Multimodal semi-supervised learning for image classification

Matthieu Guillaumin, Jakob Verbeek and Cordelia Schmid
LEAR, INRIA Grenoble, Laboratoire Jean Kuntzmann
firstname.lastname@inria.fr

Abstract

In image categorization the goal is to decide if an image belongs to a certain category or not. A binary classifier can be learned from manually labeled images; while using more labeled examples improves performance, obtaining the image labels is a time consuming process.

We are interested in how other sources of information can aid the learning process given a fixed amount of labeled images. In particular, we consider a scenario where keywords are associated with the training images, e.g. as found on photo sharing websites. The goal is to learn a classifier for images alone, but we will use the keywords associated with labeled and unlabeled images to improve the classifier using semi-supervised learning. We first learn a strong Multiple Kernel Learning (MKL) classifier using both the image content and keywords, and use it to score unlabeled images. We then learn classifiers on visual features only, either support vector machines (SVM) or least-squares regression (LSR), from the MKL output values on both the labeled and unlabeled images.

In our experiments on 20 classes from the PASCAL VOC'07 set and 38 from the MIR Flickr set, we demonstrate the benefit of our semi-supervised approach over only using the labeled images. We also present results for a scenario where we do not use any manual labeling but directly learn classifiers from the image tags. The semi-supervised approach also improves classification accuracy in this case.

1. Introduction

The goal of image classification is to decide whether an image belongs to a certain category or not. Different types of categories have been considered in the literature, *e.g.* defined by presence of certain objects, such as cars or bicycles [7], or defined in terms of scene types, such as city, coast, mountain, *etc*. [12]. To solve this problem, a binary classifier can be learned from a collection of images manually labeled to belong to the category or not. Increasing the quantity and diversity of hand-labeled images improves the performance of the learned classifier, however, labeling images is a time consuming task. Although it is possible to label large amounts of images for many categories for research purposes [6], this is often unrealistic, *e.g.* in personal photo organizing applications. This motivates our interest in using other sources of information that can aid the learning process using a limited amount of labeled images.

Tags: desert,nature,landscape,sky
Labels: clouds, plant life, sky, tree

Tags: rose, pink
Labels: flower, plant life

Tags: india
Labels: cow

Tags: aviation, airplane, airport
Labels: aeroplane

Figure 1. Example images from MIR Flickr (top row) and VOC'07 (bottom row) data sets with their associated tags and class labels.

In this work we consider a scenario where the training images have associated keywords or tags, such as found on photo sharing websites like Flickr. Our goal is to learn a classifier for images alone, but we will use the tags associated with labeled and unlabeled images to improve the classifier using a semi-supervised approach. Image tags tend to be noisy in the sense that they might not directly relate to the image content, and typically only a few of many possible tags have been added to each image, as shown in Figure 1. Despite the noisy relation between tags and image content,

they have been found a useful additional feature for fully supervised image categorization [13, 23].

We propose a semi-supervised learning approach to leverage the information contained in the tags associated with unlabeled images in a two-step process. First, we use the labeled images to learn a strong classifier that uses both the image content and tags as features. We use the multiple kernel learning (MKL) framework [18] to combine a kernel based on the image content with a second kernel that encodes the tags associated with each image. This MKL classifier is used to predict the labels of unlabeled training images with associated tags. In the second step we use both the labeled data and the output of the classifier on unlabeled data to learn a second classifier that uses only visual features as input. Our work is different from most work on semi-supervised learning as our labeled and unlabeled data have additional features that are absent for the test data. A schematic overview of the approach is given in Figure 2.

We perform experiments using the PASCAL VOC'07 and MIR Flickr data sets [7, 11] that were both collected from the Flickr website and for which user tags are available. The image sets have been manually annotated for 20 and 38 categories respectively. We measure performance using average precision on these manual annotations. In our experiments we confirm that the tags are beneficial for categorization, and that our semi-supervised approach can improve classification results by leveraging unlabeled images with associated tags. We also consider a weakly-supervised scenario where we learn classifiers directly from the images tags, and do not use any manual annotation. Also in this case our approach can improve the classification performance by identifying images that are erroneously tagged.

In the next section we discuss the most relevant related work, and in Section 3 we present our method in detail. In Section 4 we present the data sets we used in our experiments and the feature extraction procedure. The experimental results follow in Section 5, and we conclude in Section 6.

2. Related work

Given the increasing amount of images that are currently available on the web with weak forms of annotation, there has been considerable interest in the computer vision community to leverage this data to learn recognition models. Examples are work on filtering images found using web image search, or images found on photo sharing sites using keyword based queries [3, 8, 9, 10, 19]. Others have used image captions to learn face recognition models without manual supervision [2], or to learn low dimensional image representations by predicting caption words and can be transferred to other image classification problems [17]. A related approach was taken in [24] where classifiers were learned to predict the membership of images to Flickr groups, and the difference in class membership probabilities were used to define a semantic image similarity.

