Multi-label Visual Classification with Label Exclusive Context

Xiangyu Chen[†‡§], Xiao-Tong Yuan[§], Qiang Chen[§], Shuicheng Yan[§†], Tat-Seng Chua[†‡]
[†]NUS Graduate School for Integrative Sciences and Engineering
[§]Department of Electrical and Computer Engineering, [‡]School of Computing
National University of Singapore
{chenxiangyu,eleyuanx,chenqiang,eleyans,dcscts}@nus.edu.sg

Abstract

We introduce in this paper a novel approach to multi-label image classification which incorporates a new type of context — label exclusive context — with linear representation and classification. Given a set of exclusive label groups that describe the negative relationship among class labels, our method, namely LELR for Label Exclusive Linear Representation, enforces repulsive assignment of the labels from each group to a query image. The problem can be formulated as an exclusive Lasso (eLasso) model with group overlaps and affine transformation. Since existing eLasso solvers are not directly applicable to solving such an variant of eLasso in our setting, we propose a Nesterov's smoothing approximation algorithm for efficient optimization. Extensive comparing experiments on the challenging real-world visual classification benchmarks demonstrate the effectiveness of incorporating label exclusive context into visual classification.

1. Introduction

Multi-label visual classification aims to solve the problem where each image sample can be assigned with multiple class labels simultaneously. As in a fixed data set, many concepts are semantically related, the class labels may correlate to each other. For instance, as shown in Figure 1(a), the objects {"tree", "grass", "sky"} or {"tower", "sky", "cloud"} are frequently contained in the same image and thus form two groups of co-occurrent labels. It is reasonable to make use of such a correlated context of labels for predicting class labels of the query image sample. In the last decade, co-occurrent label context has been widely exploited in multi-label learning for image annotation. Ueda and Saito [23] proposed a generative model for multi-label learning that explicitly incorporates the pairwise correlation between any two class labels. A Bayesian model is introduced in [11] to assign labels through underlying latent representations. Zhu *et al*. [35] suggested a maximum entropy model for exploring the label correlation for multi-label learning. For large-scale propagation, Chen *et al*. [3] proposed the Kullback-Leibler divergence based multi-label learning, which encodes the label information of an image as a unit label confidence vector, naturally imposes inter-label constraints and manipulates labels interactively.

(a) Co-occurrent labels (b) Exclusive labels

Figure 1. Two types of label context in real-scene images. The label co-occurrent context as in (a) describes the *positive* correlation among labels. The label exclusive context as in (b) describes the *negative* correlation among labels. In this paper, we will novelly incorporate the label exclusive context with linear representation for visual classification.

Different from this body of work motivated from label co-occurrence, we exploit in this paper a complementary type of context, the *label exclusive context*, that describes the negative relationship among class labels. In the setting of multi-label vision problems, when the number of categories is large, we may expect a negative correlations among categories. For example, as shown in Figure 1(b), the objects {"boats", "tiger", "book"} seldom simultaneously appear in a real-scene image. We call such a kind of negatively correlated labels as exclusive labels. The label exclusive context has recently been explored in [6, 4] for the object detection tasks. Here we are particularly interested in the effect of label exclusive context in the setup of multi-label image classification. Given a multi-label query image and several groups of exclusive labels learned from the training images, it is reasonable to expect that the labels in each group should be exclusively assigned to the predicted label vector. This motivates us to develop a visual classification framework with which label exclusive context may be naturally incorporated.

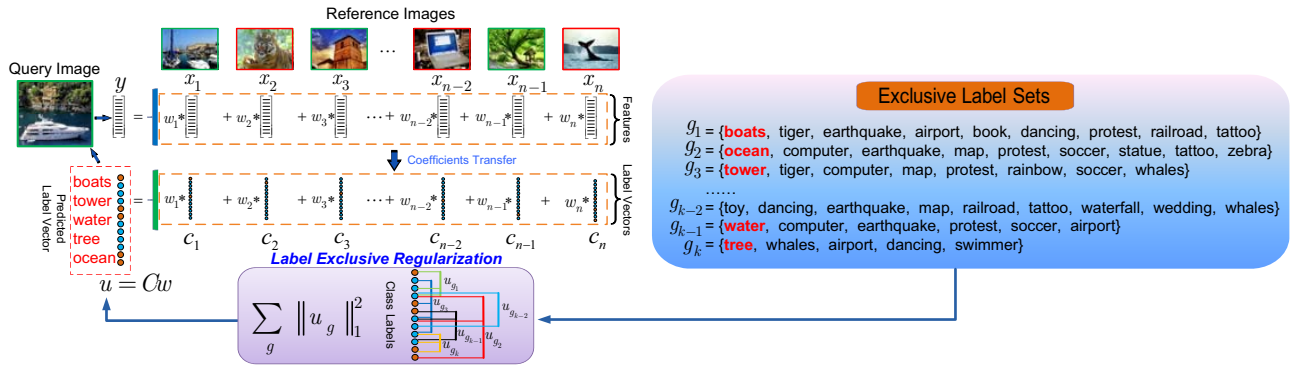

Figure 2. Flowchart of linear representation with exclusive label context. In this system, we have a dictionary of reference images $X = [x_1, ..., x_n]$ with labels $C = [c_1, ..., c_n]$, and a collection of predefined or learned exclusive label sets $\mathcal{G}$. Given a query image y, our method tends to exclusively select labels inside each label set $g \in \mathcal{G}$ to appear in the predicted label vector $u = Cw$ where w (to be learned) best reconstructs the query image, i.e., $y \approx Xw$. This model can be cast as an exclusive Lasso problem with group overlaps and affine transformation. For better viewing, please see original color pdf file.

