Learning Image Representations for Efficient Recognition of Novel Classes

Alessandro Bergamo   Lorenzo Torresani
Computer Science Department, Dartmouth College
6211 Sudikoff Lab, Hanover, NH 03755, U.S.A.
{aleb, lorenzo}@cs.dartmouth.edu

Introduction

In this work we consider the problem of efficient object-class recognition in large image collections. We are specifically interested in scenarios where the classes to be recognized are not known in advance. The motivating application is "object-class search by example" where a user provides at query time a small set of training images defining an arbitrary *novel* category and the system must retrieve from a large database images belonging to this class. This application scenario poses challenging requirements on the system design: the object classifier must be learned efficiently at query time from few examples; recognition must have low computational cost with respect to the database size; finally, compact image descriptors must be used to allow storage of large collections in memory rather than on disk for additional efficiency.

Traditional object categorization methods do not meet these requirements as they employ high-dimensional descriptors and they typically use non-linear kernels which render them computationally expensive to train and test. For example, the LP-β multiple kernel combiner described in [2] achieves state-of-the-art accuracy on several categorization benchmarks but it requires over 23 Kbytes to represent each image and it uses 39 feature-specific nonlinear kernels. Note that the use of this recognition model in our application would require costly *query-time* kernel evaluations for each image in the database since the training set varies with every new query and thus pre-calculation of kernel distances is not possible.

We propose to address these storage and efficiency requirements by learning a compact binary image representation optimized to yield good categorization accuracy with linear (i.e., efficient) classifiers. The binary entries in our image descriptor are thresholded nonlinear projections of low-level visual features extracted from the image, such as descriptors encoding texture or the appearance of local image patches. Each non-linear projection can be viewed as implementing a nonlinear classifier using multiple kernels. The intuition is that we can then use these pre-learned multiple kernel combiners as a **classification basis** to define recognition models for arbitrary novel categories: the final classifier for a novel class is obtained by linearly combining the binary outputs of the basis classifiers, which we can pre-compute for every image in the database, thus enabling efficient novel object-class recognition even in large datasets.

The idea of describing images in terms of basis classes is evocative of the use of attributes [1, 3] which are fully-supervised classifiers trained to recognize certain properties in the image such as "has beak", "near water". The recognition model for each class is then defined by hand-specifying its associated attributes.

The approach proposed here is most closely related to our prior work [5], where we have introduced the "classeme" image descriptor for general class recognition: the entries of this descriptor are the outputs of a large set of weakly-trained basis classifiers evaluated on the image. We have shown that linear classifiers trained on classeme vectors can recognize novel classes with near state-of-the-art accuracy even when the descriptor is compressed down to less than 200 bytes per image. In subsequent work [4], Li et al. have proposed using the localized outputs of object detectors as image representation. The advantage of this representation is that it encodes spatial information; furthermore, object detector are more robust to clutter and uninformative background than classifiers evaluated on the entire image. These prior methods work under the assumption that an "overcomplete" representation for classification can be obtained by pre-learning classifiers for a large number of basis classes, some of which will be related to those encountered at test-time. Such high-dimensional representations are then compressed down using quantization, dimensionality reduction or feature selection methods in a stage subsequent the learning of the basis classifiers [5, 4].

Unlike prior work, where the basis classifiers are learned disjointly and each is optimized to recognize a predefined basis class, we propose instead to jointly seek the basis classifiers such that linear combinations of these classifiers yield optimal classification accuracy on a given training set. In other words, our basis classifiers are recognizers of "abstract" (as opposed to predefined) object categories and they are collectively trained to produce good classification when used as a basis in a linear combination. This allows us to decouple the number of basis classifiers from the number of categories in the training set and thus to learn

Topics: vision; learning algorithms. Preference: oral/poster.

the basis classifiers for any specified target dimensionality, without resorting to a subsequent suboptimal compression of the descriptor. Finally, we learn all parameters with respect to the binarized outputs of the basis classifiers, therefore solving for a compact *binary* encoding of the images. Our work shares similarities with the approach of Weiss et al. [7] where binary image codes are learned such that the Hamming distance between codewords approximates the Euclidean distance between GIST descriptors. In contrast, we seek binary codes such that a linear combination of the bits associated to an image can produce accurate classification. Furthermore, our bits implement powerful non-linear projections of multiple low-level features.

