

SZTAKI @ ImageCLEF 2011*

Bálint Daróczy Róbert Pethes András A. Benczúr

Data Mining and Web search Research Group, Informatics Laboratory
Computer and Automation Research Institute of the Hungarian Academy of Sciences
{daroczyb, rpethes, benczur}@ilab.sztaki.hu

Abstract

We participated in the ImageCLEF 2011 Photo Annotation and Wikipedia Image Retrieval Tasks. Our approach to the ImageCLEF 2011 Photo Annotation is based on a kernel weighting procedure using visual Fisher kernels and a Flickr-tag based Jensen-Shannon divergence based kernel. We trained a Gaussian Mixture Model (GMM) to define a generative model over the feature vectors extracted from the image patches. To represent each image with high-level descriptors we calculated Fisher vectors from different visual features of the images. These features were sampled at various scales and partitions such as Harris-Laplace detected patches, scale and spatial pyramids. We calculated distance matrices from the descriptors of train images to combine different high-level descriptors and the tag based similarity matrix. With this uniform representation we had the possibility to learn the natural weights for each category over the different type of descriptors. This re-weighting resulted 0.01838 MAP increase over the average kernel results. We used the weighted kernels for learning linear SVM models for each of the 99 concepts independently. For the Wikipedia Image Retrieval Task we used the search engine of the Hungarian Academy of Sciences as our information retrieval system that is based on Okapi BM25 ranking. We calculated light Fisher vectors to represent the content of the images and performed nearest-neighbour search on them.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [Database Management]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

SIFT, Color moments, Gaussian mixtures, Fisher vector, kernel methods, SVM

1 Introduction

In this paper we describe our approach to the ImageCLEF 2011 Photo Annotation and Wikipedia Image Retrieval evaluation campaigns [6, 7]. Our image classification is based on a combination

*This work was supported by the EU FP7 Projects LAWA (Large-Scale Longitudinal Web Analytics).

of visual and textual (Flickr tag) information defining uniform kernel matrices. We measured the similarity between the set of image tags using the Jensen-Shannon divergence. We extracted normalized Fisher vectors to calculate the distance kernel for the content of the images. In Wikipedia Image Retrieval Task we used the search engine of the Hungarian Academy of Sciences [2] as our information retrieval system that is based on Okapi BM25 [8]. We extracted light Fisher vectors for each image to measure image similarity.

2 Photo Annotation Task

To reduce the effect of noise on codebook generation and bag-of-words modeling, we smoothed the images. Although the images are not of the same resolution, we did not rescale them. One of the reasons was to avoid adding noise. In addition, with a properly normalized bag-of-words modeling we did not need to calculate fixed number of samples per image. We used feature vectors to describe the visual content of an image by approximately 13000 descriptors per image per modality. We sampled the patches with dense multi-scale grid and Harris-Laplace point detection. We calculated SIFT (Scale Invariant Feature Transform[5]) and RGB color descriptors for each patch. For each type of low-level descriptor we trained a Gaussian Mixture Model (GMM) with 256 Gaussians. The training of GMM models with about 3 million training points took 20 minutes per descriptor with our open-source CUDA GMM implementation [3]. For the SIFT descriptors the training was performed after reducing the dimension of descriptors to 80 by PCA. By the Color moments the dimension reduction resulted performance loss so we did not adopted PCA for it. The normalized Fisher gradient vector computed from GMM of SIFT descriptors is a well known technique to represent an image with only one vector per pooling (we used 1x1, 1x3 and 2x2 spatial pyramids [9]). We also calculated Fisher vectors on the Harris-Laplacian detected corner descriptors. Our overall procedure is shown in Fig. 1.

Our GMM procedure is based on the standard expectation maximization (EM) algorithm with a non-hierarchical structure. We resolved the well known vulnerability of EM algorithm to underflow especially computing the conditional probabilities with large (50+) codebooks in fp32/fp64 precision. This is a limitation of GPGPU cards, additionally in fp64 they are usually more than twice as slow as in fp32. Since it is not sufficient enough to use logarithm instead of values we implemented a magnitude summation algorithm. Read the details in our paper[3]. Our source code along with preprocessed GMM models and codes for Fisher vector calculation is available free for research use at <http://datamining.sztaki.hu/?q=en/GPU-GMM>.

