Exploiting Low-rank Structure for Discriminative Sub-categorization

Zheng Xu[1]
xuzhustc@gmail.com
Xue Li[1]
echolixue@gmail.com
Kuiyuan Yang[2]
kuyang@microsoft.com
Tom Goldstein[1]
tomg@cs.umd.edu

[1] Department of Computer Science, University of Maryland, College Park, USA
[2] Microsoft Research, Beijing, China

Abstract

In visual recognition, sub-categorization has been proposed to deal with large intra-class variance of samples in a category. Instead of learning a single classifier for each category, discriminant sub-categorization approaches divide a category into several sub-categories and simultaneously train classifiers for each sub-category. In this paper, we propose a novel approach for discriminative sub-categorization. Our method jointly trains the exemplar classifier for each positive sample to address the intra-variance of a category and exploits the low rank structure to preserve common information while discovering sub-categories. We formulate the problem as a convex objective function and introduce an efficient solver based on alternating direction method of multipliers. Comprehensive experiments on various datasets demonstrate the effectiveness and efficiency of the proposed method in both sub-category discovery and visual recognition.

1 Introduction

Visual recognition is one of the central challenges in computer vision, which is often evaluated by the performance of classifying images into pre-defined categories from the vocabulary of a lexical database [8, 36]. However, a category in the real world contains large intra-variance, which increases the difficulty in modeling and classification. The sub-categorization approach, which divides a category into several sub-categories to deal with the intra-variance challenge, has been successfully applied to many applications including object detection and classification [7, 9, 11, 15, 22, 26, 27]. Sub-categories are often automatically discovered by unsupervised clustering [14, 16, 24, 28, 32, 38]. Meanwhile, recent discriminant sub-categorization approaches utilize samples that do not belong to the category under consideration as negative data for supervision, and cluster positive samples of the category into sub-categories, then simultaneously train the corresponding classifier for each sub-category [1, 15, 22, 41].

In the unified clustering and classification framework of previous methods [15, 22], the classifier for each sub-category is trained by using samples hard-assigned to the sub-category.

Pages 149.1-149.12
DOI: https://dx.doi.org/10.5244/C.29.149

However, it is difficult to separate samples based on the hard boundary between two sub-categories while training, since the intra-variance of a category is caused by complex factors in the real world. The visual appearance of samples changes continuously and some samples could contribute to the training of several sub-categories. Moreover, sub-categories are closely related since they are discovered from the same category. The common information among these sub-categories is beneficial for the training of classifiers. To utilize common information when learning multiple related classifiers has been studied from the perspective of multi-task learning [2, 6, 25]. Instead of using pre-defined tasks in multi-task learning, sub-categorization needs to discover tasks (sub-categories) during learning.

In this paper, we propose a new approach for discriminative sub-categorization, which adopts the exemplar based method to address the intra-variance in each category, and exploits low rank structure to preserve common information while discovering sub-categories. Our approach builds up the exemplar-LDAs [21], which generate a set of exemplar classifiers with each classifier trained by a single positive sample and all the negative samples. The extreme case of sub-categorization is the atomic sub-category, which has only one positive sample, and is a compact set for training and modeling. In order to share common information among exemplar classifiers while preserving diversity, we jointly train the exemplar-LDAs for all the positive samples and introduce the trace-norm regularizer on the matrix of weights of exemplar-LDAs, as we assume the weights lie on a union of subspaces such that the matrix of weights is low-rank. To solve the convex formulation in our model, we propose an efficient algorithm based on the scaled form of alternating direction method of multipliers (scaled ADMM) [4]. We conduct comprehensive experiments on various datasets to validate the effectiveness of our approach in sub-category discovery and visual recognition.

2 Low-rank least squares exemplar-LDAs

In this section, we introduce the proposed low-rank least squares exemplar-LDAs (LRLSE-LDAs) for sub-categorization. We start from conducting the least squares form of exemplar-LDAs, and then form the LRLSE-LDAs by adding the trace-norm regularizer on the weights of exemplar classifiers. After that, we introduce an efficient algorithm based on scaled ADMM to solve the optimization problem of LRLSE-LDAs. Finally, we discuss LRLSE-LDAs based clustering and classification method for sub-category discovery and visual recognition.

