$, single-size trained	1	10	35.06	14.04
(e3) PP-net$_6$, single-size trained	1	10	34.98	14.14
(e4) PP-net$_6$, single-size trained	1	10	34.60	13.64
(e5) PP-net$_6$, single-size trained	1	8+2full	**34.16**	**13.57**

Table 1. Error rates in the validation set of ImageNet 2012. All the results are based on **a single network**. The number of views in Overfeat depends on the scales and strides, for which there are several hundreds at the finest scale.