UPMC/LIP6 at ImageCLEFannotation 2010

Ali Fakeri-Tabrizi, Sabrina Tollari, Nicolas Usunier, Massih-Reza Amini, and Patrick Gallinari

Université Pierre et Marie Curie - Paris 6,
Laboratoire d'Informatique de Paris 6 - UMR CNRS 7606
4 place Jussieu, 75252 Paris, France
firstname.lastname@lip6.fr

Abstract. In this paper, we present the LIP6 annotation models for the ImageCLEFannotation 2010 task. We study two methods to train and merge the results of different classifiers in order to improve annotation. In particular, we propose a multiview learning model based on a RankingSVM. We also consider the use of the tags matching the visual concept names to improve the scores predicted by the models. The experiments show the difficulty of merging several classifiers and also the interest to have a robust model able to merge relevant information. Our method using tags always improves the results.

Key words: SVM, Multi-Class Multi-Label Image Classification, Imbalanced Class Problem, Semi-Supervised Learning, Transductive Learning, Visual Concepts, Ranking SVM

1 Introduction

Last year, in ImageCLEFannotation 2009, we focused on how to deal with imbalanced data [2]. Instead of training a standard SVM, we have used a Ranking SVM in which the chosen loss function is helpful in the case of imbalanced data. This year, in ImageCLEFannotation 2010 [5], we additionally focus on how to use different visual feature spaces in the same model using supervised and semi-supervised learning. We also consider to use the tags associated to the images.

In this work, we consider two models that merge the predictions of several classifiers. The first model takes the mean of the predicted score of the classifier, where each classifier is trained on a specific visual feature space using the labeled data provided for the competition. The second model makes use of additional unlabeled data to train the classifier in the semi-supervised, multiview paradigm [1]. In our case, the representation of an image in a given visual feature space is a *view* of the image. Semi-sueprvised learning is carried out by first learning classifiers on each view with the labeled data, and then enforcing these classifiers to make similar predictions on the unlabeled data. As in our first model, the final prediction is the mean of the scores predicted by the different classifiers.

In addition to visual descriptors, the text associated to an image is often relevant to improve image retrieval. There are only a few works [7] studying the

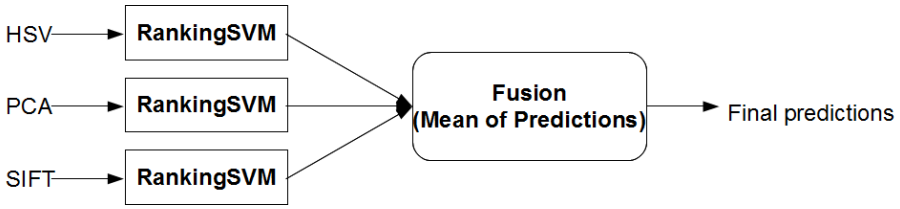

Fig. 1. Schema of the fusion model

correlation between the names of the visual concepts and text. The ImageCLEFannotation 2010 task gives us the opportunity to study the links between the names of visual concepts and tags.

The remainder of the paper is organized as follows. In Section 2, we propose our models for image annotation using multiple views. Section 3 describes our method which uses tags to improve annotation. The experiences are illustrated in Section 4. The conclusion and perspectives are presented in Section 5.

2 Annotation Models

2.1 Using RankingSVM in Imbalanced Dataset Case

The data for image annotation is often highly imbalanced: for many classes, there are only very few positive (or negative) examples. As we showed in [2], standard learning algorithms like SVMs may be biased towards the majority class in such cases, but more involved algorithms like RankingSVM may help to overcome this problem. The Ranking SVM does not strive to separate the two classes, but rather learns a score function that gives greater scores to positive examples than to negative ones. We choose to optimize the Area under the ROC Curve (AUC) as in [4]. The AUC is the probability that a positive example has a greater score than a negative one.

Strictly speaking, a Ranking SVM does not learn a classifier. A classifier can however be obtained by comparing the scores with an appropriate threshold. In the following, the classifier is obtained by comparing the score to 0: if an observation x is such that $\langle w, \phi(x) \rangle > 0$, then we predict that x is in the positive class, otherwise we predict that it is in the negative class. Although this choice may not be optimal, it is a simple decision rule that gives good results in practice.