In the following sections, vectors/matrices are denoted by lowercase/uppercase letters in boldface. The transpose of a vector/matrix is denoted by using superscript $'$. $\mathbf{A} = [a_{ij}] \in \mathbb{R}^{m\times n}$ defines a matrix $\mathbf{A}$ with a_{ij} being its (i,j)-th element for $i = 1,\ldots,m$ and $j = 1,\ldots,n$, $\mathbf{1}_n \in \mathbb{R}^n$ represents a vector with all entries being 1. $\mathbf{0}_n \in \mathbb{R}^n$ represents a vector with all entries being 0, and $\mathbf{I}_n \in \mathbb{R}^{n\times n}$ is an identity matrix.

2.1 Least squares exemplar-LDA

We conduct the least squares form of exemplar-LDAs in this section. Exemplar-SVMs and exemplar-LDAs have been wildly adopted in visual recognition [3, 7, 21, 29, 39]. Each exemplar classifier is learned from one positive sample and all the negative samples. We adopt exemplar classifiers to represent atomic sub-categories and preserve intra-variance in each category. Particularly, our approach builds up the exemplar-LDAs [21], which has a closed-form solution, requires little effort to train, and achieves similar performance as exemplar-SVMs.

Let $\mathcal{S} = \mathcal{S}^+ \cup \mathcal{S}^-$ denote the set of training samples, in which $\mathcal{S}^+ = \{\mathbf{s}_1^+, \ldots, \mathbf{s}_n^+\}$ is the set of positive training samples, and $\mathcal{S}^- = \{\mathbf{s}_1^-, \ldots, \mathbf{s}_m^-\}$ is the set of negative training samples. Each training sample $\mathbf{s}^+$ or $\mathbf{s}^-$ is a d-dimensional column vector, *i.e.*, $\mathbf{s}^+, \mathbf{s}^- \in \mathbb{R}^d$. Each exemplar-LDA for a positive sample $\mathbf{s}_i^+$ is a linear classifier represented by weight vector $\mathbf{w}_i$ trained by maximizing the following Fisher criterion:

$$J_{Fisher}(\mathbf{w}_i) = \frac{\mathbf{w}_i'(\mathbf{s}_i^+ - \boldsymbol{\mu}^-)(\mathbf{s}_i^+ - \boldsymbol{\mu}^-)'\mathbf{w}_i}{\mathbf{w}_i'\mathbf{Q}\mathbf{w}_i}, \tag{1}$$

where $\boldsymbol{\mu}^-$ is the mean of negative samples in $\mathcal{S}^-$, $\mathbf{Q} = \sum_{s \in \mathcal{S}^-}(\mathbf{s} - \boldsymbol{\mu}^-)(\mathbf{s} - \boldsymbol{\mu}^-)'$ is a scaled covariance matrix of the negative samples. The Fisher criterion aims to simultaneously maximize between-class distance and minimize within-class distance. The training of each exemplar-LDA benefits from the closed-form solution of the two class linear discriminant analysis [13, 21]. As a special case of LDA with only one positive sample, exemplar-LDA has a closed form solution, *i.e.*,

$$\mathbf{w}_i = \mathbf{Q}^{-1}(\mathbf{s}_i^+ - \boldsymbol{\mu}^-) \tag{2}$$

The Fisher criterion in Eq. 1 is a non-convex function. However, it is known that the LDA solution can be achieved by a linear regression model with the proper choice of regression labels [13, 30, 40]. Following [13, 30, 40], we now conduct the least squares formulation for exemplar-LDAs. While training the exemplar-LDA weight $\mathbf{w}_i$ for the positive sample $\mathbf{s}_i^+$, we choose the regression label for the positive sample $\mathbf{s}_i^+$ as $z_i^+ = 1$ and regression labels for all negative samples $\mathbf{s}_j^- \in \mathcal{S}^-$ as $z_j^- = -\frac{1}{m}$. We then formulate the least squares form of the objective function as:

$$J_1(\mathbf{w}_i) = \frac{1}{2}\sum_{j=1}^{m} \|(\mathbf{s}_j^- - \boldsymbol{\mu}_i)'\mathbf{w}_i + \frac{1}{m}\|^2 + \frac{1}{2}\|(\mathbf{s}_i^+ - \boldsymbol{\mu}_i)'\mathbf{w}_i - 1\|^2 \tag{3}$$

where $\boldsymbol{\mu}_i = \frac{1}{m+1}(\sum_{\mathbf{s} \in \mathcal{S}^-} \mathbf{s} + \mathbf{s}_i^+)$ is the mean of all training samples for learning $\mathbf{w}_i$.