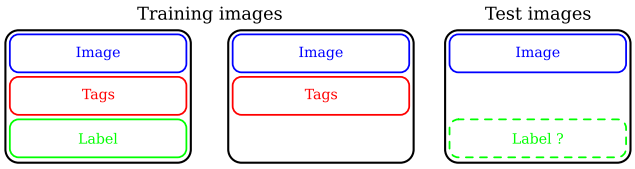

Figure 2. Overview of multimodal semi-supervised classification. Training images come with tags, and only a subset is labeled. The goal is to predict the class label of test images without tags.

Two recent papers that use tagged images to improve image classification performance are closely related to our work. In [13] image tags were used as additional features for the classification of touristic landmarks. We also use image tags to improve the performance of our classifiers, but we do not assume their availability for test images. Wang *et al*. [23] use a large collection of up to one million tagged images, to obtain a textual representation of images without tags. This is achieved by assigning an image the tags associated with its visually most similar images in the set of tagged images. Separate classifiers were learned based on the visual and textual features, and their scores were linearly combined using a third classifier. Our approach differs in that we do not construct a new textual image representation. Rather, we use the strength of classifiers that have access to images and associated tags to obtain additional examples to train a classifier that uses only visual features, thus casting the problem as a semi-supervised learning problem.

There is a large literature on semi-supervised learning techniques. For sake of brevity, we discuss only two important paradigms, and we refer to [5] for a recent book on the subject. When using generative models for semi-supervised learning a straightforward approach is to treat the class label of unlabeled data as a missing variable, see *e.g*. [1, 15]. The class conditional models over the features can then be iteratively estimated using the EM algorithm. In each iteration the current model is used to estimate the class label of unlabeled data, and then the class conditional models are updated given the current label estimates.

This idea can be extended to our setting where we have variables that are only observed for the training data [21]. The idea is to jointly predict the class label and the missing text features for the test-data, and then marginalize over the unobserved text features. These methods are known to work well in cases where the model fits the data distribution, but can be detrimental in cases where the model has a poor fit.

Current state-of-the-art image classification methods are discriminative ones that do not estimate the class conditional density models, but directly estimate a decision function to separate the classes. However, using discriminative classifiers, the EM method of estimating the missing class labels used for generative models does not apply: the EM iterations immediately terminate at the initial classifier.

Co-training [4] is a semi-supervised learning technique that does apply to discriminative classifiers, and is designed for settings like ours where the data is described using several different feature sets. The idea is to learn a separate classifier using each feature set, and to iteratively add training examples for each classifier based on the output of the other classifier. In particular, in each iteration the examples that are most confidently classified with the first classifier are added as labeled examples to the training set of the second classifier, and vice-versa.

A potential drawback of the co-training is that it relies on the classifiers over the separate feature sets to be accurate, at least among the most confidently classified examples. In our setting we find that for most categories one of the two feature sets is significantly less informative than the other. Therefore, using the classifier based on the worse performing feature set might provide erroneous labels to the classifier based on the better performing feature set, and its performance might be deteriorated. In the next section we present a semi-supervised learning method that uses both feature sets on the labeled examples, and we compare it with co-training in our experiments.

3. Multimodal semi-supervised learning

In this section we first present the supervised classification setup (Section 3.1), which forms the basis for the semi-supervised approach (Section 3.2).

3.1. Supervised classification

For our baseline image classification system we follow state-of-the-art image categorization methods [7], and use support vector machines (SVM) with non-linear kernels based on several different image features. The kernel function $k(\cdot,\cdot)$ can be interpreted as a similarity function between images and is the inner product in an induced feature space. The SVM is trained on labeled images to find a classification function of the form

$$f(x) = \sum_i \alpha_i k(x, x_i) + b. \quad (1)$$

For a test image, the class label $y \in \{-1, +1\}$ is predicted as $\text{sign}(f(x))$.

In order to combine the visual and textual representations we adopt the multiple kernel learning (MKL) framework [18], although not making use of its full power. Denoting the visual kernel by $k_v(\cdot,\cdot)$ and the textual kernel by $k_t(\cdot,\cdot)$, we can define a combined kernel as a convex combination of these: $k_c(\cdot,\cdot) = d_v k_v(\cdot,\cdot) + d_t k_t(\cdot,\cdot)$, where $d_v, d_t > 0$ and $d_v + d_t = 1$. The MKL framework allows joint learning of the kernel combination weights d_v, d_t and the parameters $\{\alpha_i\}$ and b of the SVM based on the combined kernel. The parameters are found by solving a convex, but non-smooth objective function.[1] Below, we will use f_v, f_t, and f_c to differentiate between classification functions based on the different kernels.

3.2. Semi-supervised classification

Given these different classifiers, we now consider how we can apply them in a semi-supervised setting. We use $\mathcal{L}$ to denote the set of labeled training examples, and $\mathcal{U}$ to refer to the set of unlabeled training examples. As noted above, we assume that our training images have associated tags, but that our final task is to classify images that do not have such tags. We proceed by learning a first classifier on the labeled examples in $\mathcal{L}$, and then use it to predict the class labels for the unlabeled examples in $\mathcal{U}$.