1.1. Overview of Our Work

The major contribution of this work is a label exclusive context regularized linear representation and classification method. The problem is formulated as an exclusive Lasso [34] model which is a recent advance in sparse learning. Figure 2 depicts the working mechanism of our method. For a given query image feature y, we seek for a linear representation coefficient vector w that best reconstructs y from reference image features X. The predicted label vector u of the query image is the linear combination of the reference image label vectors (zero-one vector indicating multi-label) $C = [c_1, ..., c_n]$ using the same coefficient vector w (to be estimated), i.e., $u = Cw$. Given a set of exclusive label groups $\mathcal{G}$, we expect that at most one label inside each exclusive label set $g \in \mathcal{G}$ will be non-zero in the predicted label vector u. The problem can be cast as an exclusive Lasso with group overlaps and affine transformation. To optimize such an variant of eLasso, we develop a Nesterov-type smoothing approximation [18] method to convert the non-smooth problem to a smooth problem and then solve it using the Accelerated Proximal Gradient method [22]. Moreover, in our application, the exclusive label groups are automatically learned using the dense subgraph searching method [17]. Empirical studies on the challenging visual classification tasks validate the effectiveness of our label exclusive linear representation and classification method.

1.2. Related Work

We briefly review in this subsection several closely related sparse learning techniques utilized in this work.

Sparse Linear Representation for Classification Linear representation with sparse inducing regularizer has enjoyed considerable popularity in recent multi-class visual recognition applications [27, 28, 30, 21, 25]. Given a query image feature and a dictionary of reference features, the objective of sparse linear representation is to select a small set of reference images to reconstruct the query image. Such a sparse representation scheme is typically free of model training and robust to sparse noise. In this work, we show that the label exclusive context can be elegantly integrated into linear representation to boost classification performance.

Group Sparse Inducing Regularization Learning models regularized by group sparse inducing penalties have been widely studied in both machine learning [29, 32] and signal processing fields [13, 10]. Let $w \in \mathbb{R}^n$ be the n parameters to be regularized. Denote $\mathcal{I} = \{1, ..., n\}$ the variable index and $\mathcal{G} = \{g_i \subseteq \mathcal{I}\}_{i=1}^{l}$ a set of variable index groups. The group formation varies according to the given grouping or hierarchical structure. Denote $\|w_{\mathcal{G}}\|_{p,q} := \sum_{g \in \mathcal{G}} \|w_g\|_p^q$ the $\ell_{p,q}$-norm defined over groups $\mathcal{G}$, where $\|w_g\|_p^q := \left(\sum_{j \in g} |w_j|^p\right)^{q/p}$. The $\ell_{2,1}$-norm regularizer is used in group Lasso [29] which encourages the sparsity on group level. Jacob *et al.* [12] proposed the overlap group Lasso and graph Lasso as variants of group Lasso to handle overlapping groups. Another group sparsity inducing regularizer is the $\ell_{\infty,1}$-norm which is widely used in multi-task learning problems[16, 31, 26].

Exclusive Lasso When $p = 1$, $q = 2$, the $\ell_{1,2}$-norm has recently been studied in the exclusive Lasso (eLasso) regression [34] for the multi-task learning. Given a set of observed data $\mathcal{D} = \{X, y\}$ in which $X \in \mathbb{R}^{m \times n}$ is the design matrix of predictors, and $y \in \mathbb{R}^m$ is a response vector. The eLasso is defined (in our notations) as solving the following $\ell_{1,2}$-regularized least squares problem

$$\min_w \frac{1}{2}\|y - Xw\|_2^2 + \frac{\lambda}{2}\sum_{g \in \mathcal{G}} \|w_g\|_1^2, \qquad (1)$$

where λ is a user-specified term trade-off parameter. Unlike

the group Lasso[29] regularizer that assumes covariant variables in groups, the eLasso regularizer models the scenario where variables in the same group are exclusively selected in the output. It is assumed in [34] that the groups in $\mathcal{G}$ are *disjoint*. In our work, motivated by the practice of multi-label visual classification, we will investigate the optimization of an important variant of eLasso with group overlap and affine transformation of parameter vector.

1.3. Paper Organization

The remainder of this paper is organized as follows: We present the label exclusive linear representation and classification framework in Section 2. The optimization procedure is described in Section 3. Section 4 states a kernel-view extension of our method in the setting where features are given in form of kernel matrices. The experimental results on several benchmark visual classification tasks are given in Section 5. We conclude this work in Section 6.

2. Label Exclusive Linear Representation and Classification

The reference image set is represented as a matrix $X = [x_1, ..., x_n] \in \mathbb{R}^{m\times n}$ where m is the feature dimension and n is the sample number. The class labels of the reference images are encoded in a matrix $C = [c_1, ..., c_n] \in \mathbb{R}^{p\times n}$, where p is the number of classes and the elements of label vector c_i are set to be 1 or 0 according to whether image x_i containing the object(s) of the jth class. Here we consider multiple labels, i.e., more than one entries of c_i can be 1.