For exemplar-LDA, the number of negative samples is much larger than the single positive sample, so $\boldsymbol{\mu}_i$ can be approximated by the mean $\boldsymbol{\mu}^-$ of the negative samples. We represent the centered sample as:

$$\begin{aligned} \mathbf{x}_i^+ &= \mathbf{s}_i^+ - \boldsymbol{\mu}^-, \forall \mathbf{s}_i^+ \in \mathcal{S}^+, \\ \mathbf{x}_j^- &= \mathbf{s}_j^- - \boldsymbol{\mu}^-, \forall \mathbf{s}_j^- \in \mathcal{S}^-, \end{aligned} \tag{4}$$

and write the positive and negative samples in the matrix form as $\mathbf{X}_1 = [\mathbf{x}_1^+, \ldots, \mathbf{x}_n^+]$ and $\mathbf{X}_2 = [\mathbf{x}_1^-, \ldots, \mathbf{x}_m^-]$. Since the samples are centered by negative mean $\boldsymbol{\mu}^-$, we have $\mathbf{X}_2\mathbf{1}_m = \mathbf{0}_d$. Then the objective function is transformed into:

$$J_2(\mathbf{w}_i) = \frac{1}{2}\|\mathbf{X}_2'\mathbf{w}_i\|^2 + \frac{1}{2}\|\mathbf{w}_i'\mathbf{x}_i^+ - 1\|^2. \tag{5}$$

Furthermore, we ignore the quadratic term $\frac{1}{2}\|\mathbf{w}_i'\mathbf{x}_i^+\|^2$ for the positive sample as it is relatively small compared with the quadratic term for negative samples:

$$J_3(\mathbf{w}_i) = \frac{1}{2}\|\mathbf{X}_2'\mathbf{w}_i\|^2 - \mathbf{w}_i'\mathbf{x}_i^+. \tag{6}$$

The solution of minimizing $J_3(\mathbf{w}_i)$ can be achieved by setting its gradient to zero:

$$\mathbf{w}_i = \arg\min_{\mathbf{w}_i} J_3(\mathbf{w}_i) = (\mathbf{X}_2\mathbf{X}_2')^{-1}\mathbf{x}_i^+, \tag{7}$$

which is the same as the closed-form solution for exemplar-LDA in Eq. 2. The least squares form of exemplar-LDA in Eq. 6 minimizes the within-class distance by the quadratic term $\frac{1}{2}\|\mathbf{X}_2'\mathbf{w}_i\|^2$ and maximizes the between class distance by the linear term $\mathbf{w}_i'\mathbf{x}_i^+$.

In practice, the matrix $\mathbf{Q} = \mathbf{X}_2\mathbf{X}_2'$ is regularized by adding a small value to its diagonal entries to avoid unstable solutions, which is equivalent to adding an isotropic prior on $\mathbf{w}_i$ modeled by an ℓ_2 regularizer. Adding this regularizer, we finally formulate the least squares exemplar-LDA objective function for sample $\mathbf{s}_i^+$ as:

$$J^i_{LSE-LDA}(\mathbf{w}_i) = \frac{\delta}{2}\|\mathbf{w}_i\|^2 + \frac{1}{2}\|\mathbf{X}_2'\mathbf{w}_i\|^2 - \mathbf{w}_i'\mathbf{x}_i^+. \tag{8}$$

2.2 Low-rank least squares exemplar-LDAs

From the least squares exemplar-LDA objective function for a single exemplar in Eq. 8, we construct the objective function for learning the weight matrix $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_n] \in \mathbb{R}^{d\times n}$ for all the exemplars in $\mathcal{S}^+$:

$$J_{LSE-LDAs}(\mathbf{W}) = \frac{\delta}{2}\|\mathbf{W}\|_F^2 + \frac{1}{2}\|\mathbf{X}_2'\mathbf{W}\|_F^2 - \mathrm{trace}(\mathbf{X}_1'\mathbf{W}) \tag{9}$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix, $\mathrm{trace}(\cdot)$ represents the trace of a matrix.

The weight $\mathbf{w}_i$ for each exemplar-LDA is trained by the positive sample $\mathbf{s}_i^+$ and all negative samples $\mathcal{S}^-$, but does not utilize the other positive samples $\mathbf{s}_k^+ \in \mathcal{S}^+, k \neq i$. In this way, the positive training samples for each exemplar-LDA are compact and the diversity of exemplars is preserved. However, the common information is not utilized by training exemplar-LDA for each positive sample independently, as samples of a sub-category have close relation. We assume the weights of exemplar-LDAs for samples from the same sub-category lie on a subspace. Therefore the weight matrix $\mathbf{W}$ lies on a union of subspaces and should be low-rank. To discover the structure of sub-categories, we jointly learn the weight for positive samples/exemplars of the category and regularize the weight matrix with a low-rank constraint. Finally, we arrive at the objective function for our low-rank least squares exemplar-LDAs:

$$J_{LRLSE-LDAs}(\mathbf{W}) = \xi\|\mathbf{W}\|_* + J_{LSE-LDAs}(\mathbf{W}) \tag{10}$$

$\|\cdot\|_*$ is the trace norm used to constrain the weight matrix, which is a convex approximation of the rank of a matrix [20, 39]. The formulation is convex and can be solved efficiently using the algorithm introduced in the following section.

