Combining Generative and Discriminative Models for Classifying Social Images from 101 Object Categories

Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, Andrea M. Serain,
Giuseppe Serra, and Benito F. Zaccone
Media Integration and Communication Center, University of Florence, Italy

Abstract

In this paper we present a hybrid generative-discriminative approach for image categorization in real-world images, based on Latent Dirichlet Allocation and SVM classifiers. We use SVMs with non-linear kernels on different visual features in a multiple kernel combination framework. A major contribution of our work is also the introduction of a novel dataset, called MICC-Flickr101, based on the popular Caltech101 and collected from Flickr. We demonstrate the effectiveness and efficiency of our method testing it on both datasets, and we evaluate the impact of combining image features and tags for object recognition.

1 Introduction

Automatic image annotation is an important task, in which the goal is to determine the relevance of annotation terms for images. Several efforts have been made in recent years to design and develop effective and efficient algorithms for visual recognition and retrieval [4]. To this end, a common and successful approach is to quantize local visual features (e.g. SIFT) following the well-known bag-of-features paradigm [10, 3]. Then, a binary classifier (e.g. SVM) can be learned from a collection of images manually labeled as belonging to an object category or not. Increasing the quantity and diversity of labeled images improves the performance of the classifier but, unfortunately, hand-labeling images is a time consuming task.

Since nowadays photo sharing websites that let users upload and tag their images, such as Flickr and Picasa, have became very popular, a recent trend in the field is to use these huge image corpus as sources to train visual classifiers [6, 8]. Although these sites offer us great opportunity to "freely" get a large number of images with user annotations, it is recognized that many tags are noisy or overly personalized. Thus, if we are able to design and learn accurate models from these images with their associated noisy tags, content-based image retrieval and annotation should benefit much from this community contributed media collections.

In this paper we present a hybrid generative-discriminative approach for image categorization based on Latent Dirichlet Allocation [1] and non-linear SVM classifiers. We follow state-of-the-art image categorization methods, and use SVM with non-linear kernels on several bag of local visual features in a framework based on multiple kernel combination. A major contribution of our work is the introduction of a new dataset, called MICC-Flickr101[1], based on the popular Caltech101 dataset and obtained from Flickr. In our view, it can be used to compare and evaluate object categorization performance in a constrained scenario (Caltech101) and object categorization "in the wild" (MICC-Flickr101) on the same 101 categories. Moreover, since we provide also the original user tags and metadata, this dataset can be used to evaluate the impact of using social annotations combined to visual features for object recognition.

The rest of the paper is organized as follows. In Section 2, we present our dataset and we compare it with Caltech101. Section 3 describes our hybrid generative-discriminative approach for image categorization. In Section 4, we present experiments and baselines on our novel dataset. Our goal is to show that MICC-Flickr101 can serve as a useful benchmark for social image classification and object categorization in real-world images.

2 The MICC-Flickr101 Dataset

We collected our MICC-Flickr101 dataset using the 101 object categories of the popular Caltech101 dataset [5], since it is the first large-scale image dataset that served as the pivotal point for object categorization and it is probably the most used benchmark in the field. Caltech101 was obtained using the Google Image Search engine (in September 2003), it has about 40 to 800 images per category and most categories have about 50 images. The size of each image is roughly 300×200 pixels. Moreover, several images were manually flipped

[1] www.micc.unifi.it/datasets/micc-flickr-101

or rotated, so that all instances face the same direction. These are probably the main drawbacks of the dataset: most of the Caltech101 objects are of uniform size and orientation within their class, have the same spatial layout, and lack rich backgrounds. The common experimental protocol is to select 15–30 images as train set and the rest as test set. Unfortunately, some classes have around 30 images in total (e.g. "inline_skate" has 31 images, "binocular" 33) and so the dataset is quite unbalanced, especially if 30 images are selected for training.

The MICC-Flickr101 dataset was obtained by downloading images from Flickr in January 2012. As query we selected, for each object category, the name of the class in English and, depending on the class, its translation up to three other languages (i.e. Spanish, Italian, French). We have further manually inspected each category, getting rid of irrelevant images, and collected as many images as possible for each category in order to have a more balanced dataset. In total our dataset is composed of 7348 images (*vs.* 9144 of Caltech101) with at least about 40 images per class; the median of the number of elements per class is 70 (*vs.* 59 of Caltech101). Images are at high resolution, 1024×768 pixels on average, and depict objects in daily-life real scenarios. For each image we provide also a file containing user tags, the title of the image and (when available) other data such as geo-coordinates or EXIF data.