Last year [2], we focused on how to deal with imbalanced data. This year, in the ImageCLEFannotation 2010, we focus on how to train and merge the results of different classifiers to improve annotation, and specially on how to use different visual feature spaces in the same model.

To obtain a baseline, we perform several RankingSVM on different visual features, then we create a fusion of the outputs using an arithmetic mean. Figure 1 describes the *fusion* model we use in the ImageCLEFannotation2010 task.

2.2 Multiview Learning Model

Images can be described in several visual feature spaces (like SIFT, HSV, ...). Each of these representations, or *views* of an image, is very informative about the label of the image, but the different views provide rather independent information. Semi-supervised multiview learning aims at using these characteristics of the different views in order to improve the accuracy of the classifiers. The main principle is as follows: after training classifiers on each view independently (using standard supervised learning, in our case RankingSVMs), these classifiers are modified so that they predict, as much as possible, the same class labels on the unlabeled data. In our multiview learning model, we get the different views

The algorithm we use for the semi-supervised procedure is an iterative procedure. After the initial training step of the different classifiers on each view, each iteration of the algorithm consists of (1) predicting on all available unlabeled examples, then (2) adding to the training set (and removing from the unlabeled set) all examples for which all classifiers agree on a given class label. Equivalently, the unlabeled examples are given predicted labels based on the unanimous vote of the different classifiers. After step (2), the classifiers are re-trained on the new training set, and steps (1) and (2) are repeated. The algorithm stops when there are no more unlabeled examples on which all the classifiers agree. The process is described in Figure 2.

The rationale of this procedure is that when all the classifiers agree on a class label, the label is likely to be correct. Then, adding those examples to the training set gives us more training examples. Moreover, because we systematically add to the training set new examples on which the classifiers already agree, the procedure tends to increase the agreement between the different classifiers on the original unlabeled dataset.

In practice, the semi-supervised learning procedure improves performance when the classifiers are sufficiently accurate so that most training examples added at each iteration are correctly classified. If this is not the case, multiview learning may actually decrease performance, since we train new classifiers on more noisy training sets.

In the ImageCLEFannotation 2010 task, the observations are composed of the multiple views of image: each view can be considered as a feature type. We perform RankingSVM on each feature type using the training dataset. The training dataset is the labeled dataset and the test dataset is the unlabeled dataset.

3 Using Tags to Improve Visual Concept Annotation

To improve image annotation, the image associated text can be used. Free texts, tags and visual concepts are three kinds of text we can use for image annotation, but there is a big difference between them.

1. Free text is based on a very large vocabulary, it can be composed of a description of the image - this type of free text is often relevant to improve the

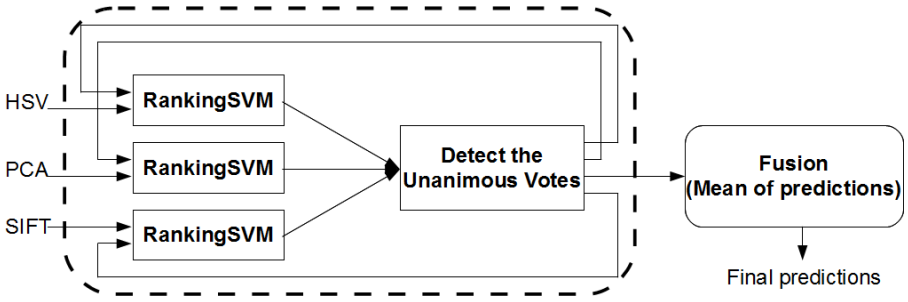

Fig. 2. Schema of our multiview model

image classification - or it can be composed of the associated text (like web pages) - this type of free text is often fewly relevant.
2. Tags are a little bit more specific then free text. When an image is associated with a tag, this tag is most of the time relevant for this image, but when a tag is not associated with an image, it should not mean that this tag is not relevant for the image.
3. Finally, visual concepts come from a very specific vocabulary, and contrary to tags, when a visual concept is not associated with an image, it means that this visual concept is not relevant for this image.

