Weakly Supervised PatchNets: Learning Aggregated Patch Descriptors for Scene Recognition

Zhe Wang, Limin Wang, Yali Wang, Bowen Zhang, Yu Qiao, and Charless Fowlkes

Abstract

*In this paper, we propose a hybrid representation, which leverages the great discriminative capacity of CNNs and the efficiency of descriptor encoding scheme scene recognition. We make three main contributions. First, we train an end-to-end **PatchNet** in a weakly supervised manner, in order to extract the discriminative deep descriptors of local patches. Second, we design a novel **VSAD** encoding approach. With the help of semantic predictions from PatchNet, it can effectively aggregate deep local-patch descriptors into a global image representation. Finally, we evaluate our approach on two standard scene recognition benchmarks to show the effectiveness, i.e., MIT Indoor67 (86.2%) and SUN397 (73.0%).*

1. Introduction

Conventional scene recognition approaches have mainly relied on local-descriptor encoding framework, such as Vector of Locally Aggregated Descriptor (VLAD) [4], and Fisher vector (FV) [7]. Recently, Convolutional Neural Networks (CNNs) have made remarkable progress on image recognition since its success in ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 [6]. Several works try to combine the encoding methods and deeply-learned representations for image recognition [2] in order to take advantage of both approaches. For instance, Dixit *et al.* [2] designed a semantic Fisher vector to aggregate features from multiple layers (both convolutional and fully-connected layers) of CNNs for scene recognition. Arandjelovic *et al.* [1] developed a new generalized VLAD layer to train an end-to-end network for instance-level recognition. Our work follows this research direction. However, it differs from these works on two aspects: (1) we design a new PatchNet architecture to learn patch-level descriptors in a weakly supervised manner; (2) we develop a new aggregating scheme, VSAD, to encode deep local-patch descriptors with the semantic predictions from PatchNet. Consequently, it can alleviate the limitation of unsupervised dictionary learning, and makes the final representation more effective for scene recognition.

2. Method

We first extract descriptors and semantic probability from PatchNet. Second, we use them to construct our semantic codebook. Finally, we use the semantic codebook, PatchNet descriptors and semantic probability to construct our final VSAD representation as in Fig. 1.

PatchNet and Its Descriptors. Our PatchNet aims to learn the patch-level descriptor from raw RGB values, by classifying them into predefined semantic categories. We apply the image-level label to each randomly-selected patch, and utilize this transferred label as supervision signal to train the PatchNet. It hierarchically extracts multiple-level representations (hidden layers, denoted as $\mathbf{f}$) from patches, and eventually outputs the probability distribution over semantic categories (output layers, denoted as $\mathbf{p}$). The final output probability $\mathbf{p}$ yields an abstracted representation of a local patch, while the hidden layer activation features $\mathbf{f}$ are capable of providing more detailed and structural information.

Semantic probability. Aggregation-based encoding methods (e.g., Fisher vector) often rely on generative models (e.g., GMMs) to calculate the posterior distribution of a local patch, indicating the probability of belonging to a codeword. A full generative model often introduces latent variables $\mathbf{z}$ to capture the underlying factors and the complex distribution of local patches $\mathbf{x}$ can be obtained by marginalization over latent variables $\mathbf{z}$ as follows: $p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$. However, from the view of aggregation process, only the posterior probabilities $p(\mathbf{z}|\mathbf{x})$ are needed to determine the (soft) assignment of a local patch $\mathbf{x}$ to these learned codewords. We directly calculate $p(\mathbf{z}|\mathbf{x})$ with our proposed PatchNet instead of relying on generative methods. Directly modeling $p(\mathbf{z}|\mathbf{x})$ with PatchNets can be robustly trained on large-scale supervised datasets, avoiding the difficulties of fitting generative models which are often sensitive to initialization. Moreover, prediction scores of PatchNet correspond to semantic categories, which are more discriminative than those of the original generative model (e.g., GMMs).

Semantic Codebook. Given a set of local patches $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, we first compute their semantic probabilities with PatchNet, denoted as $\mathcal{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N\}$