Why is it useful? Recently several large-scale image datasets have emerged. The MIR-Flickr retrieval evaluation initiative and the NUS-WIDE dataset are both obtained by crawling Flickr. They provide images, user-tags, and manual annotations for some visual concepts. Also the PASCAL VOC challenge has a large set of images from Flickr, while ImageNet is a very-large-scale web mined dataset (thousands of categories) organized following the WordNet taxonomy. Our MICC-Flickr101 dataset fixes the main drawback of Caltech101, i.e. its low intra-class variability, and provides social annotations through user tags. It builds on a standard and widely used dataset composed of a still manageable number of categories (101) and, therefore, can be used to compare and evaluate object categorization performance in a constrained scenario (Caltech101) and object categorization "in the wild" (MICC-Flickr101) on the same 101 categories. Moreover, user tags can be used to evaluate the impact of using social annotations combined to visual features for object recognition. Figure 2 shows examples of a few classes from Caltech101 and MICC-Flickr101.

3 Our Approach

We propose a hybrid generative-discriminative approach based on Latent Dirichlet Allocation and non-linear SVM classifiers. Figure 1 illustrates our full framework. We follow the work of Bosch *et al.* [2], that is based on pLSA and kNN/SVM classifiers, and we extend it by discovering latent topics using LDA and also by putting our hybrid model in a simple multiple kernel framework. The Multiple Kernel Fusion (MKF) [7] enables us to effectively combine non-linear kernels based on different image features.

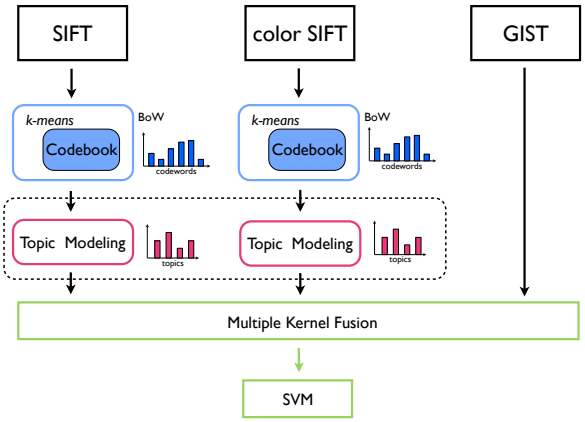

Figure 1. Our framework.

3.1 Multiple Visual Features

In order to take into account the diversity of the image content, we have used three types of visual features (in the following denoted as f_m): SIFT, color-SIFT (in particular we used *rgb*SIFT) [11], and GIST [9]. The GIST is a 980-d global feature representing dominant spatial structure of a scene by a set of perceptual dimensions such as naturalness, openness, and roughness. SIFT and color-SIFT are local features, i.e. they depict local information of the visual content, respectively in grayscale or in a specific color space. The SIFT is a 128-d descriptor while the *rgb*SIFT is a 384-d descriptor. In both cases, we adopt a dense sampling strategy for keypoint detection; features are extracted at 4 scales (0.5, 1, 1.5, and 2) with a regular grid spaced 10 pixels.

3.2 Bag-of-Features

According to the bag-of-features model, images are defined as sets of codewords obtained from the clustering of local visual descriptors. Following this approach, we first have constructed a codebook for each local descriptor (SIFT and *rgb*SIFT) separately, using the standard k-means algorithm. To limit the complexity, we clustered a subset of $150,000$ randomly selected training features and, to increase precision, we initialized k-means 4 times and kept the result with the lowest error.

3.3 Topic Modeling

Latent Dirichlet Allocation (LDA) [1] is a generative model used in the statistical text literature to dis-

cover *hidden topics* in a document represented using the bag-of-words representation. It has been used with success also in computer vision for object/scene classification in images. In our framework we have used LDA to obtain an "intermediate" compact representation of the image. In the training phase, we learned topic specific distributions $P(w|z)$ starting from bag-of-features histograms. Each image is represented by a Z-vector $P(z|d_{train})$ where Z is the number of topics learned. Further, these topic distributions can be used in the discriminative stage in place of bag-of-features histograms as input of the SVM classifiers.

3.4 Multiple Kernel Fusion

Kernel methods make use of kernel functions defining a measure of similarity between pairs of elements. Given a visual feature f_m (e.g. SIFT), a kernel function k between real vectors (i.e. the bag-of-features histograms) is defined as $k_m(x, x') = k(f_m(x), f_m(x'))$, such that the image kernel $k_m : X \times X \to \mathbb{R}$ considers similarity with respect to the image features f_m. Since we associate image features with kernel functions, a simple but effective approach for multiple feature "late fusion", is to combine kernels and sub-sequently use the resulting kernel for SVM training. To this end we have experimented *averaging kernels* and *product kernel*.