We then deduce that if the name of a visual concept tagged an image, then this image should be annotated with this concept, but the contrary is false. Our method is based on this idea. If a tag maches the name of a visual concept, then the images associated with this tag will have their prediction scores for this concept set to a given value we called UP. In order to improve the matching between the name of the visual concepts and tags, we first apply - on both concept names and tags - standard text processing, such as porter stemming algorithm [6]. For example, if a test image has a prediction score of 0.4 with the classifier of the concept *Sunny* and if its associated tags contain the text *sunni* (which is the stemmed word for *Sunny*), then its prediction score is altered to the value of UP, which is estimated using the validation set.

4 Experiments

The corpus is composed of a training set of 8000 Flickr images and a test set of 10000 Flickr images. We split the training set in a training set of 5000 images and a validation set of 3000 images[1]. Each image is annotated in average by 12 visual concepts chosen among the 93 hierarchical visual concepts.

4.1 Visual Features

We extract three different types of visual features from each image.

[1] The images in the validation set are the same as ImageCLEFannotation 2009

HSV First, we segment images into 3 horizontal regions and extract *HSV* features. For each region, we compute a color histogram in the HSV space. We believe that these visual descriptors are particularly interesting for general concepts (i.e. not objects), such as: sky, sunny, vegetation, sea and etc.

SIFT Second, we extract *SIFT* keypoints, and then we cluster them to obtain a visual dictionary: we extract the SIFT keypoints of each image. To reduce the size of the dictionary and avoid duplicate keypoints, the keypoints are clustered with a nearest-neighbors algorithm, to obtain 1024 clusters. Then, each cluster represents a visual word in the dictionary, and each image is indexed using this dictionary.

Mixed+PCA Third, we use a concatenation of various visual features from 3 labs proposed by the AVEIR consortium [3] reduced using a PCA (*Mixed + PCA*): this space is composed of the concatenation of the visual features from 3 labs: 51 dimensions HSV histograms from LIP6 lab, 150 dimensions of entropic features from LSIS and 120 dimensions features composed of a combination of color, texture and shape from PTECH lab. This space is transformed using a PCA. Then we keep the first 180 dimensions which correspond to 98% of the cumulative variance.

4.2 Text Features

We first apply - on both concept names and tags - standard text processing, principally stemming using Porter stemming algorithm [6]. Table 1 gives different information on the sets in function of the use (or not) of stemming. For example, in the training set, without stemming, there are 39 (out of 93) concepts matching at least one tag in the documents, whereas there are 69 (out of 93) concepts matching using stemming. We can see that it is difficult to match tags and concept names without stemming and also with stemming, maybe because the names of the concepts are chosen by a specialist of visual concept detection whereas tags are chosen by Flickr users. For example, a user will never tag an image with "No Visual Season", "Out of focus" or "Neutral Illumination".

Another information is the number of predicted scores modified by the tags. For example, in the validation set, there are 3000 images and 93 concepts, so there are 279000 predicted values. Among those, only 1829 are modified when we use stemming. On average over the different concepts, 19.7 predicted scores are modified. Even though the tags affect rather few values perc concept, we will see in the next section that those modifications are most often relevant.

Figure 3 compares the number of images which contain a tag corresponding to a given concept (for example, if we consider the concept "Sunny", it is the number of images which are tagged with "Sunny"), and the number of images relevant for this concept in the training set (for example, if we consider the concept "Sunny", it is the number of images which are labelled are positive for the concept "Sunny" in the ground truth of the training set). This figure shows that, contrarily to the intuition, the number of images tagged with a given concept is not correlated to the number of relevant images for this concept.

Table 1. Information on the different sets

	Training Set	Validation Set	Test Set
Number of images	5000	3000	10000
Number of concept names matching tags			
... without stemming	39	39	41
... with stemming	69	67	70
Number of prediction scores modified...			
... without stemming	2096	1189	4065
... with stemming	3286	1829	6091

4.3 Experiments on the Validation Set

As we said in the previous sections, we use RankingSVM as the basic learner. We have trained RankingSVMs on all three types of features in a fully supervised setting (the *fusion* model) or in the semi-supervised setting (the *multiview* model). For both types of training, the different classifiers are merged in a single model by taking the mean of the predicted scores. These scores are finally normalized using a gamma distribution. We optionally use tags.