The Fisher vectors can also be computed parallel in $k \cdot D$ independent calculations where D is the dimension of the low-level features and k is the number of Gaussians. If neither the dimension nor the number of clusters is more than the maximal number of threads for a GPU block, the computational time depends only on the number of low-level features. Our implementation with calculating all the gradients of the sampling points is seven-times faster than a well-tuned locally optimal Fisher vector implementation on a fast CPU and 44-times faster if we calculate the same algorithm on the CPU[3]. The calculation of Fisher vectors took about 1.5 hours per modality. We extracted 9 Fisher vectors per image for each pooling: one Fisher vector for all the patches, one for the Harris-Laplace detected points, three for the 1x3 and four for the 2x2 spatial partitions. We normalized the resulted vectors with Fisher information matrix, power and L2 normalization. Worth to mention, as our feature extraction methods eventuated high number of descriptors for each part of the image and our Fisher calculation method is not cutting the lower probabilities, the resulted Fisher vectors were highly non-sparse without any normalization.

2.1 Pre-Calculated kernel combination for linear SVM

Learning linear SVM models on Bag-of-Words models is a widely used technique[10, 1]. One of the main problem is to define the kernel function between the training instances. We used pre-computed normalized distance kernels instead of Fisher kernels because our preliminary experiments showed better annotation performance with different training sets. Beside the classification

