MODELING LABEL DEPENDENCIES IN KERNEL LEARNING FOR IMAGE ANNOTATION

Phong D. Vo
Telecom ParisTech
46 rue Barrault 75013 Paris

Hichem Sahbi
CNRS Telecom ParisTech
46 rue Barrault 75013 Paris

ABSTRACT

We introduce in this paper a novel image annotation approach based on maximum margin classification and a new class of kernels. The method goes beyond the naive use of existing kernels and their restricted combinations in order to design “model-free” transductive kernels applicable to interconnected image databases. In a first contribution of the method, we learn both a decision criterion and a kernel map that guarantee linear separability in a high dimensional space and good generalization performance.

In the second contribution of this work, we extend this class of kernels in order to include label dependency statistics that model contextual relationships between concepts into images. Experiments conducted on MSRC and Corel5k databases show that our method achieves at least comparable results with related state of the art.

Index Terms—kernel design, transduction, explicit mapping

1. INTRODUCTION

With the rapid growth of multimedia sharing spaces, such as social networks, visual contents are nowadays abundant. Searching these large collections requires a preliminary step of image annotation that translates visual contents into labels also known as keywords or concepts (see for instance [1]). The task is challenging due to the perplexity when assigning many possible labels to images and the difficulty to analyze rich and highly semantic contents.

One of the most successful image annotation approaches are based on machine learning and may be categorized into generative and discriminative. Generative methods model a prior knowledge and dependencies between image observations and their possible labels using for instance graphical models [2, 3, 4, 5]. This category of methods even though relatively successful suffers from complexity in modeling and inference especially when labels are taken from a large vocabulary. Alternative approaches are discriminative and consider image annotation as a classification problem [6, 7]. A vocabulary of labels is first defined, and a decision criterion is then learned for each label and used in order to identify images belonging to that label. More recently, ranking methods [8, 9, 10], as a branch of discriminative approaches, are used in a supervised way in order to learn distance functions from training data.

The aforementioned categories of methods may fail when concepts in images are highly semantic and difficult to model. In order to overcome these issues, recent discriminative approaches consider a prior knowledge and relationships between data and learned concepts (context, shared features, etc.) [5, 11, 12, 13]. The success of these image annotation methods also depends on cardinality of the labeled data and the choice of the appropriate setting for learning. In these methods, labeled data are usually scarce and expensive; only a very small fraction of training images is labeled and the unlabeled images may not follow the same distribution as the labeled ones, so learning using inductive techniques is clearly not appropriate. Alternatives [14, 15] may include the unlabeled data as a part of the learning process and this is known as transductive inference. The concept of transductive inference, or transduction, was pioneered by Vapnik (see for instance [15]). When applied, these transductive methods turned out to be very useful in order to overcome the limited cardinality of the labeled images in image annotation [16].

Among popular learning techniques support vector machines (SVM) [17] are well studied and proved to be performant in image annotation [18]; in SVMs, kernels are used in order to model visual similarity between images, and only images sharing the same concepts are expected to have high kernel values. The success of SVMs is therefore, highly dependent on the choice of kernels and usual ones, such the linear, the Gaussian and the histogram intersection, is therefore, highly dependent on the choice of kernels and usual ones, such the linear, the Gaussian and the histogram intersection, may not be appropriate in order to capture the actual and the semantic similarity between images for some specific concepts. Better kernels based on tuning Mahalanobis distances were obtained by minimizing the ratio between intra and inter class distances [19, 20]. In order to take extra advantage from different settings, multiple kernels (MKL) were introduced (for instance [21, 22]) and consider convex linear combinations of elementary kernels and proved to be more suitable [23]. However, MKL based design hits at least two major limitations: i) these methods are limited by the cardinality of labeled data, and ii) they are mainly restricted to linear combinations of off-the-shelf kernels.

