SZTAKI @ ImageCLEF 2012 Photo Annotation*

Bálint Daróczy Dávid Siklósi András A. Benczúr
{daroczyb, sdavid, benczur}@ilab.sztaki.hu

Data Mining and Web Search Group, Informatics Laboratory
Computer and Automation Research Institute Hungarian Academy of Sciences, MTA SZTAKI

Abstract. In this paper we describe our approach to the ImageCLEF2012 Photo Annotation task. We used both visual and textual modalities for all submissions. We described each image with a fixed length representation using different similarity measures. By this method we were able to combine, before the classification, a large variety of descriptors to improve the classification quality. This descriptor is a combination of several visual and textual similarity values between the actual image and a reference image set, containing well selected training images. We trained Gaussian Mixture Models (GMM) to define a generative model for low-level descriptors extracted from the training set using Harris-Laplacian point detection. We used two descriptors, a grayscale gradient and a color moment based one. In order to measure the visual similarity between two images, we extracted several dense Fisher vectors per image. Besides calculating visual features, we adopted a biclustering method to cluster the Flickr tags and the images at the same time. Additionally, we measured the similarity of images according to their Flickr tags using Jensen-Shannon divergence.

Keywords: image classification, biclustering, generative models, kernel methods

1 Introduction

In this paper we describe our approach to the ImageCLEF 2012 Photo Annotation task [19]. The main challenge is to select proper image processing and feature extraction methods for our given classification and pre-processing framework. Our image descriptors included spatial pooling based Fisher vectors [13, 15] calculated on point descriptors [6, 11, 9] such as Histogram of Oriented Gradients and Color moments [6, 18]. We adopted several different methods to measure the similarity of images based on their Flick tags. Beside Jensen-Shannon divergence, we used a modified version of Dhillan's biclustering algorithm [8] to explore deeper connections between the images and the Flickr tags.

In comparison to other teams our best run achieved the second highest MiAP, GMiAP and F-measure scores among the 18 participants.

2 Visual feature extraction

The GMM based Fisher gradient vector computed of SIFT [10] descriptors is a well-known technique to represent an image with only one vector per pooling [15, 13, 18].

We used low-level patch feature vectors to describe the visual content of an image by approximately $15k$ descriptors per image per modality. Our sampling strategy included a dense grid and a Harris-Laplace point detection [11]. To avoid extracting large number of local features, we downscaled all the images with proper aspect ratio. The maximal width and height was set to 500 pixels. We calculated HOG (Histogram of Oriented Gradients [6]) and RGB color descriptors for each patch using 16x16 and 48x48 pixel macroblock sizes. Both descriptors were L2 normalized and by HOG we reduced the dimension to 96 by Principal Component Analysis (PCA). The PCA model was trained on a small sample of patches extracted from training images.

* This work is supported in part by the EC FET Open project "New tools and algorithms for directed network analysis" (NADINE No 288956) and OTKA CNK 77782.

To build an efficient soft codebook, we trained a Gaussian Mixture Model (GMM) for both descriptors. The training procedure of GMM models took about 20 minutes using 3 million training points per descriptor. We used our open-source CUDA GMM implementation. Our training method was based on a standard Expectation Maximization algorithm with a non-hiherarchical structure. We avoided the well-known vulnerability of the EM algorithm to underflow by computing the condiditonal probabilities [2]. The resulted implementation is an accurate yet fast CUDA based code optimized for fp32 architectures. Our source code along with previously trained GMM models for different patch descriptors and codes for Fisher vector calculation is available free for research use at https://dms.sztaki.hu/hu/projekt/gaussian-mixture-modeling-gmm-es-fisher-vector-toolkit.

The final high-level dataset independent representation of images was the normalized Fisher gradient vector. We also calculated a separate Fisher vector on the Harris-Laplacian detected corner descriptors. As by our GMM implementation we were able to compute all the conditional probabilites for each feature vector without significant loss of time, which resulted a strongly dense Fisher vector even in fp32.

