Higher-order Integration of Hierarchical Convolutional Activations for Fine-grained Visual Categorization

Zhang, Liqiang

May 25, 2018

Deep convolutional neural networks (CNNs) have emerged as the new state-of-the-art for a wide range of visual recognition tasks. Nevertheless, it remains quite challenging to derive the effective discriminative representation for fine-grained visual categorization (FGVC), primarily due to subtle semantic differences between sub-ordinate categories. Conventional CNNs usually deploy the fully connected layers to learn global semantic representation and may not be suitable to FGVC. Therefore, leveraging local discriminative patterns in CNN is crucial to obtain more powerful representation, and recently has been intensively studied for FGVC.

In recent works [1,2], the deeper convolutional filters are regarded as weak part detectors and the corresponding activations as the responses of detection, shown in Fig. 1. Motivated by this observation, instead of part annotations and explicit appearance modeling, the author straightforwardly exploit the higher-order statistics from the convolutional activations. They provide a perspective of matching kernel to understand the widely adopted mapping and pooling schemes on convolutional activations in conjunction with linear classifier. Linear mapping and direct pooling only capture the occurrence of parts. In order to capture the higher-order relations among parts, it is better to explore local non-linear matching kernels to characterize the higher-order part interactions.

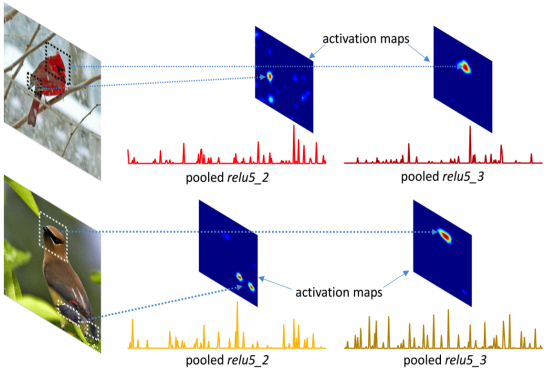

Figure 1: Visualization of several activation maps that corresponds to large responses of the sum-pooled vectors of two activation layers $relu5\_2$ and $relu5\_3$ in VGG-16 model.

Suppose that an image I is passed by a plain CNN, the author denote the 3D activations $\boldsymbol{\chi} \in R^{K\times M\times N}$ N extracted from some specific convolutional layer as a set of K-dimensional descriptors $\{x_p\}_{p\in\Omega}$, where K is the number of feature channels, x_p represents the descriptor at a particular position p over the set Ω of valid spatial locations ($\mid \Omega \mid = M \times N$). The author first consider the matching scheme $\boldsymbol{\kappa}$ for activation sets $\boldsymbol{\chi}$ and $\overline{\boldsymbol{\chi}}$ from two images, in which the set similarity is measured via aggregating all the pairwise similarities among the local descriptors:

$$\boldsymbol{\kappa}(\boldsymbol{\chi}, \overline{\boldsymbol{\chi}}) = Agg(\{k(\boldsymbol{x_p}, \boldsymbol{\bar{x}_{\bar{p}}})\}_{p\in\Omega,\bar{p}\in\Omega}) = \psi(\boldsymbol{\chi})^T\psi(\bar{\boldsymbol{\chi}}), \tag{1}$$

where $k(\cdot)$ is some kernel function between individual descriptors of two activation sets, $Agg(\cdot)$

is some set-based aggregation function, $\psi(\boldsymbol{\chi})$ and $\bar{\boldsymbol{\chi}}$ are the vector represetations for sets.

r	2	3	4	5	6
Training	9.7	7.4	5.5	4.2	2.8
Testing	29.8	23.7	18.3	14.5	10.4

Table 1: FPS with different non-homogeneous polynomial kernels.

Table. 1 lists the frame-per-second (FPS) comparison in both training and testing phases using different polynomial kernels. Since there is high computational overhead involved in the polynomial modules in the network, a large degree r will significantly slow the speed. Therefore, we suggest to adopt 2 as the practical degree in all the experiments in Section 5.3 even though degree-3 kernel can achieve slightly better results on Aircraft and Cars datasets.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, In *CVPR*, pp. 770–778, 2015. 1

[2] Y. Lecun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, Backpropagation applied to handwritten zip code recognition, In *Neural Computation*, vol. 1, no. 4, pp. 541–551, 2014. 1