2.1. Label Exclusive Linear Representation

Given a query image with feature $y \in \mathbb{R}^m$, the label exclusive linear representation (LELR) model is given by

$$\min_w \frac{1}{2}\|y - Xw\|_2^2 + \frac{\lambda}{2}\sum_{g\in\mathcal{G}} \|C_g w\|_1^2, \tag{2}$$

where w is the linear reconstruction coefficient vector, $\mathcal{G}$ is a group of label subsets, each of which contains several exclusive classes (assumed to be known here, and we will address soon in Section 2.2 how to automatically learn $\mathcal{G}$ from the reference set), and C_g is the rows of C indexed in g. The first term measures the linear reconstruction error of feature y by Xw, while the second term utilizes the $\ell_{1,2}$-norm to encourage the label exclusion behavior in the predicted label confidence vector Cw. Since both terms are convex, the objective in (2) is convex. Apparently, LELR model (2) is a variant of the standard eLasso problem (1), with the following notable differences:

- The groups in $\mathcal{G}$ may be overlapping to each other (see Section 2.2).
- The groups are defined over the affine transformed output Cw, rather than on the original parameter vector w.

We will propose shortly in Section 3 an efficient first-order method to optimize the objective in (2). Given the optimal reconstruction coefficient $\hat{w}$, the optimal $\hat{u} = C\hat{w}$ can be regarded as a label confidence vector of the query image. Such a vector can be used for performance evaluation by calculating metrics such as the average precision (AP).

2.2. Learn the Exclusive Label Sets

So far, we assume that the set $\mathcal{G}$ of exclusive label groups used in problem (2) is known as a prior. Actually, it can be automatically learned from the training labels C. Here we use the graph shift method [17] to learn a few groups of exclusive labels as dense subgraphs on a weighted graph $G = \langle V, E\rangle$ defined as follows: the node set $V := \{1, 2, ..., p\}$ contains all the class labels, and the edge set $E \subseteq V \times V$ describes the pairwise exclusiveness between nodes. The weight matrix W associated with G is given by $W_{ij} = 1$ if label i and label j do not simultaneously appear in any training image, and $W_{ij} = 0$ otherwise. The dense subgraphs of G are then determined by the graph-shift method [17]. The nodes in each dense subgraph naturally form, with high confidence, an exclusive label subset. Note that the exclusive groups learned in this way are typically overlapping to each other. Taking NUS-WIDE-LITE dataset [5] as an example, it can be seen in the right part of Figure 2 that the labels "tiger", "airport", "map", "whales", etc., all belong to more than one groups.

3. Optimization

In this section, we investigate the optimization problem associated with the LELR model (2). Since LELR is a variant of eLasso, one may wish to utilize the existing eLasso solvers for optimization. However, it comes to our notice that the eLasso solvers in literature either suffer from slow convergence rate (e.g., subgradient methods in [34]) or are particularly designed for standard eLasso (1) with disjoined groups (e.g., proximal gradient method in [14]), and thus are not directly applicable to LELR. This motivates us to seek for more suitable tools to optimizie the objective in (2). One natural thought is to approximate the non-smooth objective in (2) by a smooth function and then solve the latter by utilizing some off-the-shelf smooth optimization algorithms. Next, we derive a Nesterov's smoothing optimization method to achieve this task.

3.1. Smoothing Approximation

Let us re-express LELR (2) as follows

$$\min_w \{F(w) := f(w) + \lambda h(w)\}, \tag{3}$$

where $f(w) := \frac{1}{2}\|y - Xw\|_2^2$ is a smooth convex term and $h(w) := \frac{1}{2}\sum_{g\in\mathcal{G}} \|C_g w\|_1^2$ is convex but non-smooth. It is standard that $\|C_g w\|_1$ has a max-structure representation

$$\|C_g w\|_1 = \max_{\|u_g\|_\infty \le 1} \langle C_g w, u_g \rangle. \tag{4}$$

By utilizing the Nesterov's smoothing approximation method [18], the $\|C_g w\|_1$ in (4) can be approximated by the following smooth function

$$q_{g,\mu}(w) := \max_{\|u_g\|_\infty \le 1} \langle C_g w, u_g \rangle - \frac{\mu}{2}\|u_g\|_2^2, \tag{5}$$

where μ is a parameter to control the approximation accuracy. For a fixed w, denote $u_g(w) \in \mathbb{R}^{|g|}$ the unique minimizer of (5). It is standard that

$$u_g(w) = \min\left\{1, \max\left\{-1, \frac{C_g w}{\mu}\right\}\right\}. \tag{6}$$

Based on these preliminaries, we now propose to solve the following smooth optimization problem as an approximation to the non-smooth problem (3):

$$\min_w \{F_\mu(w) := f(w) + \lambda h_\mu(w)\}, \tag{7}$$

where h_μ is given by

$$h_\mu(w) := \frac{1}{2}\sum_{g\in\mathcal{G}} q_{g,\mu}^2(w). \tag{8}$$

Assume that $\Omega \in \mathbb{R}^n$ is a bounded feasible set of interest for w, $R := \max_{w\in\Omega} \|w\|_1$, and $\|A\|_p$ denotes the induced p-norm of a matrix A, then we have the following result on approximation accuracy of h_μ:

Proposition 1. *$h_\mu(x)$ is a μ-accurate approximation to $h(x)$, that is*

$$h_\mu(w) \le h(w) \le h_\mu(w) + (m\|C\|_1 R|\mathcal{G}|)\mu. \tag{9}$$

The proof follows the similar arguments as in [18]. Proposition 1 shows that for $\mu > 0$, the function h_μ can be seen as a uniform smooth approximation of function h.