3 Biclustering Flickr tags and image similarity

Our previous experiment [7], Jensen-Shannon divergence of Flickr tags was an excellent image similarity measure. Our goal was to expand it with determining deeper interrelations between the tags and the documents using content based similarity.

The applied biclustering was an expansion of Dhillan's information theoretic co-clustering algorithm [8]. In comparison to the original algorithm we measured document similarity with a combination of visual and textual similarity values. We chose Jensen-Shannon divergence instead of Kullback-Leibler used in the original article. Our choice was inspired by our experiences with other datasets where Jensen-Shannon divergence resulted a significantly better clustering quality instead of Kullback-Leibler [17]. In order to refine the clustering with non-textual information, we added a similarity measure based on the best performing visual features (both HOG and color Fisher vectors pooled on different partitions such as the whole image, only the detected corner points and 3x1 spatial resolution).

4 Combination of representations

Efficient combination of different feature sets based on a wide range of visual modalities is one of the main problems of image classification. This problem becomes more complex if we have additional non-visual features such as Flickr tags. Our starting point was a widely used technique: learning SVM models on textual and visual Bag-of-Words models [20, 5, 12]. The selection of the ideal kernel depends on both of the original feature space and the class variable. Therefore the selection procedure is computationally expensive. The dual form of the optimization problem of the standard Support Vector Machine (SVM) classification [4] with kernel $K(x_i, x_j)$ is the following:

$$MaximizeL_{Dual}(\alpha) = \sum_{i=1}^{m} -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{1}$$

subject to $\sum y_i \alpha_i = 0$ for all i with $\alpha_i \geq 0$.

Having multiple number of kernels due the representations via different modalities with previously selected kernel functions, we can modify the dual form into a multiple kernel learning problem:

$$MaximizeL_{Dual}(\alpha, \beta) = \sum_{i=1}^{m} -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \sum_{n=1}^{N} \beta_n K_n(x_i, x_j) \tag{2}$$

Table 1. Reference set selection

	Train. set	Ref. set	Perc.	MAP	Loss
jch5k	5000	5000	100.0%	0.3485	
p=0.10	5000	4280	85.6%	0.3485	0.0000
p=0.05	5000	3630	72.6%	0.3473	0.0012
p=0.03	5000	2414	48.3%	0.3448	0.0037
p=0.01	5000	595	11.9%	0.3082	0.0403

subject to $\sum y_i \alpha_i = 0$ for all i with $\alpha_i \geq 0$, where N is the number of the basic kernels and $K_n(xi, xj)$ is the nth kernel function.

The above problem is a special case of the Multiple Kernel Learning problems where the kernels are computed on different feature sets. In comparison to Bach [14] we assumed that all of representations are conducive to the train procedure. Bach suggested to solve the MKL problem with an iterative and sparse learning method where in each iteration they solve a standard SVM dual problem and update the weights of the basic kernels. One of the drawbacks of this solution is the increased computational time.

To avoid the computationally expensive MKL problem we used a feature transformation method. Distance from the training set, as a feature transform for classification is a well-known technique. Schölkopf et al. showed that a class of kernels can be represented as norm-based distances in Hilbert spaces [16] and Ah-Pine et al. applied L1-norm based feature transformation measuring the distance from the Fisher vectors of the training set for image classification with excellent results [1].

We defined a dense representation combining modality adaptive similarity based feature transforms. Let us consider a set of documents D (we call it as reference set) and their corresponding representations D_r. We defined the uniform representation of a document X over the set of representations R of a reference set D as

$$L_R(X, D) = [\sum_{r=1}^{R} \beta_r sim_r(X_r, D_{r1}), .., \sum_{r=1}^{R} \beta_r sim_r(X_r, D_{rd})] \tag{3}$$

where $\sum \beta_r = 1$ and sim_r denotes the selected similarity measure on basic representation r and d is the size of the reference set. The dimensionality of this representation is the cardinality of the reference set.