Method

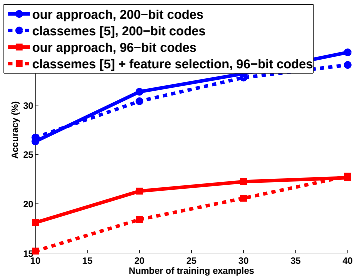
Let $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$ be the training set for learning the basis classifiers, where $\boldsymbol{x}_i$ is the i-th image example and $\boldsymbol{y}_i \in \{-1, +1\}^K$ is a vector encoding the category label out of K possible classes: $y_{ik} = +1$ iff the i-th example belongs to class k. We extract 13 distinct low-level feature descriptors from each image, corresponding to different spatial pyramid levels of GIST, HOG, SIFT and self-similarity features. Using the closed-form explicit feature maps of Vedaldi and Zisserman [6], we map each descriptor to a higher-dimensional space such that inner products in this space approximate the intersection kernel distances. Let $\hat{\boldsymbol{f}}_i$ be the vector obtained by concatenating the descriptors of image i after the explicit feature maps. We then define our c-th basis classifier to be a boolean function of the form: $h(\hat{\boldsymbol{f}}_i; \boldsymbol{a}_c) = \mathbf{1}[\boldsymbol{a}_c^T \hat{\boldsymbol{f}}_i > 0]$ where $\mathbf{1}[.]$ denotes the indicator function. Note that $h(\hat{\boldsymbol{f}}_i) \in \{0, 1\}$ computes a thresholded *nonlinear* projection of the original low-level features. We train the basis classifiers by optimizing the following learning objective, which is a trade off between a small classification error (when using the output bits of the basis classifiers as features in a one-versus-all linear SVM) and a large margin:

$$E(\boldsymbol{a}, \boldsymbol{w}, \boldsymbol{b}) = \sum_{k=1}^{K} \left\{ \frac{1}{2} \|\boldsymbol{w}_k\|^2 + \frac{\lambda}{N} \sum_{i=1}^{N} l(y_{i,k}(b_k + \sum_{c=1}^{C} w_{kc} \mathbf{1}[\boldsymbol{a}_c^T \hat{\boldsymbol{f}}_i > 0])) \right\} \tag{1}$$

where C is the total number of basis classifiers to learn and $l()$ is the traditional hinge loss function. Note that the linear SVM and the basis classifiers are learned jointly. We minimize this error function by block coordinate descent. Upon convergence, given an image with descriptor $\hat{\boldsymbol{f}}$, we use the estimated parameters $\boldsymbol{a}$ to calculate its binary basis classifier vector as $[h(\hat{\boldsymbol{f}}; \boldsymbol{a}_1) \dots h(\hat{\boldsymbol{f}}; \boldsymbol{a}_c)]$.

Experiments

We learned basis classifiers with the approach described here and compared them with classemes trained with the method of [5]. For both algorithms we used a training set of 200 ImageNet classes with 120 images drawn from each class. We then extracted binary codes from Caltech256 images for both learned representations. Finally, we trained linear SVMs on these binary descriptors for Caltech256 multi-class classification, restricting our test to 50 classes. Note that the ImageNet categories used to learn the basis classifiers are distinct from the Caltech256 classes, which represent effectively novel classes to recognize. The figure to the right reports the multiclass classification accuracy obtained with these two 200-bit descriptors. We also include results produced when setting the number of basis classes to 96 (corresponding to binary codes of only 12 bytes): while our approach can accommodate easily the case were the number of basis classes is different from the number of training categories, the classeme learning method of [5] requires a subsequent feature selection step which, as seen in the figure, greatly degrades its classification accuracy. Although the number of basis classes used in this toy experiment is too small to produce state-of-the-art accuracy, we believe that our significant relative improvement over [5] will carry over when larger training sets will be used.



References

[1] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[2] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *ICCV*, 2009.

[3] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[4] L. Li, H. Su, E. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*. 2010.

[5] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *ECCV*, 2010.

[6] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *CVPR*, 2010.

[7] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *NIPS*. 2009.