In the case where the first classifier only uses the visual kernel, we do not expect to gain from the unlabeled examples as predicting their label is as hard as it would be for any test image. This is confirmed by our experimental results presented in Section 5. Our experimental results also show that the image tags make many of the classification tasks substantially easier. Therefore, we will use MKL to learn a joint visual-textual classifier from $\mathcal{L}$, and estimate the class labels for the images in $\mathcal{U}$. Assuming that the labels predicted using the MKL classifier f_c are correct, we train a visual-only SVM classifier f_v from all training examples in $\mathcal{L} \cup \mathcal{U}$.

In practice, however, the joint classifier is not perfect, and we consider two alternative approaches to leverage the predictions of the joint classifier on the unlabeled examples in $\mathcal{U}$. In the first alternative, we only add the examples that are confidently classified using the MKL classifier and fall outside the margin, *i.e.* those with $|f_c(x)| \geq 1$, instead of adding all examples in $\mathcal{U}$. This choice is motivated by the observation that these are precisely the examples that would not change the MKL classifier if they were included among the training data for it.

Our second alternative is motivated by the observation that the only information from the MKL classifier that we use when training the final visual classifier is the sign of the examples selected from $\mathcal{U}$. Therefore, the value of $f_v(x_i)$ can arbitrarily differ from $f_c(x_i)$ provided that it is consistent with the class labels of the labeled examples, and the estimated class label of the unlabeled ones. Instead, we will directly approximate the joint classification function f_c learned using MKL. We do so by performing a least squares regression (LSR) on MKL scores $f_c(x)$ for all examples in $x \in \mathcal{L} \cup \mathcal{U}$, to find the function $f_v(x) = \sum_i \alpha_i k_v(x, x_i) + b$ based on the visual kernel. We choose to regularize LSR by projection on a lower-dimensional space using Kernel PCA [20]. We perform singular value decomposition

[1] We used the MKL implementation available at `http://www.kyb.tuebingen.mpg.de/bs/people/pgehler/ikl-webpage/`.

(SVD) to obtain a pseudo-inverse of $K_v = U\Lambda V^\top$, the centered kernel matrix for k_v such that the columns have zero mean. We invert it by suppressing dimensions with singular value in Λ below $\epsilon = 10^{-10}$. Using s to denote the vector of centered classification scores obtained with f_c, we then obtain the α_i parameters in the vector $\alpha = V\bar{\Lambda}U^\top s$, as described in Algorithm 1 below, and b is set to 0.

Algorithm 1: Procedure for learning a semi-supervised MKL+LSR visual classifier.

Input: Labeled data $\mathcal{L}$ and unlabeled data $\mathcal{U}$, visual kernel k_v and textual kernel k_t.

Output: Visual classifier α using kernel k_v.

$f_c \leftarrow \mathrm{MKL}(\mathcal{L}, \{k_v, k_t\})$ */* Learn MKL classifier */*

foreach $x \in \mathcal{L} \cup \mathcal{U}$ **do** */* Center scores */*

  $s(x) \leftarrow f_c(x) - \langle f_c(x') \rangle_{x' \in \mathcal{L} \cup \mathcal{U}}$

end

foreach $x, x' \in \mathcal{L} \cup \mathcal{U}$ **do** */* Center kernel columns */*

  $K_v(x, x') \leftarrow k_v(x, x') - \langle k_v(x, x'') \rangle_{x'' \in \mathcal{L} \cup \mathcal{U}}$

end

$U\Lambda V^\top = K_v$ */* SVD of K_v */*

for $i = 1$ **to** $|\mathcal{L} \cup \mathcal{U}|$ **do** */* Pseudo-invert K_v */*

  $\bar{\Lambda}_{ii} \leftarrow \begin{cases} 0 & \text{if } \Lambda_{ii} < \epsilon \\ \Lambda_{ii}^{-1} & \text{otherwise} \end{cases}$

end

$\alpha \leftarrow V\bar{\Lambda}U^\top s$ */* Least-squares regression of s */*

4. Datasets and feature extraction

In our experiments we use the PASCAL VOC'07 [7] and the MIR Flickr [11] data sets. Both were collected from the Flickr website. Example images are given in Figure 1. For the PASCAL VOC'07 set we used the standard train/test split, and for the MIR Flickr set we randomly split the images into equally sized test and train sets.[2]

The PASCAL VOC'07 data set contains around 10.000 images which were downloaded by querying for images of 20 different object categories in a short period of time. All the images were then annotated for each of the 20 categories. Using the image identifiers we downloaded the user tags for the 9587 images that were still available on Flickr at time of download, and assumed complete absence of tags for the remaining ones. Keeping the tags that appear at least 8 times (a minimum of 4 times in the training and test sets), a vocabulary of 804 tags was used.