4.1 Reference set selection and weight determination

The proper selection of the reference set could decrease significantly the demanding computational time of solving the standard dual problem. More precisely, we are seeking for the minimal set of documents without affecting significantly the quality of the learning procedure.

To determine the reference set we defined a ranking for the images according to their annotations. The rarer a concept, the higher the score of its positive instances will be. We cut the list where the training documents contain at least a specified quantity of positive samples for all categories. We set the minimal amount of positive samples to $p * N$ where N is the number of training images. If a category did not have the minimal amount of positive instances all the samples were included. The resulted subset of training images using $p = 0.01$ contained only 6260 images out of the original $15k$ training images. Since the dimension of the combined representation equals with the number of images in the reference set this selection reduced the dimension by more than 50%.

To identify the weight vector β of the basic representations per class we sampled the training set. We used $5k$ images for training and $5k$ images for validation. We trained binary SVM classifiers using the LibSVM package [3] separately for each representation and used grid search to find the optimal linear combination per class.

Table 2. Experimenting on visual descriptors, both the training set and the validation set contained 5k images, we used 16x16 and 48x48 macroblock sizes

	#keypoints	#Gaussians	Pooling(s)	$Dim_{FisherVec.}$	MAP
HOG	15k	512	full	98304	0.2433
HOG	2k	512	HL	98304	0.2170
HOG	15k	512	3x1	3*98304	0.2399
HOG	15k	512	full+HL+3x1	5*98304	0.2517
Color	15k	256	full	49152	0.2106
Color	2k	256	HL	49152	0.2092
Color	15k	256	3x1	3*49152	0.2131
Color	15k	256	full+HL+3x1	5*49152	0.2233
Color+HOG	15k	256 & 512	full+HL+3x1	5*(98304+49152)	0.2771

Table 3. Biclustering of Flickr tags and images

	Method	MAP
baseline	JS div.	0.2554
Bicluster	JS div.	0.2185
Bicluster	JS + Vis	0.3133

5 Experiments and Results

All of our submissions used both visual and textual features. The main differences were the number of training images used for classification and the size of the reference set. All the runs included the following basic representations: HOG based Fisher vectors (1x1,3x1,Harris-Laplacian), Color moment based Fisher vectors (1x1,3x1,Harris-Laplacian) and Jensen-Shannon divergence using Flickr tags as probability distributions (Table 5).

5.1 Experiments on the validation set

In order to determine the parameters of the combined representation we experimented on the basic features using a subset of the training set. It can be seen in Table 2 that color moment and HOG descriptors complement each other. Although the average number of keypoints detected by Harris-Laplacian was considerably less than for the rest of the poolings (average $2k$ vs. $15k$ descriptors per image), we measured small performance differences between them. For Flickr tags we tested three methods (Table 3). We selected the top 25,000 Flickr tags as vocabulary. The refined biclustering using visual similarity and Jensen-Shannon divergence outperformed Jensen-Shannon divergence and the purely tag based biclustering. We experimented with the parameter p for proper reference set selection over the best combined representation including all visual similarity values and Jensen-Shannon divergence. It can be seen in Table 1 that the performance loss was negligible even using less than half of the features. If we left only the 11.9% of the training set as reference set the performance dropped significantly.

Table 4. Dimension of the basic representations

	Dimension	sparsity
HOG Fisher vector	98304	dense
Color Fisher vector	49152	dense
Flickr tag tf	25000	very sparse
Biclustering	2000	dense

Table 5. MiAP, GMiAP and F-ex results of basic runs

	TrainSVM	RefSet	Weightn.	MiAP	GMiAP	F-ex
jchfr15k	15k	15k	fix	0.4258	0.3676	0.5731
jch10ksep	10k	6.2k	adapt.	0.4003	0.3445	0.5535
jchb10ksep	10k	6.2k	adapt.	0.3972	0.3386	0.5533

Table 6. MiAP, GMiAP and F-ex results of late fusion runs

	base runs	MiAP	GMiAP	F-ex
jchaggsep	jchfr15k + jch10ksep	0.4212	0.3655	0.5724
jchbicwelf	jchfr15k + bic	0.4173	0.3611	0.5717

5.2 Results

In $jch10ksep$ we used the ranked reference set with 6260 images and an annotation category based weighting scheme for the combination (19 different weight vectors). We trained binary SVM classifiers per class using a reduced training set containing only $10k$ images.