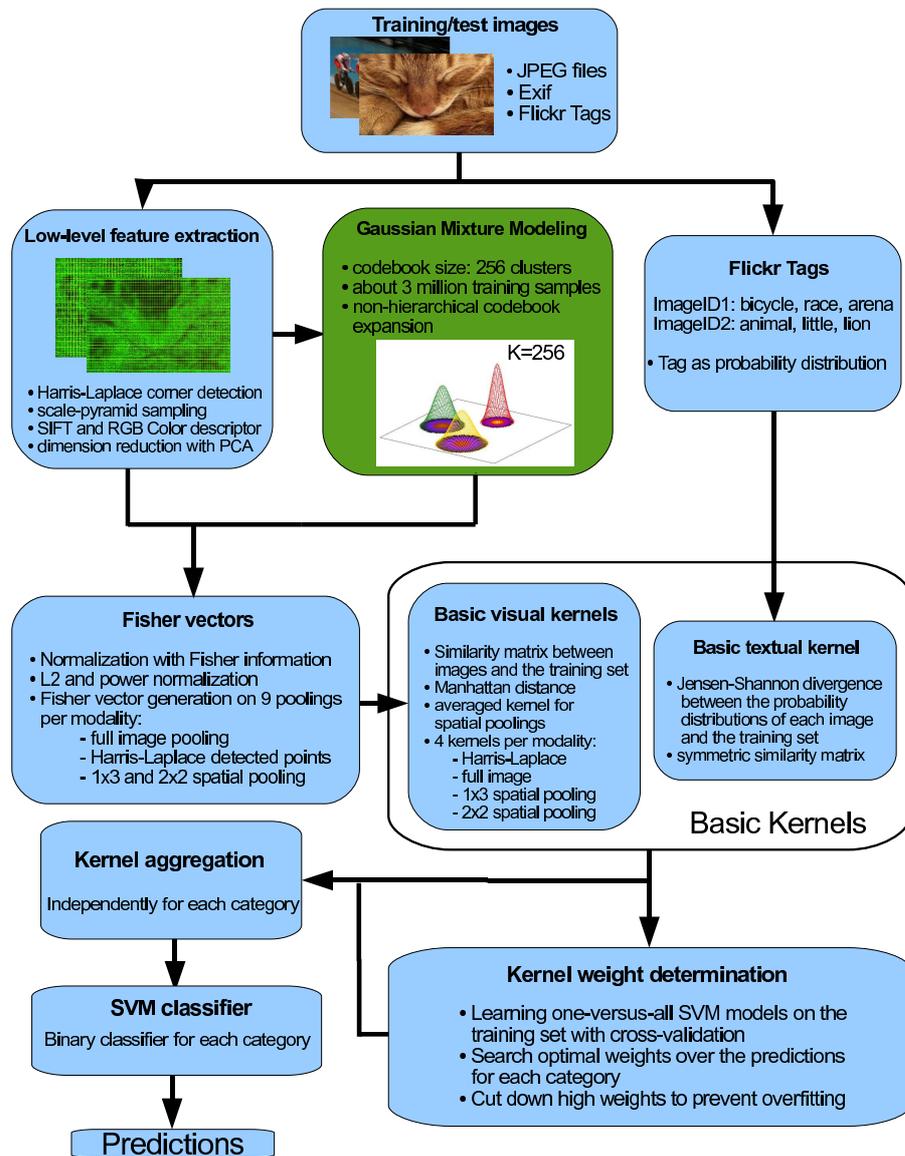


Figure 1: Feature extraction and classification procedure (Photo Annotation)

performance gain the dimension of the kernels are less than the dimension of a regular Fisher vector (*dimension of the kernel* = #training images = $8k$ vs. *dimension of Fisher vector* = $40k - 49k$). In addition, with pre-computed kernels we had the ability to combine textual and visual kernels without increasing the dimension of the learning problem. We averaged the corresponding basic kernels of the spatial pyramids. This resulted four kernels per low-level descriptor per image (Harris-Laplace, full image, 1x3 and 2x2 spatial pyramid). Additionally we measured the distance between two image according to their tags with Jensen-Shannon divergence. Our choice was inspired by the similarity properties of Jensen-Shannon against Kullback-Leibler divergence.

To learn the weights of the different basic kernels we splitted the training set into two parts on account of sparse annotation. We trained SVM models per category ([4]) on the two part of the data and evaluated it on the other half of the training set. We used the independently evaluated predictions for the basic kernels and determined a combination between them for each category. We used the existing Flickr tags to create probability distributions. If an image had no tag we set its similarity to be zero if measured against itself and one if measured against the other images. We assumed that the specified weights are linked to the ideal combination of the basic kernels. Our final weighted kernel for each category c between images X and Y was

$$K(X, Y)_c = \frac{1}{|K|} \sum_{k=1}^K \alpha_{ck} \sum_{t=1}^T K_k(X, I_t) * K_k(Y, I_t). \quad (1)$$

The $K_k(X, I_t)$ denotes the basic kernels, I_t is the t th training image and T is the number of training images of the collection. For the visual Fisher vectors we measured the similarity between two vectors with Manhattan distance. For the Flickr tag based probability distributions we adopted the Jensen-Shannon divergence as similarity measure.

$$K_{k_{visual}}(X, I_t) = \frac{dist_{Manhattan}(F_k(X), F_k(I_t))}{\max * \arg \max_t K_k(X, I_t)} \quad (2)$$

$$K_{k_{textual}}(X, I_t) = dist_{Jensen-Shannon}(X, I_t) \quad (3)$$

where $F_k(X)$ denotes the Fisher vector of X for the k th pooling.

Since the output of our classifier was a summarized values of the weighted dot-products of the support vectors and the test instances, we used the sigmoid function to map the output of the SVM classifier to a floating point prediction between zero and one.

$$Prediction_{float} = \frac{1}{1 + exp^{-1 * svm_{output}}} \quad (4)$$

Our method gained 1.8 % increase in MAP over the average sum of the basic kernels if we also included a textual based kernel beside the visual kernels. As seen in Table 1 if we adopted the kernel weighting only to the basic visual kernels we measured a 0.3 % increase.

For the example-based evaluation we needed to define a mapping from the floating point predictions into a binary annotation. We applied two strategies. In the first method we shifted the borderline between the positive and the negative samples till the annotation on the training set had the highest precision and recall. In the second method we assumed that the relative occurrence of a category in the training and the test set were similar and shifted the borderline according to it. The previous had much higher F-score (0.545341 vs. 0.593088) and higher Semantic R-Precision (0.70853 vs. 0.71928). Worth to mention, from our submissions the averaged visual kernel had the highest Semantic R-Precision (0.72902450).