In [24] authors proposed the transductive kernel learning (TKL) algorithm which is based on a constrained matrix factorization which produces a kernel map that takes image data from the input space into a high dimensional space in order to guarantee their linear separability while maximizing their margin. In that work, transductive kernel learning is not restricted to only convex linear combinations of existing kernels [22, 25]; indeed it is model-free. Beside maximizing the margin, its transductive approach includes a regularization term that enforces smoothness and low rankness in the resulting kernel map in order to diffuse label information from training to test data. Due to the availability of both the training and test data, the learning outcome obtains better generalization performance.

In this paper, we propose a multi-label transductive kernel learning algorithm for image annotation. The proposed contribution is built upon the work in [24] that initially requires learning multiple kernel maps and multiple binary classifiers in order to handle multiple classes. We propose an efficient and effective method that learns a shared kernel map together with a shared classification criterion that allow us to model dependencies between concepts into images. Besides, the proposed method handles imbalanced class distributions resulting from the natural distribution of concepts into images that may lead to biased decision criteria\(^\dagger\) (see also [5, 13, 26]).

\(^\dagger\)According to [26], re-balancing techniques are often limited to data re-sampling and asymmetric-cost learning algorithms. Such techniques do not
In this work, we include label co-occurrences (learned from a corpus of labeled training data) as a part of our kernel design. The advantage of this modeling scheme is twofold: on the one hand it promotes labels with high co-occurrences during inference, and on the other hand it reduces the effect due to imbalanced class distributions. As corroborated through image annotation experiments, our kernel design performs well on the MSRC and the Corel5k databases.

The remainder of this paper is organized as follows. We introduce our transductive kernel design approach in Section 2 and the implementation of our optimization procedure in Section 3. We illustrate in Section 4.2 the application of our method to image annotation using two datasets: MSRC and Corel5k. We conclude the paper in Section 5 while providing a possible extension for a future work.

2. METHOD

Define $X \subseteq \mathbb{R}^n$ as an input space corresponding to all possible image features and let $S = \{x_1, \ldots, x_n\}$ be a finite subset of $X$ where only the first $\ell$ label vectors of $S$, denoted $\{y_1, \ldots, y_\ell\}$ are given; here $y_i \in \{-1, +1\}^K$ and $K$ is the number of possible labels. In many real-world applications only a few data is labeled (i.e., $\ell \ll m$) and its distribution may be different from the unlabeled one. In what follows, features and labels of input data are represented as matrices $X \in \mathbb{R}^{n \times m}$ and $Y \in \mathbb{R}^{m \times m}$ in which the last $(m - \ell)$ columns of $Y$ are set to zeros; and $X_i$ (or $[X]_{i,:}$) denotes the column $i$ of $X$ while $X_{ij}$ is the entry of $X$ at row $i$ and column $j$. Additionally, $\| \cdot \|_F$ is the Frobenius norm, $\langle \cdot, \cdot \rangle$ is the trace operator, $X'$ is the transpose of $X$, and diag$(v)$ is a diagonal matrix whose diagonal entries are taken from a vector $v$.

2.1. Multi-Label Kernel Map Learning

Conventional max-margin classifiers (e.g., SVM [17]) aim to learn functions of the form $y = \langle w, \phi(x) \rangle + b$ in which $\phi(\cdot)$ maps nonlinear data $\{x_i\} \in X$ into linear ones $\{\Phi_i\} \in H$. In this model, $w \in H$ is the normal vector of the “optimal” hyperplane that separates training data $\{(x_i, y_i)\}_{i=1}^n$ while maximizing their margin $1/\|w\|$ and $H$ is a Reproducing Kernel Hilbert Space equipped with a dot-product $\langle \cdot, \cdot \rangle$ (see for instance [27]). In kernel methods [28, 27], the mapping function $\phi(\cdot)$ is not always explicitly known; it is usually associated with an implicit map whose dimension could be very large (possibly infinite) and this may result into very limited generalization performance [29]. For a better and explicit control of the dimensionality of kernel maps, a transductive kernel learning (TKL) method is introduced in [24]. In the remainder of this paper, we will show an extension of this method to multiple-labels that also models their statistical dependencies.