Addition to $jch10ksep$, in $jchb10ksep$ we added a refined biclustering representation with $2k$ clusters to the common representations. Notice that by biclustering the dimension of the representation was significantly the lowest of all (Table 4).

Our best performing run $jchfr15k$ used the entire training set as reference set and the binary SVM models were trained on the whole training set ($15k$) per class. The adopted weight vector β were the same for each class. It is worth to mention that we experienced increase in computational time in comparison to $jch10ksep$ or $jchb10ksep$. The reason for the nonlinear increase is that by $jchfr15k$ we used the whole training set as reference set and the binary SVM classifiers were trained on the entire training set.

Our second best performing run $jchaggsep$ was a combination of $jchfr15k$ and $jch10ksep$ (Table 6). We simply averaged the predictions of the runs. In $jchbicwelf$ we aggregated the output of the biclustering based classifier and the $jchfr15k$ using a linear combination learned previously on the training-validation set.

6 Conclusions

Our approach for ImageCLEF 2012 Photo Annotation task employed various representations of the images based on different visual and textual modalities. We extracted several Ficher vectors using a grayscale and a color patch descriptor. We applied a biclustering method to cluster the images and their Flickr tags. We combined the different descriptors and representations before the classification. This combination procedure included a transformation, a feature aggregation and a selection step.

References

1. J. Ah-Pine, C. Cifarelli, S. Clinchant, G. Csurka, and J. Renders. XRCEs Participation to ImageCLEF 2008. In *Working Notes of the 2008 CLEF Workshop*, 2008.
2. E. Bodzsár, B. Daróczy, I. Petrás, and András A. Benczúr. GMM based fisher vector calculation on GPGPU. http://datamining.sztaki.hu/?q=en/GPU-GMM.
3. C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011.
4. C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20, 1995.
5. G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, volume 1, page 22. Citeseer, 2004.
6. Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. *CVPR 2005*, 2005.

7. B. Daróczy, A. Benczúr, and R. Pethes. SZTAKI at ImageCLEF 2011. In *Working notes of CLEF 2011*, 2011.
8. I.S. Dhillon, S. Mallela, and D.S. Modha. Information-theoretic co-clustering. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 89–98, 2003.
9. C. Harris and M. Stephens. A combined corner and edge detector. *Alvey Vision Conference*, page 147151, 1988.
10. D.G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, volume 2, pages 1150–1157, 1999.
11. K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *IJCV*.
12. Stefanie Nowak. New Strategies for Image Annotation: Overview of the Photo Annotation Task at ImageCLEF 2010. In *Cross Language Evaluation Forum , ImageCLEF Workshop, 2010*, 2010.
13. F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07*, pages 1–8, 2007.
14. A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. Simplemkl. *Journal of Machine Learning Research*, 9, 2008.
15. C. Schmid S. Lazebnik and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New York, June 2006*, 2006.
16. Bernard Schölkopf. The kernel trick for distances. pages 301–307. MIT Press, 2000.
17. Dávid Siklósi, Bálint Daróczy, and András A. Benczúr. Content-based trust and bias classification via biclustering. In *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*, WebQuality '12, pages 41–47, New York, NY, USA, 2012. ACM.
18. G. Csurka T. Mensink, F. Perronnin, J. Snchez, and J. Verbeek. LEAR and XRCEs participation to Visual Concept Detection Task at ImageCLEF 2010. In *Working Notes for the CLEF 2010 Workshop*, 2010.
19. Bart Thomee and Adrian Popescu. Overview of the imageclef 2012 flickr photo annotation and retrieval task. CLEF 2012 working notes, Rome, Italy, 2012.
20. K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.