The MIR Flickr data contains 25.000 images collected by downloading images from Flickr over a period of 15 months. The collection contains images under the Creative Common license that scored highest according to Flickr's "interestingness" score. These images were annotated for 24 concepts, including object categories but also more general scene elements such as *sky*, *water* or *indoor*. For 14 of the 24 concepts a second, stricter, annotation was made: for each concept a subset of the positive images was selected where the concept is salient in the image. We refer to these more scrictly annotated classes by using $*$ as a suffix. In total we therefore have 38 categories for this data set. For the MIR Flickr data set we kept the tags that appear at least 50 times (*i.e.* among at least 0.2% of the images), resulting in a vocabulary of 457 tags.

We use a binary vector $t_i \in \{0, 1\}^W$ to encode the absence or presence of each of the W different tags in a fixed vocabulary in a linear kernel $k_t(t_i, t_j) = t_i^\top t_j$ which counts the number of tags shared between two images.

For each image we extracted several different visual descriptors. We then average the distances between images based on these different descriptors, and use it to compute an RBF kernel.[3] Thus, our visual kernel is defined as

$$k_v(x_i, x_j) = \exp(-\lambda^{-1} d(x_i, x_j)), \tag{2}$$

where the scale factor λ is set to the average pairwise distance, $\lambda = N^{-2} \sum_{i,j=1}^{N} d(x_i, x_j)$, and $d(x_i, x_j) = \sum_{m=1}^{M} \lambda_m^{-1} d_m(x_i, x_j)$, where $\lambda_m = \max_{i,j} d_m(x_i, x_j)$.

As in [10], we use local SIFT features [14], and local hue histograms [22], both were computed on a dense multiscale grid and on regions found with a Harris interest-point detector. We quantize the local descriptors using k-means, and represent the image using a visual word histogram. We also compute global color histograms over RGB, HSV, and LAB color spaces.

Following [12], these histogram image representations were also computed over a 3×1 horizontal decomposition of the image, and concatenated to form a new representation that also encodes some of the spatial layout of the image. Furthermore we use the GIST descriptor [16], which roughly encodes the image layout. In total we thus combine $M = 15$ different image representations, using L1 distance for the color histograms, L2 for GIST, and χ^2 for the visual word histograms.

5. Experimental results

In our experiments we measure performance using the average precision (AP) criterion for each class, and also using the mean AP (mAP) over all classes.

5.1. Supervised classification

Our first set of experimental results, presented in Table 1, compares the classification performance using the

[2] The test/train division for the MIR Flickr set and our visual and textual features described hereafter are publicly available at: `http://lear.inrialpes.fr/data/`.

[3] Although orthogonal to the focus of this paper, we could also use MKL to learn a combination of separate visual kernels for each feature set.

PASCAL VOC'07	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable
Image	0.727	0.530	0.491	0.668	0.256	0.524	0.699	0.500	0.460	0.364	**0.433**
Tags	0.667	0.407	0.608	0.375	0.197	0.292	0.513	0.664	0.153	0.393	0.076
Image+Tags	**0.879**	**0.655**	**0.763**	**0.756**	**0.315**	**0.713**	**0.775**	**0.792**	**0.462**	**0.627**	0.414

	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor	Mean
Image	0.439	0.747	0.595	0.834	0.390	0.395	0.399	0.743	0.428	0.531
Tags	0.570	0.676	0.539	0.635	0.248	0.457	0.191	0.712	0.278	0.433
Image+Tags	**0.746**	**0.846**	**0.762**	**0.846**	**0.480**	**0.677**	**0.443**	**0.861**	**0.527**	**0.667**

MIR Flickr	animals	baby	baby∗	bird	bird∗	car	car∗	clouds	clouds∗	dog	dog∗	female	female∗
Image	0.487	0.170	0.214	0.227	0.293	0.375	0.522	0.825	**0.755**	0.323	0.367	0.575	0.549
Tags	0.548	0.235	0.315	0.381	0.458	0.246	0.213	0.499	0.378	0.578	0.572	0.488	0.422
Image+Tags	**0.646**	**0.357**	**0.448**	**0.520**	**0.631**	**0.451**	**0.619**	**0.827**	0.753	**0.681**	**0.728**	**0.617**	**0.601**

	flower	flower∗	food	indoor	lake	male	male∗	night	night∗	people	people∗	plant life	portrait
Image	0.536	0.643	0.501	0.745	0.313	0.517	0.450	0.649	0.558	0.789	0.751	0.785	0.681
Tags	0.494	0.546	0.367	0.603	0.231	0.441	0.339	0.416	0.271	0.722	0.635	0.617	0.455
Image+Tags	**0.653**	**0.742**	**0.606**	**0.770**	**0.341**	**0.561**	**0.496**	**0.686**	**0.596**	**0.835**	**0.795**	**0.809**	**0.711**

	portrait∗	river	river∗	sea	sea∗	sky	structures	sunset	transport	tree	tree∗	water	Mean
Image	0.682	0.265	0.081	0.571	0.334	0.866	0.774	0.665	0.464	0.671	0.548	0.622	0.530
Tags	0.451	0.255	0.035	0.400	0.132	0.670	0.694	0.407	0.365	0.413	0.266	0.539	0.424
Image+Tags	**0.711**	**0.412**	**0.202**	**0.649**	**0.362**	**0.876**	**0.803**	**0.666**	**0.540**	**0.684**	**0.564**	**0.717**	**0.623**