Averaging Kernels. Given F different image features, we define the kernel function as the average of kernels, $k^*(x, x') = \frac{1}{F}\sum_{m=1}^{F} k_m(x, x')$, which is subsequently used in a standard SVM.

Product Kernels. In this case we obtain the combined kernel by multiplication $k^*(x, x') = (\prod_{m=1}^{F} k_m(x, x'))^{1/F}$, that, as previously, it is subsequently used as the single kernel in a SVM.

4 Experiments

4.1 Results on Caltech101 dataset

We follow the experimental setup proposed by the designers of the dataset. The performance is measured as the mean prediction rate per class, further balancing the influence of categories with a large number of test examples. We report results using 101 class (i.e. we do not use the background class), 30 images per category for training and up to 50 images for testing. The classification accuracy results (in percentage) are provided in Table 1. First we show results obtained with single and multiple kernel modalities on bag-of-features histograms, using 1000 codewords for SIFT and 1500 for *rgb*SIFT. Then we report results obtained using topic distributions as input of the SVM classifiers; in this case LDA is used to obtain an efficient representation of 200 topics, starting from codebooks with 1500 codewords. The results related to multiple kernel fusion are obtained by averaging (avg) and product (prod) of the three modalities. All kernels are computed as $\exp(\gamma^{-1} d_\chi(x, x'))$ where d_χ is the χ^2 distance between image features and γ is fixed to the mean of the pairwise distances.

It is interesting to note that the MKF on LDA gave us very similar performance with respect to the standard bag-of-features, although we used 200-d histograms instead of the original 1500-d ones; the speed-up of the process is 75% w.r.t. BoF baseline. Moreover, the reported results on Caltech101 are intended as simple baselines for our novel dataset and were obtained without spatial-pyramids that are known to significantly improve the performance.

	GIST	SIFT$_{bof}$	*rgb*SIFT$_{bof}$	avg	prod
Accuracy	48.6 ± 1.2	46.2 ± 0.5	47.6 ± 0.1	58.4 ± 1.0	57.9 ± 0.8
	GIST	SIFT$_{LDA}$	*rgb*SIFT$_{LDA}$	avg	prod
Accuracy	48.6 ± 1.2	44.3 ± 1.9	45.2 ± 0.3	58.0 ± 1.3	57.7 ± 1.3

Table 1. Results on Caltech101.

4.2 Results on MICC-Flickr101 dataset

We follow the same experimental setup used for Caltech101. Here we used vocabularies of 3000 words (for both SIFT and *rgb*SIFT) and 500 topics for LDA. As in the previous case, all kernel matrices were computed by using an exponential kernel with χ^2 distances.

Tags as feature. We also used tags as a feature to describe the image content. For the training images we excluded the class name from the representation to avoid learning a classifier that uses the class name to perfectly predict itself. Textual features are used in a standard bag-of-words model. We constructed the dictionary starting from all the tags associated to the training images and, after stemming and stop-words removal, we obtained a dictionary of 1000 terms. Further, we used LDA also on the tags and we learned 100 topics.

	GIST	SIFT$_{bof}$	*rgb*SIFT$_{bof}$	TAGS$_{bow}$
Accuracy$_{single}$	26.1 ± 1.0	31.2 ± 0.9	34.4 ± 0.6	50.9 ± 1.6
	avg$_{vis}$	prod$_{vis}$	avg$_{vis+tag}$	prod$_{vis+tag}$
Accuracy$_{MKF}$	39.3 ± 1.3	38.5 ± 1.4	58.9 ± 1.3	57.1 ± 0.3

Table 2. Results on MICC-Flickr101: single features and multiple kernel fusion on visual features only (vis) and on visual features + tags ($vis + tag$).

Results and discussion. Table 2 shows the results obtained using single visual features and tags feature, in a standard bag-of-words framework; the results on multiple kernel fusion are reported for a combination based on visual features only (the first three features) and also for visual features and tags (all the four features). Similarly, Table 3 reports the results obtained using LDA for SIFT, *rgb*SIFT, and TAGS.

Again, LDA guarantees high performance with a significant reduction of feature dimensionality (we loose around 3% in the visual MKF), leading to a speed-up of 78% w.r.t. BoF baseline. This result is in line with the previous one on Caltech101, and it is even more interesting considering that we can have several LDA models (both visual and textual) represented with the same coherent representation, despite they were originated from different data modalities.

