The Treasure beneath Convolutional Layers: Cross-convolutional-layer Pooling for Image Classification

Lingqiao Liu [1], Chunhua Shen [1,2], Anton van den Hengel [1,2]
[1] School of Computer Science, University of Adelaide, Australia. [2] Australian Centre for Robotic Vision

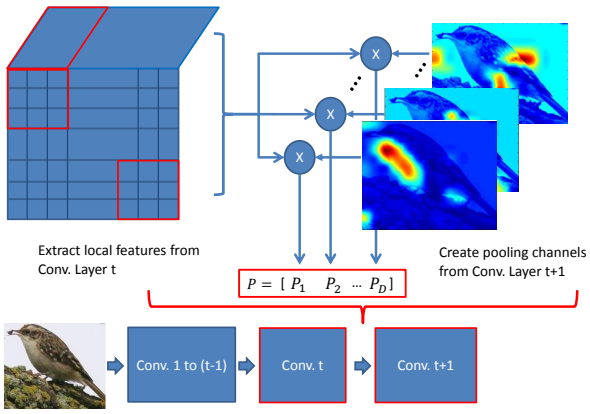

Figure 1: The overview of the proposed method.

A number of recent studies have shown that a Deep Convolutional Neural Network (DCNN) pretrained on a large dataset can be adopted as a universal image descriptor, and that doing so leads to impressive performance at a range of image classification tasks. Most of these studies, if not all, adopt activations of the fully-connected layer of a DCNN as the image or region representation and it is believed that convolutional layer activations are less discriminative.

This paper, however, advocates that if used appropriately convolutional layer activations can be turned into a powerful image representation. This is achieved by adopting a new technique proposed in this paper called cross-convolutional-layer pooling. More specifically, it consists of two major components: (1) it extracts subarrays of feature maps from the t-th convolutional layer as local features $\{\mathbf{x}_i^t\}$. (2) instead of performing pooling on $\{\mathbf{x}_i^t\}$, we use the D_{t+1} filter responses in the $t+1$-th convolutional layer to create D_{t+1} spatial weighting schemes. Then the pooling operation is applied D_{t+1} times with each time corresponding to a weighting scheme. The final image representation is obtained by concatenating these D_{t+1} pooling results. The overview of the proposed method is shown in Figure 1.

Formally, the above procedure is expressed as follows:

$$\mathbf{P}^t = [\mathbf{P}_1^{t\top}, \mathbf{P}_2^{t\top}, \cdots, \mathbf{P}_k^{t\top}, \cdots, \mathbf{P}^{t\top}_{D_{t+1}}]^\top$$
$$\text{where,}\quad \mathbf{P}_k^t = \sum_{i=1}^{N_t} \mathbf{x}_i^t a_{i,k}^{t+1}, \tag{1}$$

where $\mathbf{P}^t$ denotes the pooled feature for the t-th convolutional layer, which is calculated by concatenating the pooled feature of each pooling channel $\mathbf{P}_k^t, k = 1, \cdots, D_{t+1}$. $\mathbf{x}_i^t$ denotes the ith local feature in the tth convolutional layer. Note that feature maps of the $(t+1)$th convolutional layer is obtained by convolving feature maps of the tth convolutional layer with a $m \times n$-sized kernel. So if we extract local features $\mathbf{x}_i^t$ from each $m \times n$ spatial units in the tth convolutional layer then each $\mathbf{x}_i^t$ naturally corresponds to a spatial unit in the $(t+1)$th convolutional layer. Let's denote the feature vector in this spatial unit as $\mathbf{a}_i^{t+1} \in \mathbb{R}^{D_{t+1}}$ and the value at its kth dimension as $a_{i,k}^{t+1}$. Then we use $a_{i,k}^{t+1}$ to weight local feature $\mathbf{x}_i^t$ in the kth pooling channel.

Compared with exising methods that apply DCNNs in the local feature setting, the main advantage of the proposed method is that it avoids the input image style mismatching issue which is usually encountered when applying fully connected layer activations to describe local regions. Also, the proposed method is easier to implement since it is codebook free and does not have any tuning parameters.

Table 1: Comparison of results on Birds-200. Note that methods with "use parts" mark require parts annotations and detection while our methods (those marked with "*") do not employ these annotations so they are not directly comparable with us. CL-45 and CL-45F use different convolutional layer sizes, CL-45 uses 13×13 and CL-45F uses 26×26.

Methods	Accuracy	Remark
CNN-Global	59.2%	no parts.
CNN-Jitter	60.5%	no parts
R-CNN SCFV [3]	66.4%	no parts
*CL-45	72.4%	no parts, 13×13
*CL-45F	68.4%	no parts, 26×26
*CL-45C	**73.5%**	no parts, CL-45 + CF-45F
GlobalCNN-FT [2]	66.4 %	no parts, fine tunning
Parts-RCNN-FT [4]	76.37 %	use parts, fine tunning
Parts-RCNN [4]	68.7 %	use parts, no fine tunning
CNNaug-SVM [5]	61.8%	-
CNN-SVM [5]	53.3%	CNN global
DPD+CNN [1]	65.0%	use parts
DPD [6]	51.0%	-

We evaluate the proposed method on four popular visual classification datasets, MIT-67, Birds-200, PASCAL VOC 2007 and the H3D human attribute dataset. Through our evaluation, it is demonstrated that the proposed representation can achieve comparable or in some cases significantly better performance than existing fully-connected layer based image representations. Our method performs particularly well on the fine-grained image classification problem. As shown in Table 1, our best method achieves 73.5% classification accuracy on the Birds-200 dataset without using any parts annotation. Also, our best method achieves 71.5%, 77.8%,78.3% classification accuracy on MIT-67, PASCAL VOC07, and H3D human action recognition datasets respectively.

We also compare our method against other possible ways of using convolutional level features and our experimental results suggest that the proposed method obtains overall best performance. In addition, we show that by applying a coarse feature value quantization it is possible to greatly reduce the memory storage of our image representation without sacrificing its classification performance.

[1] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *CoRR*, abs/1310.1531, 2013.

[2] Josephine Sullivan Atsuto Maki Stefan Carlsson Hossein Azizpour, Ali Sharif Razavian. From generic to specific deep representations for visual recognition. In *arXiv:1406.5774*, 2014.

[3] Lingqiao Liu, Chunhua Shen, Lei Wang, Anton van den Hengel, and Chao Wang. Encoding high dimensional local features by sparse coding based fisher vectors. In *NIPS*, 2014.

[4] Ross Girshick Trevor Darrell Ning Zhang, Jeff Donahue. Part-based r-cnns for fine-grained category detection. In *ECCV*, 2014.

[5] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. *CVPR Workshop*, 2014.

[6] Ning Zhang, Ryan Farrell, Forrest Iandola, and Trevor Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *ICCV*, December 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.