Table 1. The AP scores for the supervised setting on both data sets, with the visual kernel alone (Image), a linear SVM on tags (Tags), and the combined kernel (Image+Tags) obtained by Multiple Kernel Learning. The best classification results for each class are marked in bold.

visual representation and the tags, and their combination with MKL. We observe for both data sets that for many classes the visual classifier is stronger than the textual one, yielding a 10% higher mAP score. Also on both data sets, the combined MKL classifier is significantly improving the classification results, the mAP score increases by more than 13% on the VOC classes and by more than 9% on the MIR classes. Interestingly, the mAP of 0.667 obtained by combining visual features and tags is also significantly above the 0.594 winning score of the VOC'07 which used a visual classifier alone.

These results are in line with those of [13], where visual features and tags were combined for landmark classification. A difference is that we find the visual features to be stronger on average, where the situation was reversed in [13]. This might be due to the fact that they used a weaker linear classifier on the visual features, or due to the different type of classification problems: landmarks might be more likely to be tagged than classes such as *diningtable*. Wang *et al*. [23] also found textual features to improve the performance of visual classifiers, but only for relatively weak visual classifiers and not for strong non-linear classifiers.

5.2. Semi-supervised classification

In this section we present results for semi-supervised learning. We compare the following methods:

- SVM: visual classifier learned on labeled examples,
- MKL+SVM(0): MKL classifier learned on the labeled examples, followed by a visual SVM trained on all training examples using the MKL label prediction,
- MKL+SVM(1): same as MKL+SVM(0) but excluding the unlabeled examples in the margin of the MKL classifier to train the SVM,
- MKL+LSR: uses least-squares regression on the MKL scores for all examples to obtain the visual classifier,
- SVM+SVM(0): same as MKL+SVM(0) but using the visual SVM to predict the class of unlabeled examples.
- Co-training: iterative learning of textual and visual classifiers using the co-training paradigm.

The regularization parameters of the SVM and MKL algorithms can be set using cross-validation, but for the sake of efficiency we adopted the constant value of $C = 10$ for all experiments after observing that this value was selected for many classes and settings in initial experiments. We do not expect major differences when performing cross-validation per class and experiment.

The co-training approach has a number of additional parameters to set: the number of iterations T in which examples are added, and the number of positive and negative examples to add in each iteration, which we denote as p and n respectively. Setting these parameters using cross-validation is relatively costly as each co-training iteration requires re-training of the visual and textual SVM classifiers. For two classes of the VOC'07 set we evaluated the performance over the first 200 iterations using $p = 1, n = 1$ and $p = 1, n = 3$, the latter reflecting the fact that for each class there are many more negative than positive examples.

From the results shown in Figure 3, we observe that using many iterations seems to have a detrimental effect on performance. This might be explained by the small number

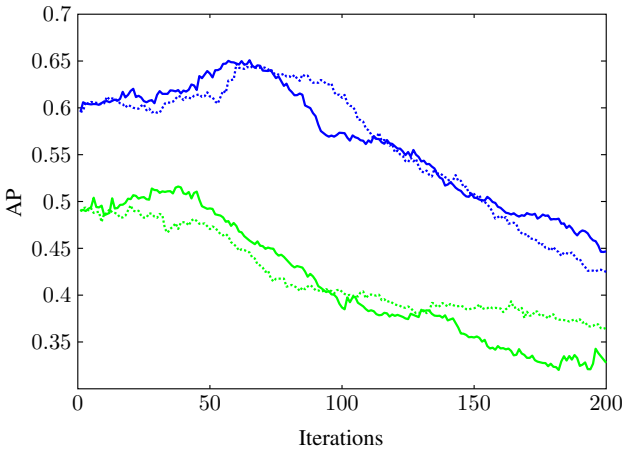

Figure 3. AP scores for the classes *aeroplane* (blue) and *boat* (green), using co-training with $p = 1, n = 3$ (solid) and $p = 1, n = 1$ (dashed), with varying number of co-training iterations.

of positive examples in the unlabeled set. Given these results we used $p=1, n=3$ and compared $T=30$ and $T=50$ for the VOC'07 data set. Since only little difference was observed between the two options in terms of performance, we later opted for $T = 30$ for the MIR Flickr data set in order to reduce the computational load of the experiment.

We evaluated the performance for different amounts of labeled training images. In one set of experiments we randomly selected $k \in \{20, 50, 100\}$ positive and the same number of negative examples for each class. In another set of experiments we use a fraction $r \in \{10\%, 25\%, 50\%\}$ of the positive and negative examples from each class, *i.e*. with $r = 10\%$ and for a class with 2.500 positive images and 10.000 negative ones we randomly select 250 positive examples and 1.000 negative examples. Note that using 10% of the labeled images means that we use a total of 500 and 1.250 labeled images for the VOC and MIR sets respectively, that is, many more than in the $k = 100$ setting.