Table 2. Results on validation set

Run	EER	AUC	MAP
SIFT	0.326	0.731	0.251
Fusion	0.305	0.754	0.278
Multiview	0.321	0.732	0.260
SIFT+tags	0.309	0.753	0.312
Fusion+tags	0.287	0.776	0.330
Multiview+tags	0.301	0.760	0.310

Table 2 compares the results of the different models on the validation set. The fusion model gives better results than the SIFT and multiview models. Our method using tags always improves significantly the scores.

Figure 4 compares the AUC scores between fusion and fusion+tags on validation set for each concept. We can notice that the use of the tags never decreases the AUC of a concept. This means that when a tag is put on an image, it is most often relevant for this image.

Figure 5 compares the AUC obtained with and without stemming depending on the value of UP. This figure shows that using Porter stemming algorithm significantly improves the results. We can notice that the best scores are obtained for $UP = 1$. This confirms that if a tag corresponding to a concept is associated to an image, than this image should be labeled by this concept.

Figure 6 compares the behaviour of the different models depending on UP. We remark that the fusion model always gives the best results. We also note

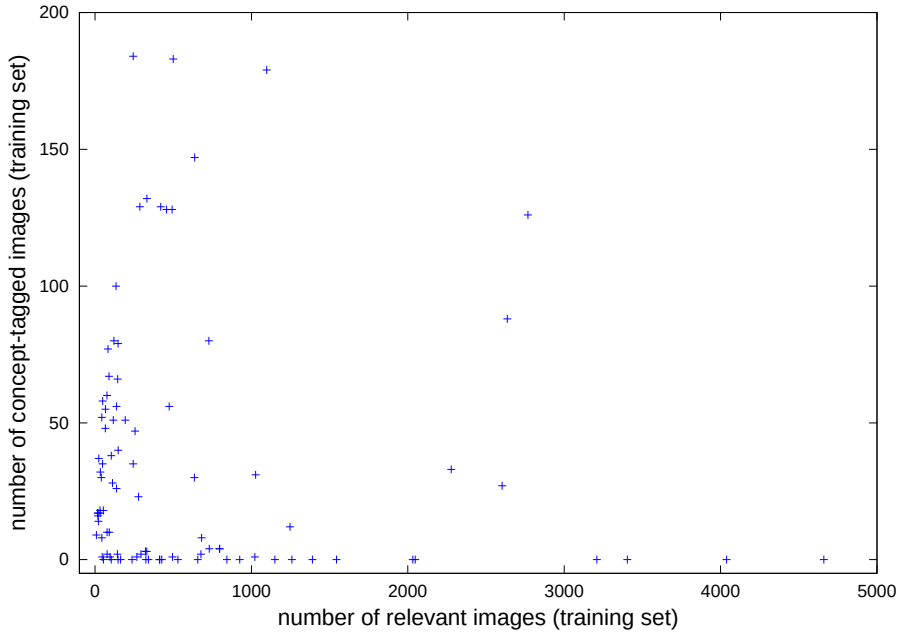

Fig. 3. Comparison of the number of images in the training set relevant for a given concept, and of the number of images which contain a tag corresponding to this given concept in the training set. Each cross corresponds to a concept

that the multiview model is less accurate than the Fusion or Sift models. This may be due to the use of unanimous vote.

4.4 Submitted Runs and Official Results

In ImageCLEFannotation 2010, we submitted the following 5 runs :

Run1 We perform a RankingSVM using only SIFT features.

Run2 We perform a RankingSVM for each type of features (HSV, SIFT, Mixed + PCA). Then we merge the prediction scores using an arithmetic mean.

Run3 We perform the multiview learning by using three views (HSV, SIFT, Mixed + PCA); each view represents a type of visual features.

Run4 Same as Run2, but here we also consider tags. The prediction score given by the classifier for a given concept is increase up to 1, if the image is tagged with the name of the visual concept.