Motivated from [18, Theorem 1], we derive the following result stating that h_μ is differentiable with Lipschitz continuous gradient:

Theorem 1. *Function $h_\mu(w)$ is well defined, convex and continuously differentiable. Moreover, its gradient*

$$\nabla h_\mu(w) = \sum_{g\in\mathcal{G}} q_{g,\mu}(w)(C_g^T u_g(w)) \tag{10}$$

is Lipchitz continuous with the constant

$$L_\mu = \left(m + \frac{\|C\|_1 R}{\mu}\right)\|C\|_2^2|\mathcal{G}|. \tag{11}$$

The proof is given in Appendix A.

3.2. Smooth Minimization via APG

Given a fixed $\mu > 0$, by Theorem 1 it is easy to see that the objective F_μ is differentiable with gradient

$$\nabla F_\mu(w) = X^T(Xw - y) + \lambda \nabla h_\mu(w), \tag{12}$$

which is Lipschitz continuous with constant

$$\tilde{L}_\mu = \|X^T X\|_2 + \lambda L_\mu. \tag{13}$$

Therefore, we employ the Accelerated Proximal Gradient method [22] to optimize the smoothed LELR problem (7). The algorithm is formally described in Algorithm 1. For a fixed μ, it is shown that APG has $\mathcal{O}(1/t^2)$ asymptotical convergence rate bound, where t is the time instance. If we describe convergence in terms of the number of iterations needed to reach an ϵ solution, i.e., $|F_\mu(w) - \min F_\mu| \le \epsilon$, then by choosing $\mu \approx \epsilon$ the rate of convergence is $O(1/\epsilon)$. It is noteworthy that the convergent complexity of Algorithm 1 depends on constant $1/\tilde{L}_\mu$ which is dominated by the factor μ when it is small. To further accelerate Algorithm 1 for extremely small μ, one may resort to the continuation technique as suggested in [1].

Inputs : $X \in \mathbb{R}^{m\times n}$, $y \in \mathbb{R}^m$, C, $\mathcal{G}$, λ, μ.
Output: $w \in \mathbb{R}^n$
Initialization: Calculate $\tilde{L}_\mu$ by (13). Initialize w_0, v_0 and let $\alpha_0 \leftarrow 1$, $t \leftarrow 0$.
repeat
  $u_t = (1-\alpha_t)w_t + \alpha_t v_t$,
  Calculate $\nabla h_\mu(u_t)$ according to (10),
  $v_{t+1} = v_t - \frac{1}{\alpha_t \tilde{L}_\mu}\left(X^T(Xu_t - y) + \lambda\nabla h_\mu(u_t)\right)$,
  $w_{t+1} = (1-\alpha_t)w_t + \alpha_t v_{t+1}$,
  $\alpha_{t+1} = \frac{2}{t+1}$, $t \leftarrow t+1$.
until *Converges* ;

Algorithm 1: Smooth minimization for LELR

4. A Kernel-view Extension

So far, the smooth minimization Algorithm 1 only applies to LELR (2) with raw image features (y, X). However, in the practice of visual classification, the descriptors are often encoded as similarities or kernel matrix, without the raw features available. For the purpose of utilizing feature kernels for LELR, we present in this subsection an extension of LELR to Reproducing Kernel Hilbert Space (RKHS). The intuition of such a kernel trick is to use a nonlinear function ϕ to map the reference and query samples from the original space to a higher dimensional RKHS in which we have $\phi(x_i)^T\phi(x_j) = k(x_i, x_j)$ for certain kernel function $k(\cdot,\cdot)$. In this new space, we can write the problem (2) as:

$$\min_w \frac{1}{2}\|\phi(y) - \phi(X)w\|_2^2 + \frac{\lambda}{2}\sum_{g\in\mathcal{G}}\|C_g w\|_1^2, \tag{14}$$

where $\phi(X) = [\phi(x_1), ..., \phi(x_n)]$. Note that the calculation in APG iteration of Algorithm 1 is characterized by inner product of features, and thus can be straightforwardly extended to solve problem (14). Let $K = \phi(X)^T\phi(X)$ be the reference feature kernel matrix, and $z = \phi(X)^T\phi(y)$ be the query kernel vector. The kernel-view of Algorithm 1 for LELR is given in Algorithm 2.

Inputs : $K \in \mathbb{R}^{n\times n}$, $z \in \mathbb{R}^n$, C, $\mathcal{G}$, λ, μ.
Output: $w \in \mathbb{R}^n$
Initialization: Calculate $\tilde{L}_\mu$ by (13) with X^TX replaced by K. Initialize w_1, v_1 and let $\alpha_0 \leftarrow 1$, $t \leftarrow 0$.
repeat
  $u_t = (1-\alpha_t)w_t + \alpha_t v_t$,
  Calculate $\nabla h_\mu(u_t))$ according to (10),
  $v_{t+1} = v_t - \frac{1}{\alpha_t \tilde{L}_\mu}(Ku_t - z + \lambda\nabla h_\mu(u_t))$,
  $w_{t+1} = (1-\alpha_t)w_t + \alpha_t v_{t+1}$,
  $\alpha_{t+1} = \frac{2}{t+1}$, $t := t+1$,
until *Converges* ;

Algorithm 2: Smooth minimization for LELR in kernel-view

5. Experiments

To evaluate the effectiveness of LELR for object classification, we systematically compare it with representative state-of-the-art methods on several multi-label object classification benchmarks.