In Table 2 we report the mAP scores for both data sets for the different learning algorithms with varying amounts of labeled data. Due to lack of space we report the individual AP of the 58 classes only when using 50 labeled training examples per class, see Table 3.

We observe that overall semi-supervised learning significantly improves the performance of the baseline visual-only SVM, in particular when little labeled training data is available. However, it does so only when using the textual features; the visual-only SVM+SVM(0) approach performs worse than the baseline on average and consistently for almost all classes and amount of labeled data. In cases with up to 100 positive and negative examples, MKL+SVM(0) seems to generalize better than MKL+SVM(1), and the MKL+LSR method clearly outperforms all other semi-supervised approaches, including co-training. As larger sets of labeled examples are available, all the methods except

PASCAL VOC'07	20	50	100	10%	25%	50%
SVM	0.268	0.294	0.370	0.345	0.427	0.468
MKL+SVM(0)	0.284	0.314	0.352	0.410	**0.458**	**0.482**
MKL+SVM(1)	0.278	0.322	0.371	0.367	0.440	0.478
SVM+SVM(0)	0.244	0.266	0.328	0.303	0.395	0.455
MKL+LSR	**0.336**	**0.366**	**0.406**	**0.413**	**0.458**	**0.482**
Co-training(30)	0.287	0.323	0.381	0.360	0.438	0.475
Co-training(50)	0.285	0.328	0.377	0.374	0.441	0.476

MIR Flickr	20	50	100	10%	25%	50%
SVM	0.276	0.333	0.370	0.412	0.462	0.501
MKL+SVM(0)	0.272	0.334	0.365	**0.441**	**0.479**	0.505
MKL+SVM(1)	0.283	0.340	0.373	0.424	0.471	0.504
SVM+SVM(0)	0.267	0.319	0.358	0.392	0.444	0.490
MKL+LSR	**0.316**	**0.367**	**0.395**	0.431	0.475	**0.510**
Co-training(30)	0.286	0.351	0.380	0.420	0.471	0.504

Table 2. Performance in mAP on the two data sets for different learning methods and various amounts of labeled training images.

SVM+SVM(0) tend to perform similarly. From the per-class results in Table 3, we observe that the gain varies strongly across classes. For four out of the 38 MIR Flickr classes the baseline supervised classifier performs best: *male*$*$, *river*$*$, *tree* and *tree*$*$. However, this is largely compensated for by the improvements on the 34 other classes obtained by our MKL+LSR method.

5.3. Learning classes from Flickr tags

In our third set of experiments we consider learning classifiers without using any manually labeled examples. For this purpose we use the 18 classes of the MIR Flickr set for which the class name also belongs to the tag dictionary. For the training images we exclude the class name from the textual representation to avoid learning a degenerate classifier that uses the tag to perfectly predict itself. As before, performance is measured using AP based on the manual ground truth class labels on the test set. Our baseline approach takes all images tagged with the class name as positives, and all other images as negatives.

The tags have a noisy relation to the class labels since the tags are not always relevant to the image content, and most images have only a few tags and lack many relevant ones. The positive examples from tag annotation are relatively clean (82.0% precision averaged over all 18 classes), but a large portion of the true positive images is not tagged (17.8% recall on average).

As in the semi-supervised setting, we first learn a joint visual-textual MKL classifier, albeit from all 12.500 images in this case, and then use it to learn a visual only classifier. In this setting we use our semi-supervised approach to remove examples that are likely to be incorrectly tagged, rather than to add unlabeled examples. Given that the positive examples have a relatively low label noise, and that we have many more negative examples than positives, we will remove only the negative examples with the highest scores

PASCAL VOC'07	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable
SVM	0.387	0.218	0.217	0.462	0.150	0.213	0.439	0.271	0.265	0.112	0.258
MKL+SVM(0)	0.549	0.163	0.271	0.409	**0.169**	0.253	0.453	0.311	**0.310**	0.127	0.261
MKL+SVM(1)	0.479	0.218	0.248	0.466	0.145	0.233	0.467	0.296	0.297	**0.186**	0.247
MKL+LSR	**0.592**	**0.324**	**0.376**	**0.519**	0.154	0.278	**0.501**	**0.366**	0.300	0.117	0.255
SVM+SVM(0)	0.326	0.185	0.201	0.398	0.142	0.205	0.444	0.233	0.299	0.108	0.233
Co-training (30)	0.475	0.199	0.299	0.400	0.158	**0.326**	0.497	0.306	0.209	0.148	**0.299**

	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor	Mean
SVM	0.318	0.347	0.321	0.651	0.199	0.182	0.175	0.451	0.239	0.294
MKL+SVM(0)	0.280	0.452	0.251	0.685	0.181	0.213	0.183	0.550	0.219	0.314
MKL+SVM(1)	0.306	0.464	0.326	0.652	0.209	**0.239**	**0.193**	0.522	0.238	0.322
MKL+LSR	**0.331**	**0.637**	**0.383**	**0.703**	**0.212**	0.218	0.191	**0.617**	0.236	**0.366**
SVM+SVM(0)	0.310	0.215	0.249	0.647	0.143	0.218	0.164	0.403	0.197	0.266
Co-training (30)	0.289	0.517	0.362	0.662	0.148	0.233	0.170	0.517	**0.249**	0.323