	GIST	SIFT_{LDA}	$\textit{rgb}\text{SIFT}_{LDA}$	TAGS_{LDA}
Accuracy_{single}	26.1 ± 1.0	29.3 ± 0.9	32.5 ± 0.6	36.4 ± 1.3
	avg_{vis}	prod_{vis}	$\text{avg}_{vis+tag}$	$\text{prod}_{vis+tag}$
Accuracy_{MKF}	39.0 ± 1.0	38.6 ± 1.0	56.1 ± 0.7	55.0 ± 0.8

Table 3. Results on MICC-Flickr101 using LDA.

MICC-Flickr101 vs Caltech101. Our dataset is significantly more complex and challenging than Caltech101, as demonstrated by the drop in performance. A global feature such as GIST suffers a lot in a realistic and general scenario (-22.5%), while bag-of-features shows a drop of around 15%. The use of tags seems very promising, since in this case the MKF gives a result close to that obtained on the easier Caltech101 (e.g. 56.1 vs 58.0, using LDA and averaging kernels).

5 Conclusions

We presented a generative/discriminative method for image categorization, based on LDA and SVM classifiers, in a multiple kernel combination framework, that greatly reduces the computational cost. We also introduced the novel MICC-Flickr101 dataset based on the popular Caltech101. We demonstrated the effectiveness of our approach testing it on both datasets, and we evaluated the impact of combining image features and user tags for object recognition in real-world images. We hope that our preliminary results, and the public availability of our dataset, will encourage other researchers to test their algorithms for image annotation on weakly-labeled images obtained from social media.

Acknowledgments. This work was supported in part by the EU "euTV" Project (Contract FP7-226248).

References

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[2] A. Bosch, A. Zisserman, and X. Muñoz. Scene classification using a hybrid generative/discriminative approach. *IEEE TPAMI*, 30(4):712–727, 2008.

[3] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Proc. of ECCV-SLCV*, 2004.

[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM CSUR*, 40(2):1–60, 2008.

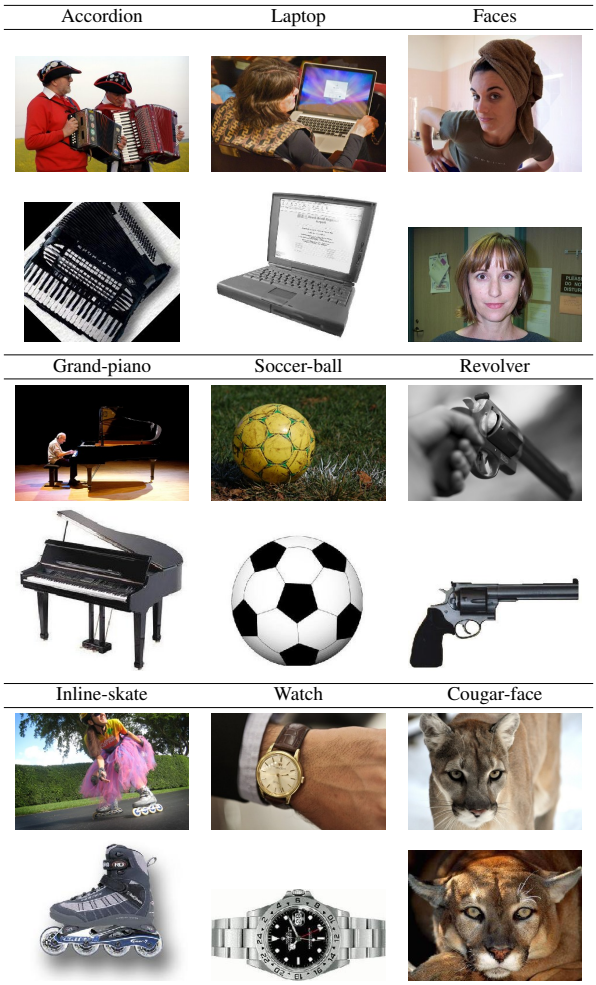

Figure 2. Sample images from MICC-Flickr101 (odd rows) and Caltech101 (even rows) datasets.

[5] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In *Proc. of CVPR-GMBV*, 2004.

[6] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from Google's image search. In *Proc. of ICCV*, 2005.

[7] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *Proc. of CVPR*, 2009.

[8] X. Li and C. G. M. Snoek. Visual categorization with negative examples for free. In *Proc. of ACM MM*, 2009.

[9] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[10] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. of ICCV*, 2003.

[11] K. E. A. van De Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE TPAMI*, 32(9):1582–1596, 2010.