5.1. Datasets and Features

The PASCAL VOC 2007&2010 are two challenging databases from the PASCAL Visual Object Classes Challenge (VOC) [9]. A total number of 20 object classes are collected from four main categories, i.e. *Person*, *Animal*, *Vehicle* and *Indoor*. VOC 2007 and VOC 2010 datasets contain 9,963 and 21,738 images respectively. Both datasets are split into 50% for training/validation and 50% for testing. The distributions of images and objects by class are approximately equal across the training/validation and test sets. We utilize the training set as reference image set. We extract several low-level features including SIFT and its variants [19], LBP and HOG by dense sampling strategy in three scales. Each image is represented by Bag-of-Word model with spatial pyramid matching [15]. These features are first transformed to kernel space using χ^2 distance and further combined with a detection kernel as in [2].

The NUS-WIDE-LITE [5] is a lite version of NUS-WIDE database which contains 269,648 images and the associated 5,018 tags. This lite data set consists of 55,615 images randomly selected from the NUS-WIDE data set. For each image, an 81-D label vector is maintained to indicate its relationship to 81 distinct concepts (tightly related to tags yet relatively high-level). For evaluation, we construct a reference image set of size 27,807 whilst the rest are used for testing. We extract multiple types of global visual features which include 225-D block-wise color moments, 128-D wavelet texture and 75-D edge direction histogram. These features are transformed to kernel space using χ^2 distance and linearly combined into a mean feature kernel.

5.2. Evaluation Criteria

Following [33], the criteria to evaluate the performance include *Average Precision* (AP) for each label (or concept) and *Mean Average Precision* (MAP) over all labels. The former is a well-known gauge widely used in the field of image retrieval, whilst the latter is developed to handle the multi-class and multi-label problems. All experiments are conducted on a common PC equipped with 2 Intel quad-core 3.0 GHz CPU and 32GB RAM.

5.3. Results on PASCAL VOC 2007&2010

On VOC 2007, a total number of 11 exclusive label groups are learned, and each group contains 6 labels in average. We compare LELR with two state-of-the-art methods: Locality-constrained Linear Coding (LLC) [24] and Super Vector Coding (SVC) [33], and two reported top ranked solutions [7]: the INRIA_Flat and INRIA_Genetic. Moreover, we are interested in the performance comparison between label exclusive context and label co-occurrence context in linear representation and classification. To do this, we simply replacing the eLasso-type regularizer $\sum_{g\in\mathcal{G}} \|C_g w\|_1^2$ in LELR with a graph Laplacian regularizer that enforces label co-occurrence

$$\min_w \frac{1}{2}\|y - Xw\|^2 + \frac{\lambda}{2} w^T C^T L C w, \qquad (15)$$

where $L = D - W$, W is a label co-occurrence matrix with the entry W_{ij} counting the number of times an object with label i appears in a training image with an object with label j, and D is a diagonal matrix with $D_{ii} = \sum_j W_{ij}$. We call such a model as label co-occurrence linear representation (LCLR). The objective in (15) is quadratic and thus can be optimized with closed form solution.

Table 1 lists the quantitative results. As can be seen that our LELR solution outperforms the competing methods in MAP and APs on 18 out of the 20 object classes. On comparison between LELR and LCLR, since both utilize the same features, the improvement of the former over the latter is supposed to stem from the fact that label exclusive context is more helpful than label con-occurrence context in linear representation and classification. The per query time of LELR is about 0.13 second.

On VOC 2010, a total number of 11 exclusive label groups are learned on the average of 8 labels per group. The comparing results on VOC 2010 are listed in Table 2. In this table, we compare our approach with the

Table 1. The APs and MAPs of different image classification algorithms on the PASCAL VOC 2007 dataset. The **INRIA_F** and **INRIA_G** stand for INRIA_Flat and INRIA_Genetic, respectively.

AP %	INRIA_F	LLC	INRIA_G	SVC	LCLR	LELR
aeroplane	74.8	74.8	77.5	79.4	79.7	**83.7**
bicycle	62.5	65.2	63.6	72.5	76.7	**81.2**
bird	51.2	50.7	56.1	55.6	52.7	**57.8**
boat	69.4	70.9	71.9	73.8	71.2	**75.2**
bottle	29.2	28.7	33.1	34.0	52.0	**53.0**
bus	60.4	68.8	60.6	72.4	73.5	**75.7**
car	76.3	78.5	78.0	83.4	86.0	**90.3**
cat	57.6	61.7	58.8	63.6	62.5	**63.8**
chair	53.1	54.3	53.5	56.6	58.9	**61.4**
cow	41.1	48.6	42.6	52.8	53.8	**54.0**
dining table	54.9	51.8	54.9	**63.2**	54.3	57.2
dog	42.8	44.1	45.8	**49.5**	43.3	42.9
horse	76.5	76.6	77.5	80.9	82.5	**87.4**
motorbike	62.3	66.9	64.0	71.9	73.8	**77.1**
person	84.5	83.5	85.9	85.1	90.1	**92.9**
potted plant	36.3	30.8	36.3	36.4	48.1	**48.7**
sheep	41.3	44.6	44.7	46.5	56.8	**57.6**
sofa	50.1	53.4	50.9	59.8	60.7	**66.2**
train	77.6	78.2	79.2	83.3	78.8	**84.4**
tvmonitor	49.3	53.5	53.2	58.9	68.0	**70.9**
MAP %	57.5	59.3	59.4	64.0	66.2	**69.1**

Table 2. Performance comparison of different image classification algorithms on the PASCAL VOC 2010 dataset. Red score indicates that LELR outperforms NUSPSL_MFDETSVM and blue LELR+SVM outperforms NUSPSL_KERNELREGFUSING.