MIR Flickr	animals	baby	baby∗	bird	bird∗	car	car∗	clouds	clouds∗	dog	dog∗	female	female∗
SVM	0.299	0.043	0.162	0.057	0.094	0.204	0.246	0.569	0.481	0.155	0.181	0.431	0.319
MKL+SVM(0)	0.278	0.055	0.151	**0.141**	0.065	0.210	0.228	0.573	0.503	0.124	0.170	0.436	0.321
MKL+SVM(1)	0.300	0.037	0.159	0.085	0.077	0.220	0.242	0.597	0.508	0.160	0.176	0.425	0.324
MKL+LSR	0.310	**0.075**	**0.161**	0.124	**0.163**	**0.229**	**0.305**	**0.612**	**0.537**	0.182	**0.212**	0.440	0.313
SVM+SVM(0)	0.266	0.038	0.146	0.054	0.073	0.196	0.224	0.560	0.446	0.151	0.169	0.432	0.312
Co-training (30)	**0.345**	0.035	0.136	0.076	0.097	0.199	0.287	0.597	0.471	**0.187**	0.194	**0.443**	**0.357**

	flower	flower∗	food	indoor	lake	male	male∗	night	night∗	people	people∗	plant life	portrait
SVM	0.264	0.359	0.295	0.518	0.139	0.358	**0.296**	0.471	0.289	0.588	0.529	0.602	0.443
MKL+SVM(0)	0.278	0.360	0.267	0.522	0.137	0.319	0.295	0.439	0.259	0.612	**0.553**	0.600	0.441
MKL+SVM(1)	0.353	0.387	0.297	0.516	0.132	0.312	0.281	**0.482**	0.285	0.615	0.545	0.617	**0.477**
MKL+LSR	**0.373**	**0.424**	**0.333**	0.514	0.159	0.366	0.255	0.471	**0.368**	0.629	**0.554**	0.613	0.474
SVM+SVM(0)	0.197	0.343	0.289	0.519	0.122	0.358	0.267	0.460	0.222	0.586	0.528	0.602	0.441
Co-training (30)	0.359	0.419	0.282	**0.559**	**0.172**	**0.380**	0.249	0.466	0.289	**0.634**	0.544	**0.634**	0.465

	portrait∗	river	river∗	sea	sea∗	sky	structures	sunset	transport	tree	tree∗	water	Mean
SVM	0.404	0.154	**0.054**	0.361	0.166	0.661	0.614	0.470	0.285	**0.461**	**0.254**	0.378	0.333
MKL+SVM(0)	0.413	0.150	0.047	0.410	0.209	0.670	0.649	0.503	0.278	0.458	0.155	0.413	0.334
MKL+SVM(1)	0.421	0.149	0.041	0.423	0.158	0.673	0.643	0.515	0.279	0.439	0.178	0.406	0.340
MKL+LSR	0.429	**0.234**	0.047	**0.437**	**0.255**	0.693	**0.655**	**0.543**	**0.321**	0.453	0.231	**0.452**	**0.367**
SVM+SVM(0)	0.391	0.135	0.043	0.357	0.147	0.655	0.615	0.448	0.275	0.454	0.230	0.374	0.319
Co-training (30)	**0.432**	0.181	0.050	0.417	0.213	**0.705**	0.636	0.493	0.273	0.443	0.209	0.426	0.351

Table 3. AP scores for the 38 classes of the two data sets using 50 positive and 50 negative labeled examples for each class.

according to the MKL classifier. We experimented with removing between 2.000 and 10.000 negative examples from the total 12.500 training examples.

In Table 4 we show the performance of the baseline visual-only SVM and of the MKL+LSR approach for various numbers of negative examples that were removed. Not surprisingly, when learning from the user tags, AP scores are lower than those obtained using manual annotations for training, *c.f.* results for "Image" in Table 1. However also in this more difficult scenario our semi-supervised approach improves on average over the performance of the baseline that directly learns a visual classifier from the noisy labels.

As before, the results vary strongly among the classes: for 5 classes the baseline is better (up to 5.6% on *baby*), while for 13 classes our MKL+LSR approach improves results (up to 9.8% on *night*). On average, the improvement is 2.2%. On the same subset of 18 classes, the supervised approach has a mAP of 53.0% compared to 40.7% for MKL+LSR, demonstrating the significant gain obtained by adding supervised information.

6. Conclusion and Discussion

We have considered how learning image classifiers can benefit from unlabeled examples in the case where the training images have associated tags. We presented a novel semi-supervised approach that operates in two stages. First, we learn a strong classifier from the labeled examples that uses both visual features and tags as inputs. The first classifier is then evaluated on both the labeled and unlabeled training examples. In the second stage we learn a visual-only classifier by fitting a function on the scores of the strong classifier, or re-training a classifier.