The goal of TKL is to build a finite-dimensional kernel map $\Phi \in \mathbb{R}^{n \times m}$ and a basis $B \in \mathbb{R}^{m \times p}$, which are the elements of the factorization $B\Phi = X$, together with the parameters $W$ of a decision criterion that guarantees the max-margin property. By sharing $\Phi$ with $K$ classes, we extend the setting in [24] to multiple-labels as

$$\min_{W, \Phi, B} \frac{\alpha}{2} \|B\Phi - X\|^2_F + \frac{\mu}{2} \|\Phi\|^2_F + \frac{1}{2} \|W\|^2_F \quad \text{s.t.} \quad B\Phi = Y,$$

(1)

here $\alpha, \mu \geq 0$ and $C \in \mathbb{R}^{m \times m}$ is a diagonal matrix (with $C_{ii} = 1_{\{1 \leq i \leq \ell\}}$) used in order to guarantee the consistency of obtained labels with those of training data. The Frobenius norms of $\Phi$ and $W$ control the dimensionality of the kernel map and the complexity of classifiers respectively. The $p$ equality constraints generate basis vectors in $B$ with a unit magnitude and prevent $\Phi$ from growing infinitely. Solving the above optimization problem consists in factorizing $X$ into $B\Phi$, while reducing the effective dimension (rank) of $\Phi$ and also separating the labeled training data; initially, the rank of $\Phi$ should be overestimated to $p = \max(n, m) + 1$ in order to guarantee the linear separability of data in $\{(x_i, y_i)\}_{i=1}^n$.

In order to diffuse labels from training to test data, a smoothness term is added to the above formulation. This term enforces similar label predictions to visually similar images. Considering the matrix $A$, with each entry $A_{ij}$ being proportional to the similarity between $x_i$ and $x_j$, we add the following criterion to (1)

$$\frac{1}{2} \sum_{i,j=1}^m \|\Phi_i - \Phi_j\|^2_F A_{ij} = \text{tr}\left(W'\Phi L \Phi' W\right).$$

(2)

The right-hand side of (2) corresponds to the discrete Laplacian operator [30] and $L = D - A$ with $D = \text{diag}(A_{1:n})$ and $I_n$ is the all-ones vector of length $m$. The affinity matrix $A$ characterizes the local structure of data where each data point is connected to its $k$ most similar neighbors (see Experiments).

2.2. Modeling Label Dependency

The idea is based on modeling the dependency between labels in a given image $I$: if a label $c$ is likely to be present into $I$ and $c$ highly correlates with another label $c'$, then $c'$ is also likely to be present in $I$. Let us denote $p(c|c')$ as the probability of the occurrence of $c$ given that of $c'$, minimizing the following term encourages labels $c$ and $c'$ to co-occur:

$$\frac{1}{2} \sum_{c=1}^K \sum_{c'=1}^K \|W'c - W'c'\|^2_2 \quad \text{P}_{cc'} = \text{tr}\left(W'QW'\Phi\right).$$

(3)

In the above equation, $\text{P}_{cc'} = p(c|c')$ and $Q = \text{diag}(\text{P}_{1:K}) - \text{P}$ and $1_K$ is the all-one vector of size $K$. Given a training database, the conditional probability $p(c|c')$ is computed as the ratio between the number of images annotated with both labels $c, c'$ and the number of images annotated with $c'$:

$$p(c|c') = \frac{\sum_{i=1}^n 1\{y_{ci}=1\}1\{y_{c'i}=1\}}{\sum_{i=1}^n \sum_{c''=1}^K 1\{y_{c''i}=1\}1\{y_{c'i}=1\}}$$

(4)

The remainder of this paper is organized as follows. We introduce our transductive kernel design approach in Section 2 and the implementation of our optimization procedure in Section 3. We illustrate in Section 4.2 the application of our method to image annotation using two datasets: MSRC and Corel5k. We conclude the paper in Section 5 while providing a possible extension for a future work.