AP %	NUSPSL_MFD.	LELR	NUSPSL_KER.	LELR+SVM
aeroplane	91.9	**93.3**	93.0	**93.1**
bicycle	77.1	**78.8**	79.0	**79.3**
bird	69.5	**71.0**	71.6	**72.0**
boat	74.7	**76.7**	77.8	**77.9**
bottle	52.5	**52.6**	54.3	54.1
bus	84.3	**85.2**	85.2	**85.5**
car	77.3	**78.5**	78.6	78.6
cat	76.2	**78.1**	78.8	**78.9**
chair	63.0	**64.6**	64.5	**64.9**
cow	**63.5**	62.5	**64.0**	63.7
dining table	62.9	**63.0**	62.7	**63.0**
dog	65.0	**67.8**	69.6	**70.0**
horse	79.5	**81.7**	82.0	**82.2**
motorbike	83.2	**84.9**	84.4	**84.7**
person	91.2	**91.4**	91.6	91.6
potted plant	45.5	**46.9**	48.6	48.6
sheep	65.4	**67.4**	64.9	**71.5**
sofa	55.0	**57.6**	59.6	**60.0**
train	87.0	**88.9**	89.4	89.4
tvmonitor	**77.2**	75.5	76.4	**76.6**
MAP %	72.1	**73.3**	73.8	**74.3**

Winner'10 system from NUS-PSL team [8]: the rank-one algorithm NUSPSL_KERNELREGFUSING and the rank-two algorithms NUSPSL_MFDETSVM. We also fuse the results of LELR and a standard SVM classifier trained on the same kernel to further improve the final performance as used in [2]. As can be seen from Table 2, LELR outperforms NUSPSL_MFDETSVM in MAP and APs on 18 out of 20 classes, and LELR+SVM outperforms NUSPSL_KERNELREGFUSING (VOC 2010 Winner) in MAP and APs on 14 out of 20 classes. Here we do not report the results by LCLR since it is inferior to the state-of-the-art and also for ease of presentation of the table. The per query time of LELR is about 0.2 second.

5.4. Results on NUS-WIDE-LITE

On NUS-WIDE-LITE dataset, a total number of 47 exclusive label groups are learned with averagely 9 labels per group (see the right part of Figure 2 for some exemplar groups). We compare LELR with the following five algorithms: KNN, SVM, LCLR, Entropic Graph Semi-Supervised Classification (EGSSC) [20] and Large-scale Multi-label Propagation (LSMP) [3]. The last two are semi-supervised methods which make use of the feature information of test samples. All the algorithms utilize the same features as described in Section 5.1.

The MAP results obtained under varying reference set sizes (in percentages of the training set) are shown in Figure 3. Figure 4 illustrates the detailed APs for each of the 81 concepts, with the whole training set as reference set. Our observations from Figure 3 and Figure 4 are: (i) under different reference set sizes, LELR consistently outperforms all the baseline algorithms in MAP; and (ii) in Figure 4, LELR and LCLR significantly outperform the other comparing algorithms on some rare concepts (e.g., map, horses, swimmers, waterfall, etc.). This is because LELR and LCLR are a linear representation model which is free of explicit model training and thus is relatively insensitive to the imbalance issue. The LELR per query processing time is about 0.75 second.

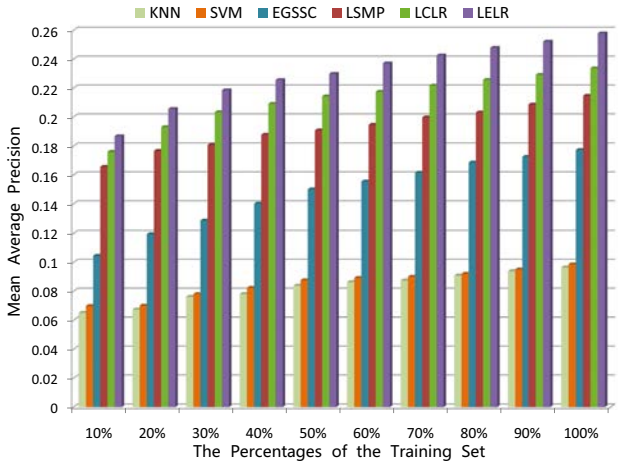

Figure 3. The MAP results of our LELR algorithm and the four baselines with varying reference image set sizes (in percentage) on NUS-WIDE-Lite dataset.