Table 1: Photo Annotation results

	Kernel aggregation	MAP	EER	AUC
visual + textual run3	weighted	0.438744	0.243574	0.827621
visual + textual run2 (we_cssj)	weighted	0.436294	0.241691	0.827747
visual + textual run1 (avg_cssj)	average	0.420406	0.243885	0.828322
visual run2	weighted	0.369688	0.263449	0.806691
visual run1 (avg_cns)	average	0.367054	0.264328	0.805142
textual run1 (jensen)	only one	0.345616	0.338127	0.717966

3 The Wikipedia Image Retrieval Task

We used the Hungarian Academy of Sciences search engine [2] as our information retrieval system based on Okapi BM25 ranking [8]. We applied the English, German and French annotations and articles independently for indexing. We made no differentiation between the title and the body of the annotation.

Since file names often contain relevant keywords and also often as substring, we gave score proportional to the length of the matching substring. Since the indexing of all substrings is infeasible, we only performed this step for those documents that already matched at least one query term in their body.

For the WikipediaMM task we also deployed query expansion by an online WordNet¹. We added groups of synonyms with reduces weight so that only the score of the first few best performing synonym was added to the final score to avoid overscoring long lists of synonyms.

Nearest neighbor search was performed over light Fisher vector. We call it light Fisher vectors because we calculated descriptors only on a dense grid (16 pixel step sampling) and we did not extracted different poolings. We used the same Gaussian Mixture Model for the RGB color moment descriptors as in the Photo Annotation task.

Our Relevance Feedback method used the first 10 results of the aggregated score of the three language query results. We calculated the Jensen-Shannon distance from the first 10 hits to the lower ranked hits with weight according to their rank. For the documents ranked lower ($i > 9$) the new score was

$$score_i = \frac{1}{i+1} * \sum_{j=0}^9 (1 - dist_{Jensen-Shannon}(D_i, D_j)) \quad (5)$$

where D_i and D_j denotes the probability distribution of the i th and j th document.

We found the text based score more accurate. Therefore we adopted our Relevance Feedback procedure using the visual similarity of the first hundred hits to re-rank the documents. This resulted 0.3 % increase in MAP (0.2167 vs. 0.2136).

4 Conclusions

For Photo Annotation Task, we successfully applied visual and textual kernel matrices as instance matrices for SVM learning. We used our own implementations for Low-level feature extraction, to train GMM models and calculate Fisher vectors. Our kernel weighting procedure resulted 1.8 % improvement over the average combination of the textual and visual kernels. We are planning to including new pre-computed kernels in the system and applying generative models to improving the determination of weights for kernel aggregation.

For image retrieval task, our Relevance Feedback method improved the baseline Okapi BM25 based textual system. We also plan to strengthen our retrieval results by using more sophisticated methods for text and image retrieval fusion.

¹<http://wordnet.princeton.edu/>

References

- [1] J. Ah-Pine, S. Clinchant, G. Csurka, and Y. Liu. XRCEs Participation in ImageCLEF 2009. 2009.
- [2] András A. Benczúr, Károly Csalogány, Eszter Friedman, Dániel Fogaras, Tamás Sarlós, Máté Uher, and Eszter Windhager. Searching a small national domain—preliminary report. In *Proceedings of the 12th World Wide Web Conference (WWW)*, Budapest, Hungary, 2003.
- [3] 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>.
- [4] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [5] D.G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, volume 2, pages 1150–1157, 1999.
- [6] S. Nowak and M. Huiskes. New strategies for image annotation: Overview of the photo annotation task at imageclef 2010. *Working notes of CLEF*, 2010, 2010.
- [7] A. Popescu, T. Tsirikia, and J. Kludas. Overview of the wikipedia retrieval task at imageclef 2010. *Working Notes of CLEF*, 2010, 2010.
- [8] Stephen E. Robertson and Karen Sparck Jones. Relevance weighting of search terms. In *Document retrieval systems*, pages 143–160. Taylor Graham Publishing, London, UK, UK, 1988.
- [9] 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.
- [10] 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.