The right-hand side of (2) corresponds to the discrete Laplacian operator [30] and $L = D - A$ with $D = \text{diag}(A_{1:n})$ and $I_n$ is the all-ones vector of length $m$. The affinity matrix $A$ characterizes the local structure of data where each data point is connected to its $k$ most similar neighbors (see Experiments).

2.2. Modeling Label Dependency

The idea is based on modeling the dependency between labels in a given image $I$: if a label $c$ is likely to be present into $I$ and $c$ highly correlates with another label $c'$, then $c'$ is also likely to be present in $I$. Let us denote $p(c|c')$ as the probability of the occurrence of $c$ given that of $c'$, minimizing the following term encourages labels $c$ and $c'$ to co-occur:

$$\frac{1}{2} \sum_{c=1}^K \sum_{c'=1}^K \|W'c - W'c'\|^2_2 \quad \text{P}_{cc'} = \text{tr}\left(W'QW'\Phi\right).$$

(3)

In the above equation, $\text{P}_{cc'} = p(c|c')$ and $Q = \text{diag}(\text{P}_{1:K}) - \text{P}$ and $1_K$ is the all-one vector of size $K$. Given a training database, the conditional probability $p(c|c')$ is computed as the ratio between the number of images annotated with both labels $c, c'$ and the number of images annotated with $c'$:

$$p(c|c') = \frac{\sum_{i=1}^n 1\{y_{ci}=1\}1\{y_{c'i}=1\}}{\sum_{i=1}^n \sum_{c''=1}^K 1\{y_{c''i}=1\}1\{y_{c'i}=1\}}$$

(4)

In which the training label $1\{y_{ci}=1\}$ if $y_{ci}=1$ and $1\{y_{ci}=1\}=0$ if $y_{ci} \neq 1$. Since we do not consider the case where $c = c'$, then $p(c|c')$ is out of interest and $\text{P}_{cc} = 0$. In practice, we found that better performance is obtained if $P$ is symmetric, i.e., $\text{P}_{cc'} = \frac{1}{2} (p(c'|c) + p(c|c))$.

3. OPTIMIZATION

By replacing the constraint $\|W'\Phi C - Y\|^2_F$, grouping it with the factorization term and plugging (2), (3) into (1), our complete formula is the following

$$\min_{W, \Phi, B} \frac{\alpha}{2} \left(\|W'\Phi C - Y\|^2_F + \|B\Phi - X\|^2_F\right) + \frac{\mu}{2} \|\Phi\|^2_F + \frac{1}{2} \text{tr}\left(W'QW'\Phi\right) + \frac{1}{2} \text{tr}\left(W'(I + \beta FL\Phi') W\right)$$

(5)

s.t. $\|B_i\|^2_F = 1, \forall i = 1, \ldots, p$.
where $\beta$ and $\gamma$ control the effect of smoothness and label dependency in the solution. In order to solve (5), we use an EM-like optimization procedure (see for instance [31]). In what follows, we add the superscript $(t)$ in order to show the evaluation of different parameters w.r.t. different iterations of the optimization process.

**Updating Classifiers and Basis.** Since classifiers in $W$ depend on each other, an iterative optimization procedure is required in order to converge to a stationary solution. Assuming fixed $\Phi^{(t)}$ (denoted simply as $\Phi$) and enforcing the gradient of (5) to vanish (w.r.t $W$) leads to $W^{(t)} = V$ with $V = \lim_{\epsilon \to \epsilon_{\max}} V^{(\epsilon)}$ and

$$V^{(\epsilon)} = (I + \Phi (\alpha C + \beta L + \gamma M) \Phi')^{-1} \left[ \alpha \Phi C Y + \gamma \Phi \Phi' \gamma^{(\epsilon-1)} P \right] c,$$

here $V^{(0)} = W^{(t-1)}$ and $M_{cc} = \sum_{c'} p(c|c')$. Assuming $\Phi^{(t)}$ fixed, we find $B^{(t+1)}$ as

$$\min_B \frac{1}{2} \| X - B \Phi \|_F^2 \quad \text{s.t.} \quad \| B \|_F^2 = 1, i = 1, \ldots, p, \tag{7}$$

which can be solved using Lagrange multiplier method [32].