6. Conclusions

The LELR model is proposed to incorporate label exclusive context into a multi-label linear representation framework for visual classification. The problem can be formu-

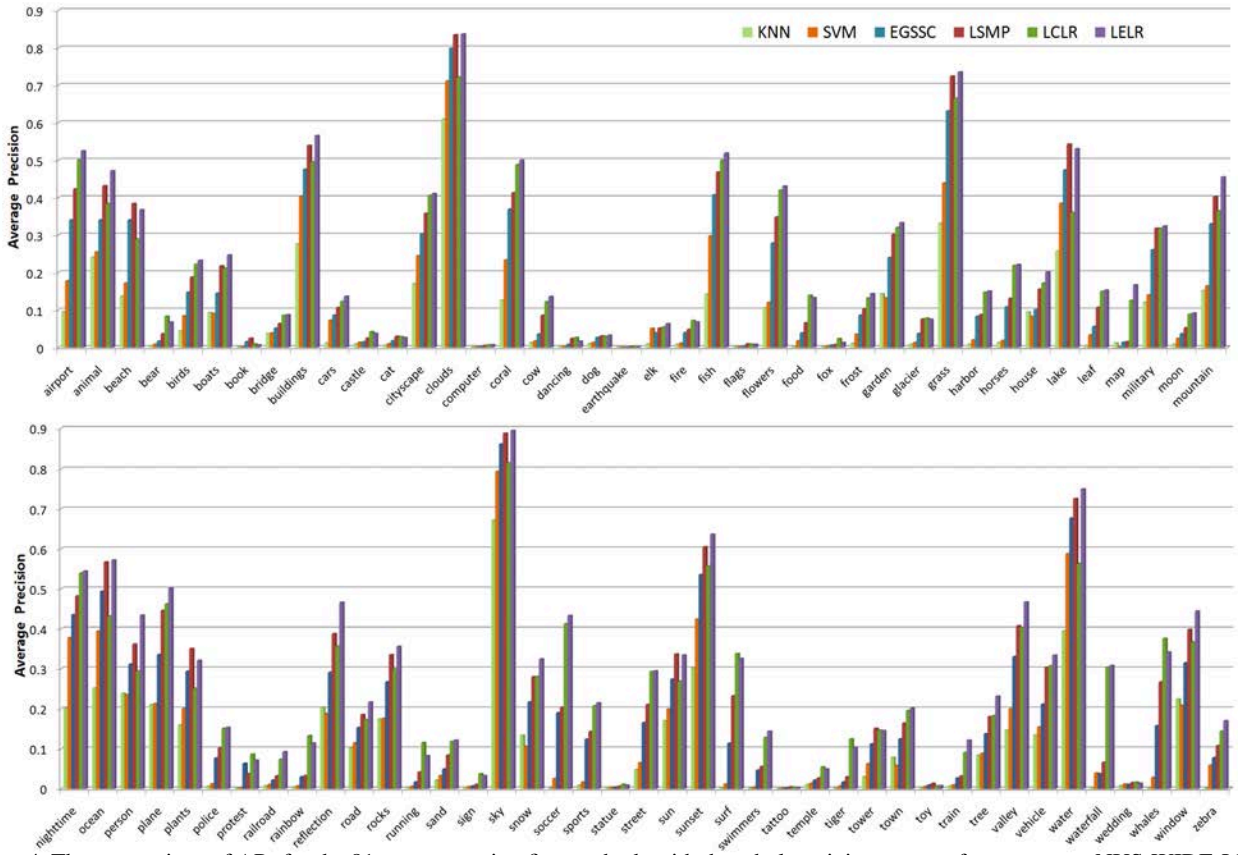

Figure 4. The comparison of APs for the 81 concepts using five methods with the whole training set as reference set on NUS-WIDE-LITE.

lated as an eLasso model with group overlaps and affine transformation. Such a variant of eLasso can be efficiently optimized with Nesterov-type smoothing approximation method. Extensive comparing experiments on the challenging real-world visual classification tasks validate that LELR is a powerful model to boost the performance of linear representation and classification.

7. Acknowledgements

We would like to acknowledge to support of "NExT Research Center" funded by MDA, Singapore, under the research grant: WBS:R-252-300-001-490.

Appendix

A. Proof of Theorem 1

Proof. From the standard results (see, e.g. [18, Theorem 1]) we have that $q_{g,\mu}(w)$ is well defined and continuously differentiable, and its gradient $\nabla q_{g,\mu}(w) = C_g^T u_g(w)$ is Lipschitz continuous with constant

$$L_{g,\mu} = \frac{\|C_g\|_2^2}{\mu} \leq \frac{\|C\|_2^2}{\mu}. \tag{A.1}$$

Since $h_\mu(w)$ is the summation of the *squares* of $q_\mu(w_g)$, it is also well defined with gradient given by

$$\nabla h_\mu(w) = \sum_{g \in \mathcal{G}} q_{g,\mu}(w) \nabla q_{g,\mu}(w). \tag{A.2}$$