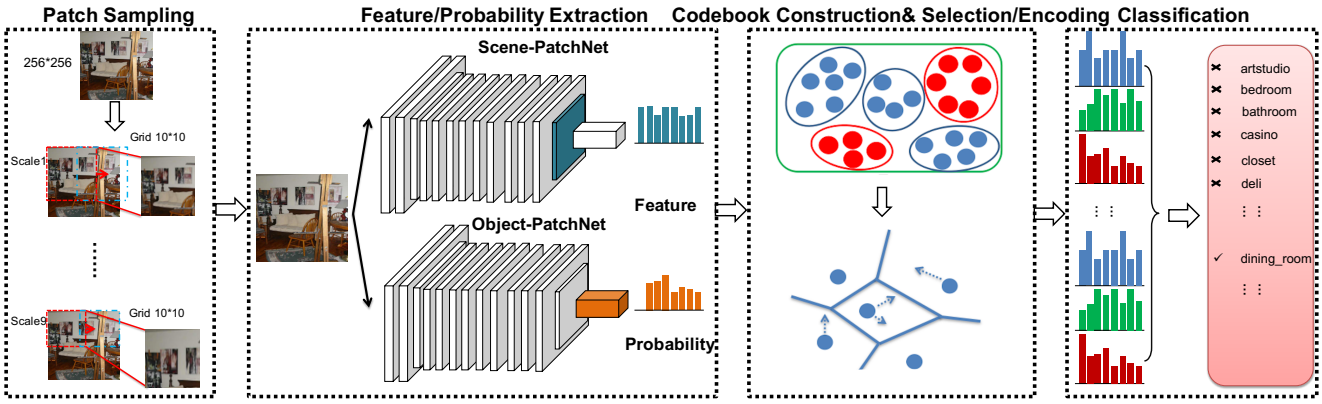

Figure 1. **Pipeline of our method**. We first densely sample local patches. Then, we utilize two kinds of PatchNets to describe patch contents (Scene-PatchNet feature) and encode patch descriptors (Object-PatchNet probabilities), respectively. Based on our learned semantic codebook, these local patches are aggregated into a global representation with VSAD encoding scheme. Finally, these global representations are utilized for scene recognition with a linear SVM.

and extract patch-level descriptors $\mathcal{F} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N\}$. Then, we generate **semantic mean** (center) for each codeword as follows: $\mu_k = \frac{1}{N_k}\sum_{i=1}^{N} \mathbf{p}_i^k \mathbf{f}_i$, where $\mathbf{p}_i^k$ is the k^{th} dimension of $\mathbf{p}_i$, and N_k is calculated as follows: $N_k = \sum_{i=1}^{N} \mathbf{p}_i^k$, $\pi_k = \frac{N_k}{N}$. We can interpret N_k as the prior distribution over the semantic categories and μ_k as the category template in this feature space $\mathbf{f}$. Similarly, we can calculate the **semantic covariance** for each codeword by the weighted sample estimate: $\Sigma_k = \frac{1}{N_k}\sum_{i=1}^{N} \mathbf{p}_i^k (\mathbf{f}_i - \mu_k)(\mathbf{f}_i - \mu_k)^\top$. The semantic mean and covariance constitute our semantic codebook.

VSAD. The procedure is illustrated in Fig 1. After the description of PatchNet and semantic codebook, we are able to construct a hybrid visual representation, namely *vector of semantically aggregating descriptor* (VSAD). Given a set of local patches with descriptors $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_T\}$, we aggregate both first order and second order information of local patches with respect to semantic codebook as follows: $\mathcal{S}_k = \frac{1}{\sqrt{\pi_k}}\sum_{t=1}^{T} \mathbf{p}_t^k \left(\frac{\mathbf{f}_t - \mu_k}{\sigma_k}\right)$, $\mathcal{G}_k = \frac{1}{\sqrt{\pi_k}}\sum_{t=1}^{T} \mathbf{p}_t^k \left[\frac{(\mathbf{f}_t - \mu_k)^2}{\sigma_k^2} - 1\right]$, where $\{\pi, \mu, \sigma\}$ is semantic codebook defined above, $\mathbf{p}$ is the semantic probability calculated from PatchNet, $\mathcal{S}$ and $\mathcal{G}$ are first and second order VSAD, respectively. Finally, we concatenate these subvectors from different codewords to form our VSAD representation: $[\mathcal{S}_1, \mathcal{G}_1, \mathcal{S}_2, \mathcal{G}_2, \cdots, \mathcal{S}_K, \mathcal{G}_K]$.

3. Experiments

We examine the effectiveness of our VSAD representation on MIT indoor [5] and SUN397 [9]. It achieves the-state-of-the-art performance on two datasets, which verifies the power of our VSAD representation.

Method	MIT indoor	SUN397
Semantic Fisher vector [2]	79.0	61.0
Data Bias [3]	81.0	66.3
Human Performance [9]	-	68.5
Ours[8]	**86.2**	**73.0**

Table 1. Our method secures the state-of-the-art performance. Details are reported in our paper [8].

References

[1] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *CVPR*, 2016. 1

[2] M. Dixit, S. Chen, D. Gao, N. Rasiwasia, and N. Vasconcelos. Scene classification with semantic fisher vectors. In *CVPR*, 2015. 1, 2

[3] L. Herranz, S. Jiang, and X. Li. Scene recognition with cnns: objects, scales and dataset bias. In *CVPR*, 2016. 2

[4] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *PAMI*, 2012. 1

[5] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009. 2

[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 1

[7] J. Sánchez, F. Perronnin, T. Mensink, and J. J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 2013. 1

[8] Z. Wang, L. Wang, Y. Wang, B. Zhang, and Y. Qiao. Weakly supervised patchnets: Describing and aggregating local patches for scene recognition. *TIP*, 2017. 2

[9] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 2