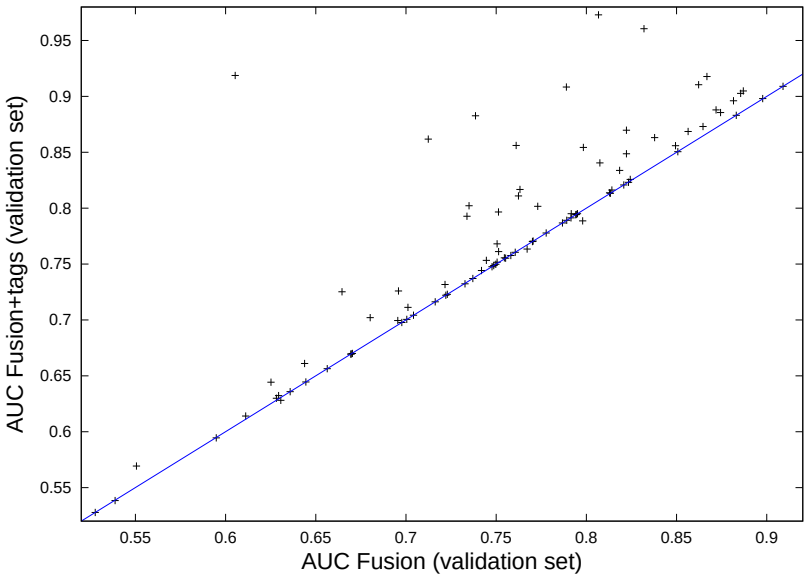

Fig. 4. Comparison of AUC scores between Fusion and Fusion+tags on validation set. Each cross represents the scores for a concept

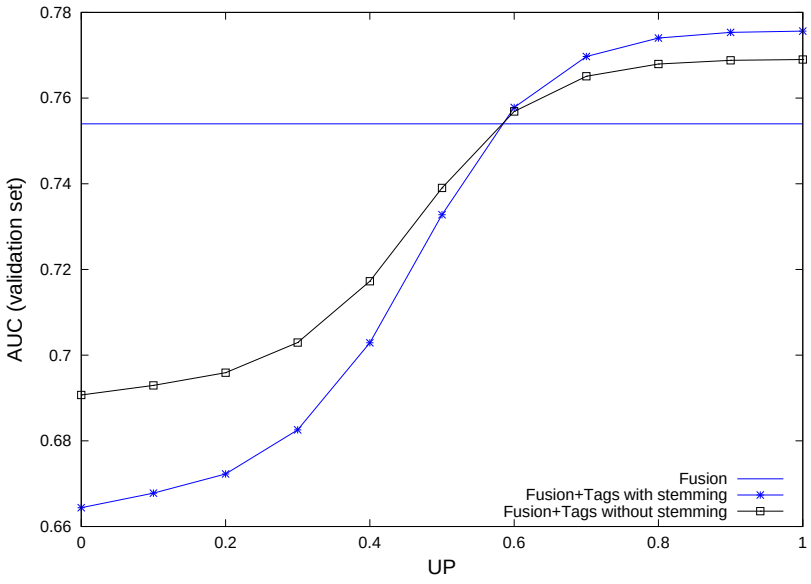

Fig. 5. Comparison of AUC scores with and without stemming in function of the value of UP for the Fusion model