To prove the Lipschitz continency of $\nabla h_\mu(w)$, we first show the Lipschitz continuousness of $q_{g,\mu}(w)\nabla q_{g,\mu}(w)$:

$$\begin{aligned} & \|q_{g,\mu}(w_1)\nabla q_{g,\mu}(w_1) - q_{g,\mu}(w_2)\nabla q_{g,\mu}(w_2)\|_2 \\ = \; & \|q_{g,\mu}(w_1)\nabla q_{g,\mu}(w_1) - q_{g,\mu}(w_1)\nabla q_{g,\mu}(w_2) \\ & + q_{g,\mu}(w_1)\nabla q_{g,\mu}(w_2) - q_{g,\mu}(w_2)\nabla q_{g,\mu}(w_2)\|_2 \\ \leq \; & |q_{g,\mu}(w_1)| \cdot \|\nabla q_{g,\mu}(w_1) - \nabla q_{g,\mu}(w_2)\|_2 \\ & + \|\nabla q_{g,\mu}(w_2)\|_2 \cdot |q_{g,\mu}(w_1) - q_{g,\mu}(w_2)| \\ \leq \; & \left(\frac{\|C\|_2^2 \|C\|_1 R}{\mu} + \|C\|_2^2 m \right) \|w_1 - w_2)\|_2, \end{aligned} \tag{A.3}$$

where the last inequality follows the basic facts: (i) (A.1), (ii) $q_{g,\mu}(w_1) \leq \|C_g w_1\|_1 \leq \|C\|_1 R$, (iii)

$\|\nabla q_{g,\mu}(w_2)\|_2 = \|C_g^T u_g(w_2)\|_2 \leq \|C\|_2\sqrt{m}$, and (iv) $|q_{g,\mu}(w_1) - q_{g,\mu}(w_2)| \leq \|C\|_2\sqrt{m}\|w_1 - w_2\|_2$ (due to the boundness of $\nabla q_{g,\mu}$ in (iii)). By combining (A.2) and (A.3) we establish the validity of (11). □

References

[1] S. Becker, J. Bobin, and E. Candes. NESTA: A fast and accurate first-order method for sparse recovery. *SIAM J. on Imaging Sciences*, 4(1):1–39, 2011. 4

[2] Q. Chen, Z. Song, S. Liu, X. Chen, X.-T. Yuan, T.-S. Chua, S. Yan, Y. Hua, Z. Huang, and S. Shen. Boosting classification with exclusive context. http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2010/workshop/nuspsl.pdf. 5, 6

[3] X. Chen, Y. Mu, S. Yan, and T.-S. Chua. Efficient large-scale image annotation by probabilistic collaborative multi-label propagation. In *ACM MM*, 2010. 1, 6

[4] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In *CVPR*, 2010. 1

[5] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. NUS-WIDE: A real-world web image database from national university of singapore. In *CIVR*, 2009. 3, 5

[6] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, 2009. 1

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 5

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html. 6

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2), 2010. 5

[10] M. Fornasier and H. Rauhut. Recovery algorithm for vector-valued data with joint sparsity constraints. *SIAM Journal on Numerical Analysis*, 46(2):577–613, 2008. 2

[11] T. Griffiths and Z. Ghahramani. Infinite latent feature models and the indian buffet process. In *NIPS*, 2005. 1

[12] L. Jacob, G. Obozinski, and J. Vert. Group lasso with overlap and graph lasso. In *ICML*, 2009. 2

[13] M. Kowalski. Sparse regression using mixed norms. *Applied and Computational Harmonic Analysis*, 27(3):303–324, 2009. 2

[14] M. Kowalski and B. Torreesani. Sparsity and persistence: mixed norms provide simple signals models with dependent coefficient. *Signal, Image and Video Processing*, 2008. 3

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 5

[16] H. Liu, M. Palatucci, and J. Zhang. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *ICML*, 2009. 2

[17] H. Liu and S. Yan. Robust graph mode seeking by graph shift. In *ICML*, 2010. 2, 3

[18] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005. 2, 4, 7

[19] K. Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *TPAMI*, 32(9):1582–1596, 2010. 5

[20] A. Subramanya and J. Bilmes. Entropic graph regularization in non-parametric semi-supervised classification. In *NIPS*, 2009. 6

[21] J. Tang, S. Yan, R. Hong, G.-J. Qi, and T.-S. Chua. Inferring semantic concepts from community-contributed images and noisy tags. In *ACM MM*, 2009. 2

[22] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal of Optimization*, 2008. 2, 4

[23] N. Ueda and K. Saito. Parametric mixture models for multi-labeled text. In *NIPS*, 2002. 1

[24] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010. 5

[25] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, and Y. Song. Unified video annotation via multi-graph learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(5):733–746, 2009. 2

[26] M. Wang, X.-S. Hua, J. Tang, and R. Hong. Beyond distance measurement: Constructing neighborhood similarity for video annotation. *IEEE Transactions on Multimedia*, 11(3):465–476, 2009. 2

[27] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *TPAMI*, 31(2):210–226, 2009. 2

[28] S. Yan and H. Wang. Semi-supervised learning by sparse representation. In *SDM*, 2009. 2

[29] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68(1):49–67, 2006. 2, 3

[30] X.-T. Yuan and S. Yan. Visual classification with multi-task joint sparse representation. In *CVPR*, 2010. 2

[31] J. Zhang. A probabilistic framework for multi-task learning. Technical report, CMU-LTI-06-006, 2006. 2

[32] P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497, 2009. 2

[33] X. Zhou, K. Yu, T. Zhang, and T. Huang. Image classification using super-vector coding of local image descriptors. In *ECCV*, 2010. 5

[34] Y. Zhou, R. Jin, and S. C. Hoi. Exclusive lasso for multi-task feature selection. In *AISTATS*, 2010. 2, 3

[35] S. Zhu, X. Ji, W. Xu, and Y. Gong. Multi-labelled classification using maximum entropy method. In *ACM SIGIR*, 2005. 1