**Updating Kernel Map.** Considering fixed $B^{(t+1)}$ and $W^{(t+1)}$ (denoted simply as $B$, $W$), and the previous kernel map solution $\Phi^{(t)}$, our goal is to find $\Phi^{(t+1)}$ by solving (5). The optimization problem in (5) admits a unique solution $\Phi^{(t+1)} = \tilde{\Psi}$ where $\tilde{\Psi} = \lim_{\tau \to \tau_{\max}} \Psi^{(\tau)}$ and

$$\tilde{\Psi}^{(\tau)} = \left( \mu I + \alpha B' B + W (\nu I + \gamma Q) W' \right)^{-1} \left[ \alpha (B' X + WYC) + \beta WW' \Psi^{(\tau-1)} A \right], \tag{8}$$

where $\nu_1 = \alpha C_{ii} + \beta D_{ii}$. The process (8) allows us to recursively diffuse the kernel maps from the labeled to the unlabeled data. We iterateously use Eqs (6), (7) and (8) until convergence is reached (i.e., when the difference between current and previous estimates of $W$, $\Phi$, $B$ are below a given threshold $\epsilon$) or when $t$ reaches $t_{\max}$.

### 4. EXPERIMENTS

#### 4.1. Experimental Setup

Our method is tested in multi-label image annotation using two standard datasets MSRC and Corel5k. The former includes 591 images from 23 categories mixing man-made and natural objects; the dataset is randomly partitioned into two equal subsets for training and testing. The latter contains 5000 images manually annotated with 260 labels (each image has at least one label and may include up to 5 labels); the dataset is divided into two parts: 4500 images for training and 500 images for testing.

**Features.** Every image is divided into blocks using three grids of size $1 \times 1, 2 \times 2$, and $1 \times 3$; every block is represented by a bag-of-word histogram based on 512 visual words. The latter result from the quantization of densely sampled SIFT descriptors extracted from training images. We use six variants of SIFT descriptors [33] in order to obtain better visual discrimination: SIFT, rgbSIFT, rgsSIFT, hsvSIFT, cSIFT, and opponentSIFT. Every image is described with a super-descriptor corresponding to histograms of 8 blocks; this super-descriptor is normalized in order to guarantee that its $L_2$-norm is equal to 1.

We define a similarity $A_{ij}$ between any two samples $x_i, x_j$ as $\exp(-\frac{1}{2\sigma^2} \sum_{q=1}^Q d(\psi_q(x_i), \psi_q(x_j))/Z_q)$ where $d(\cdot, \cdot)$ is a distance function and $\psi_q(x_i)$ corresponds to the $q^{th}$ feature vector of $x_i$. In practice, $Q = 6$ and $Z_q = \sum_{i,j} d(\psi_q(x_i), \psi_q(x_j))/m^2$ with $d(\psi_q(x_i), \psi_q(x_j))$ being inversely proportional to the histogram intersection similarity between $\psi_q(x_i)$ and $\psi_q(x_j)$. The bandwidth $\sigma$ is set to $\sum_{i,j \in N(i)} \sum_{q=1}^Q d(\psi_q(x_i), \psi_q(x_j))/m^2 Z_q$, here $N(i)$ corresponds to the indices of the $k$-nearest neighbors of $x_i$. When $j \notin N(i)$, $A_{ij} = 0$. As $A$ is not necessarily symmetric, we consider $(A_{ij} + A_{ji})/2$ instead of $A_{ij}, \forall i, j \in \{1, \ldots, m\}$.

**Evaluation Measures.** Every test image is annotated by five keywords corresponding to the top five scores obtained from the classification results. Then precision (denoted P), recall (denoted R) and the number of keywords having positive recall (denoted N+) are computed for each keyword. We also evaluate performance using the break-even point (BEP) criterion [18]; it measures precision of a given keyword $c$ ($c = 1 \ldots K$) among its top $N_c = \sum_{i=1}^t Y_{ci}$ relevant images. We also compute the mean average precision (denoted mAP) in order to measure ranking quality.