Our experiments compared several variants of this semi-supervised approach with a co-training approach. From the results we conclude the following: (i) The tags provide a useful feature that improves classification performance for most classes when combined with visual features. (ii) Classifiers learned from limited amounts of labeled training can be improved by using unlabeled training images, but only when additional information in the form of tags is available.

	Removed	animals	baby	bird	car	clouds	dog	flower	food	lake	night
SVM	0	0.304	**0.133**	**0.180**	0.288	0.621	0.249	0.438	0.402	**0.256**	0.465
MKL+LSR	0	0.279	0.082	0.167	0.298	0.628	0.237	0.437	0.405	0.237	0.485
MKL+LSR	2000	0.279	0.073	0.173	0.304	0.662	0.255	0.464	**0.429**	0.207	0.525
MKL+LSR	4000	0.285	0.078	0.128	**0.307**	0.679	**0.258**	**0.468**	0.427	0.254	0.544
MKL+LSR	8000	0.299	0.077	0.129	0.305	0.695	0.256	0.462	0.419	0.216	0.563
MKL+LSR	10000	**0.313**	0.076	0.114	0.293	**0.698**	0.250	0.454	0.414	0.208	**0.565**

	Removed	people	portrait	river	sea	sky	sunset	tree	water	Mean
SVM	0	0.556	0.440	**0.216**	0.353	0.656	0.600	0.368	0.403	0.385
MKL+LSR	0	0.578	0.450	0.214	0.336	0.650	0.593	0.370	0.402	0.380
MKL+LSR	2000	0.582	0.480	0.164	0.362	0.665	0.615	0.372	0.430	0.391
MKL+LSR	4000	0.589	0.503	0.182	0.380	0.676	0.613	0.388	0.445	0.400
MKL+LSR	8000	0.606	**0.517**	0.181	0.418	0.695	**0.614**	0.413	**0.463**	**0.407**
MKL+LSR	10000	**0.616**	**0.517**	0.178	**0.432**	**0.708**	0.604	**0.428**	0.461	**0.407**

Table 4. AP scores for 18 of the MIR Flickr classes when learning from image tags using a visual-only SVM approach and our MKL+LSR approach that also uses the image tags. For the latter, we removed varying amounts of negative examples to obtain the visual-only classifier.

(iii) Our semi-supervised method that uses regression to learn the second visual-only classifier outperforms the other approaches we considered. (iv) When learning from noisy image tags rather than manual labeling we can improve the performance by using our multimodal semi-supervised approach to remove noisy negative examples.

In parallel, we also considered learning the textual-visual classifier and the visual-only classifier *jointly*, rather than *sequentially* as presented in this paper. However, it appeared unclear how to make the combined classifier benefit from the visual classifier.

In future work, we want to explore more powerful text representations than the current linear kernel over binary tag absence/presence vectors. In addition, we will consider automatically adding unlabeled training data from Flickr. Using these in combination with the existing labeled data we hope to improve state-of-the-art performance on these benchmarks without additional manual labeling.

References

[1] S. Baluja. Probabilistic modeling for face orientation discrimination: Learning from labeled and unlabeled data. In *NIPS*, 1998.

[2] T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y. Teh, E. Learned-Miller, and D. Forsyth. Names and faces in the news. In *CVPR*, 2004.

[3] T. Berg and D. Forsyth. Animals on the web. In *CVPR*, 2006.

[4] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, 1998.

[5] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, 2006.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[7] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007). http://www.pascal-network.org/challenges/voc/voc2007/workshop/index.html.

[8] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for Google images. In *ECCV*, 2004.

[9] R. Fergus, Y. Weiss, and A. Torralba. Semi-supervised learning in gigantic image collections. In *NIPS*, 2009.

[10] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, 2009.

[11] M. Huiskes and M. Lew. The MIR Flickr retrieval evaluation. In *ACM MIR*, 2008.

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[13] Y. Li, D. Crandall, and D. Huttenlocher. Landmark classication in large-scale image collections. In *ICCV*, 2009.

[14] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[15] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.

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

[17] A. Quattoni, M. Collins, and T. Darrell. Learning visual representations using images with captions. In *CVPR*, 2007.

[18] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *JMLR*, 9:2491–2521, 2008.

[19] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In *ICCV*, 2007.

[20] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge Univ Pr, 2004.

[21] R. Tibshirani and G. Hinton. Coaching variables for regression and classification. *Statistics and Computing*, 8:25–33, 1998.

[22] J. van de Weijer and C. Schmid. Coloring local feature extraction. In *ECCV*, 2006.

[23] G. Wang, D. Hoiem, and D. Forsyth. Building text features for object image classification. In *CVPR*, 2009.

[24] G. Wang, D. Hoiem, and D. Forsyth. Learning image similarity from Flickr groups using stochastic intersection kernel machines. In *ICCV*, 2009.