Learning Coarse-to-Fine Sparselets for Efficient Object Detection and Scene Classification

Gong Cheng[1], Junwei Han[1,*], Lei Guo[1], Tianming Liu[2]
[1]School of Automation, Northwestern Polytechnical University, Xi'an, China; [2]Department of Computer Science, The University of Georgia

Part model-based methods have been successfully applied to object detection and scene classification and have achieved state-of-the-art results. Their success is largely owing to the usage of a number of part detectors to explicitly capture some discriminative visual concepts. However, as the number of part detectors grows, a major bottleneck to their wide applications is the heavy computational load caused by the exhaustive convolution operation between image feature pyramid and the large number of part detectors. This severely prevents them from scaling up to dealing with a large number of object or scene categories.

To tackle this problem, more recently the "sparselets" work [1-3] were introduced to serve as a shared intermediate representation for multi-class object detection, resulting in notable speedup. In this application, the sparselets are defined as a universal set of shared basis learned from a number of part filters. With this representation, the part filter responses can be reconstructed as sparse combinations of the sparselets with their corresponding activation vectors instead of exhaustive convolutions. However, although the existing sparselets work have obtained a great computational saving, some intrinsic drawbacks still exist in these methods.

To learn more effective sparselets, we propose a novel solution by constructing a coarse-to-fine training framework, as shown in Fig. 1. It consists of two stages: coarse sparselets training and fine sparselets and discriminative activation vectors training. In the first stage, coarse sparselets are trained to exploit the redundancy existing among different part detectors by using an unsupervised single-hidden-layer auto-encoder. The parameters between input layers and single-hidden-layers denote the to-be-learned coarse sparselets and the number of neurons in the hidden layer corresponds to the sparselets dictionary size. In the second stage, we simultaneously train fine sparselets and discriminative activation vectors in a unified framework, using a single-hidden-layer neural network with L0-norm sparsity constraint, to adequately explore the discriminative information hidden in the part detectors.

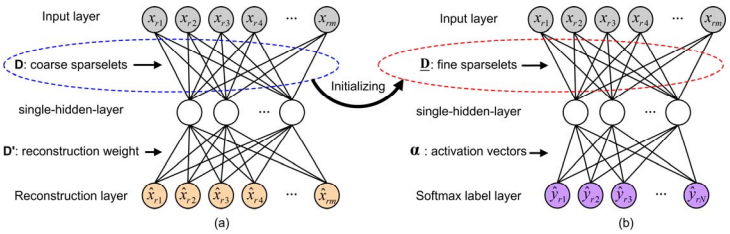

Figure 1. Framework of the proposed coarse-to-fine sparselets training: (a) coarse sparselets training; (b) fine sparselets and discriminative activation vectors training.

We comprehensively evaluated our method on three benchmarks: the PASCAL VOC 2007 dataset, the MIT Scene-67 dataset, and the UC Merced Land Use dataset, where the former one is used for 20-class object detection and latter two are used for 67-class indoor scene classification and 21-class aerial scene classification. The performance of object detection and scene classification is evaluated by mean Average Precision (AP) and average classification accuracy (ACA), respectively.

PASCAL VOC 2007 dataset: We adopted the widely used deformable part models (DPMs) [4] as our baseline and used the off-the-shelf part filters of DPMs from voc-release4 to train the sparselets. Table 1 presents the comparison results.

Table 1. Results on PASCAL VOC 2007 dataset

Methods	mean AP (%)	Speedup factor
DPM voc-release4 [4]	32.30	baseline
Reconstructive sparselets [1]	24.13-30.64	2.43-1.59
Discriminative sparselets [2]	29.21-31.89	
Our coarse-to-fine sparselets	30.32-32.25	

MIT Scene-67 dataset: We adopted the work of [5] as our baseline since [5] has demonstrated state-of-the-art performance on this challenging dataset by learning 200 part detectors for the most frequently occurring elements per class, for a total of 13,400 part detectors. In our work we directly used these off-the-shelf part detectors of [5] to train the sparselets. Table 2 presents the comparison results.

Table 2. Results on MIT Scene-67 dataset

Methods	ACA (%)	Speedup factor
ROI + Gist [6]	26.05	--
Object Bank [7]	37.60	--
D-patches [8]	38.10	--
Mid-level visual elements [5]	64.03	baseline
Reconstructive sparselets [1]	43.12-56.45	24.72-7.75
Discriminative sparselets [2]	53.84-61.21	
Our coarse-to-fine sparselets	59.87-64.36	

UC Merced Land Use dataset: We adopted the work of [16] as our baseline, in which there are total 3,093, 3,140, 3,072, 3,110, and 2,822 part detectors on all five held-out sets, respectively. Table 3 presents the comparison results.

Table 3. Results on UC Merced Land Use dataset

Methods	ACA (%)	Speedup factor
SPCK++ [9]	77.38	--
BRSP [10]	77.80	--
COPD [11]	91.33	baseline
Reconstructive sparselets [1]	64.52-83.24	14.78-4.44
Discriminative sparselets [2]	79.32-88.02	
Our coarse-to-fine sparselets	85.23-91.46	

[1] H. O. Song, S. Zickler, T. Althoff, R. Girshick, M. Fritz, C. Geyer, P. Felzenszwalb, and T. Darrell. Sparselet models for efficient multiclass object detection. In *ECCV*, 2012.
[2] R. Girshick, H. O. Song, and T. Darrell. Discriminatively activated sparselets. In *ICML*, 2013.
[3] H. Song, R. Girshick, S. Zickler, C. Geyer, P. Felzenszwalb, and T. Darrell. Generalized sparselet models for real-time multiclass object recognition. *TPAMI,* 2014.
[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI,* 32(9): 1627-1645, 2010.
[5] C. Doersch, A. Gupta, and A. A. Efros. Mid-level visual element discovery as discriminative mode seeking. In *NIPS*, 2013.
[6] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009.
[7] L. Li, H. Su, E. P. Xing, and F. Li. Object Bank: a high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, 2010.
[8] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012.
[9] Y. Yang and S. Newsam. Spatial pyramid co-occurrence for image classification. In *ICCV*, 2011.
[10] Y. Jiang, J. Yuan, and G. Yu. Randomized spatial partition for scene recognition. In *ECCV*, 2012.
[11] G. Cheng, J. Han, P. Zhou, and L. Guo. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS J. Photogramm. Remote Sens.,* 98: 119-132, 2014.