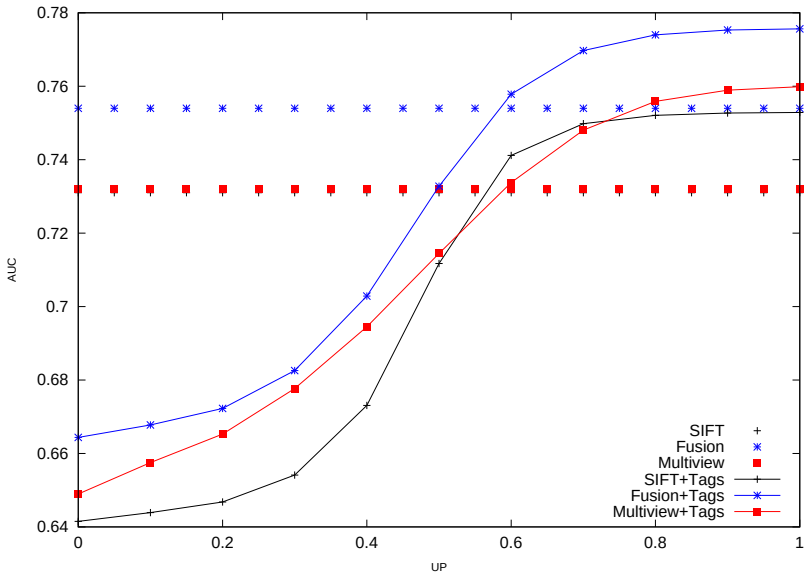

Fig. 6. Comparison of AUC scores in function of the value of UP for the different models using stemming

Run5 Same as Run3, but here we also consider tags. The prediction score given by the classifier for a given concept is increase up to 1, if the image is tagged with the name of the visual concept.

Table 3 gives the official results on test set. We note that the results are close to the results obtain by a random run. We can conclude that there might be some mistakes in our process (maybe in the final step where we have to make the prediction scores in a given format), but not in our methods because the results on validation set are reasonable (see Table 2). In Table 3, we can notice that the use of the tags improves significantly the results.

Table 3. Official results

	Run	MAP	Average F-ex	Ontology score	EER	AUC
run1	SIFT	0.145	0.127	0.328	0.502	0.497
run2	Fusion	0.146	0.174	0.348	0.497	0.504
run3	Multiview	0.148	0.173	0.348	0.498	0.502
run4	Fusion+tags	0.182	0.184	0.351	0.463	0.559
run5	Multiview+tags	0.180	0.186	0.351	0.464	0.557

5 Conclusion

We proposed two models to train and merge the results of different classifiers in order to improve annotation. We described a multiview learning method for image annotation in which each view is a visual feature type. We also merged the predicted scores for all feature types. The experiments showed that our multiview model is close to, but less effective than, the fusion model. Maybe the voting part should be modified to obtain higher performances.

We also considered the use of the tags matching the visual concept names to improve the scores predicted by the models. Our method using tags always improved the results of the classifiers. Our study of the links between visual concept names and tags showed that when an image is tagged with a concept name this image should be labeled with this concept.

As perspectives, we can modify the voting part in the multiview model to avoid adding noisy examples to the training set. Moreover, we intend to try our models with other feature types, such as text, to study the evolution of the performances.

Acknowledgment

This work was partially supported by the French National Agency of Research (ANR-06-MDCA-002 AVEIR project).

References

1. C. Christoudias, Raquel Urtasun, and Trevor Darrell. Multi-view learning in the presence of view disagreement. In *Uncertainty in Artificial Intelligence (UAI-08)*, pages 88–96, Corvallis, Oregon, 2008. AUAI Press.
2. A. Fakeri-Tabrizi, S. Tollari, N. Usunier, and P. Gallinari. Improving image annotation in imbalanced classification problems with ranking svm. In *Multilingual Information Access Evaluation Vol. II Multimedia Experiments: Proceedings of the 10th Workshop of the Cross–Language Evaluation Forum (CLEF 2009)*, 2010.
3. H. Glotin, A. Fakeri-Tabrizi, P. Mulhem, M. Ferecatu, Z.-Q. Zhao, S. Tollari, G. Quenot, H. Sahbi, E. Dumont, and P. Gallinari. Comparison of various aveir visual concept detectors with an index of carefulness. In *CLEF working notes*, 2009.
4. Thorsten Joachims. A support vector method for multivariate performance measures. In *International Conference on Machine Learning (ICML)*, 2005.
5. S. Nowak and M. Huiskes. New strategies for image annotation: Overview of the photo annotation task at imageclef 2010. In *Working Notes of CLEF 2010*, 2010.
6. M. F. Porter. An algorithm for suffix stripping. *Readings in information retrieval*, pages 313–316, 1997.
7. S. Tollari, M. Detyniecki, C. Marsala, A. Fakeri-Tabrizi, M.-R. Amini, and P. Gallinari. Exploiting visual concepts to improve text-based image retrieval. In *European Conference on Information Retrieval (ECIR)*, 2009.