#### 4.2. Results and Discussion

In what follows, TKL and wTKL refer to the optimization problem (5) with $\gamma = 0$ and $\gamma > 0$ respectively. In all these experiments the neighborhood size $k = 3$, the graph Laplacian is normalized $L = I - D^{-1} A$; other settings are $\alpha = \beta = \gamma = 1, \mu = 10^{-8}$, $\epsilon = 10^{-2}, \tau_{\max} = 5, \gamma_{\max} = 20$, and $t_{\max} = 5$.

##### 4.2.1. Does Label-Dependency Help?

Empirical results, shown in Fig. 2(b), demonstrate that modeling the dependencies between classes contributes positively in detecting rare concepts from the Corel5k dataset (see examples in Fig. 3). Indeed, as shown in Table. 1, $N+$ of our method increases from 140 to 165. Besides, recall increases from 35% to 42% when $\gamma$ varies from $10^{-2}$ to 1 with a slight decrease of the precision from 26% to 26% and also BEP (see Fig. 1). However, continuing to increase $\gamma$ enhances neither recall nor mAP. Compared with Corel5k, using label-dependency in MSRC dataset is not so effective (Fig. 2(a)). It turns out that label-dependency is useful if the dataset consists of many classes which are highly imbalanced and correlated with each other. In those cases, the label-dependency term may increase the recall for concepts which are rare or not sufficiently discriminative when using their visual features.

##### 4.2.2. Comparison with Related Work

**Inductive methods.** We consider three state-of-the-art methods: (i) standard SVM classifier [17, 34] with 4 kernel choices (linear, RBF, $\chi^2$, Histogram Intersection); (ii) Multiple Kernel Learning implemented by SMO-MKL [25]; (iii) multi-label SVM implemented by M3L [35]. Parameters of each method are optimally tuned using k-fold cross validation on the training data. Note that SMO-MKL is
Fig. 1. The evolution of the evaluation measures with respect to the label-dependency parameter $\gamma$. The label-dependency term allows our algorithm to retrieve more keywords ($N^+$ is shown in Table 1) which are the main factors that increase the average recall.

![Graph showing precision, recall, mAP, and BEP](image)

Fig. 2. The comparative results between our methods (TKL and wTKL) and the best configurations of the related work.

![Graph comparing precision, recall, and BEP](image)

Our algorithm is extensively trained using 36 Gram matrices taken from the combination of the 6 kernels (linear, $\chi^2$, Histogram Intersection, and RBF with 3 scales values) and all 6 visual descriptors mentioned earlier.

Transductive methods. LapSVM [14] and TranSVM [36] are extensions of SVM for semi-supervised and transduction approaches. While they are based on max-margin approach, LapSVM is more related to our method since both share smoothness regularization term. The two methods are also tested with four choices of kernels mentioned above.

Evaluation results with MSRC (Fig. 2(a)) and Corel5k (2(b)) datasets show that transductive methods are better at recall and mAP while inductive ones are better at precision and BEP. Our method outperforms other transductive approaches (mainly TransSVM in three out of four evaluation criteria) and it is competitive with M3L. Finally, we compare annotation performance of our method against evaluations reported in some related work (see Table 2). In general our method is among the top performant ones. For the state-of-the-art TagProp [8], our method is competitive: it has similar recall, better at $N^+$ and mAP, and slightly worse at precision and BEP.

5. SUMMARY

We introduced in this paper a multi-label transductive kernel learning algorithm and demonstrated its use for image annotation. Our method also includes label-dependency statistics which exploit relationships between concepts into images. Evaluations on the MSRC and Corel5k datasets show that the proposed method is very competitive compared to the state of the art. As a future work, we plan to apply this method to images on social networks by considering auxiliary information such as tags, user profiles, and geographical locations in order to further improve the quality of annotation.

![Table showing comparisons](image)

**Table 2.** Comparisons with related works on Corel5k dataset.