2.3 Scaled ADMM for LRLSE-LDAs

In this section, we discuss how to minimize the objective function for our LRLSE-LDAs model in Eq. 10. We adopt the alternating direction method of multipliers (ADMM) in scaled form [2, 17] to solve the minimization problem. Minimizing objective function

$J_{LRLSE-LDAs}(\mathbf{W})$ in Eq. 10 can be equivalently transformed into an equality-constrained convex optimization problem by introducing an intermediate variable $\mathbf{F}$:

$$\min_{\mathbf{W},\mathbf{F}} J_{LSE-LDAs}(\mathbf{W}) + \xi\|\mathbf{F}\|_* \quad \text{s.t. } \mathbf{W} = \mathbf{F} \tag{11}$$

Then the augmented Lagrangian for the formulation in Eq. 11 can be written as:

$$L(\mathbf{W},\mathbf{F},\mathbf{\Lambda}) = J_{LSE-LDAs}(\mathbf{W}) + \xi\|\mathbf{F}\|_* + \frac{\tau}{2}(\|\mathbf{W}-\mathbf{F}+\mathbf{\Lambda}\|_F^2 - \|\mathbf{\Lambda}\|_F^2) \tag{12}$$

where $\mathbf{\Lambda}$ is the scaled dual parameter matrix, and τ is the penalty parameter. We iteratively update variables $\mathbf{W},\mathbf{F},\mathbf{\Lambda}$ to solve Eq. 12, where $\mathbf{W},\mathbf{F}$ are updated by solving two subproblems both with closed-form solutions, and $\mathbf{\Lambda}$ is updated by dual ascent. We describe the steps in detail as follows.

Update W: Each iteration begins by fixing variables $\mathbf{F}$ and $\mathbf{\Lambda}$ and updating the weight matrix $\mathbf{W}$ by solving the subproblem

$$\mathbf{W} = \arg\min_{\mathbf{W}} J_{LSE-LDAs}(\mathbf{W}) + \frac{\tau}{2}\|\mathbf{W}-\mathbf{F}+\mathbf{\Lambda}\|_F^2 \tag{13}$$

$$= (\mathbf{X}_2\mathbf{X}_2' + (\delta+\tau)\mathbf{I}_d)^{-1}(\mathbf{X}_1 + \tau(\mathbf{F}-\mathbf{\Lambda})). \tag{14}$$

The closed-form solution of Eq. 14 is derived by setting the gradient of the objective function in Eq. 13 to zero.

Update F: After updating $\mathbf{W}$, we fix the weight matrix $\mathbf{W}$ and dual parameter $\mathbf{\Lambda}$, and update the low-rank matrix $\mathbf{F}$ by solving the subproblem:

$$\mathbf{F} = \arg\min_{\mathbf{F}} \xi\|\mathbf{F}\|_* + \frac{\tau}{2}\|\mathbf{W}-\mathbf{F}+\mathbf{\Lambda}\|_F^2 \tag{15}$$

Eq. 15 can be solved by the singular value thresholding (SVT) method [5]. We first compute the singular value decomposition for matrix $\mathbf{W}+\mathbf{\Lambda}$ as $\mathbf{U}\mathbf{\Gamma}\mathbf{V}'$, where $\mathbf{U},\mathbf{V}$ are two orthogonal matrices, and $\mathbf{\Gamma}$ is a diagonal matrix of singular values. We represent the singular values as $\{\gamma_1,\ldots,\gamma_l\}$ where $l = \min\{n,d\}$. We update intermediate variable $\mathbf{F}$ with a closed-form solution by SVT as:

$$\mathbf{F} = \mathbf{U}\mathcal{D}(\mathbf{\Gamma},\xi/\tau)\mathbf{V}' \tag{16}$$

where $\mathcal{D}(\cdot,\cdot)$ is the singular value shrinkage operator. The result of $\mathcal{D}(\mathbf{\Gamma},\xi/\tau)$ is a diagonal matrix with shrinked nonnegative singular values $\{\gamma_1^*,\ldots,\gamma_l^*\}$ as diagonal elements, where $\gamma_k^* = \max\{0,\gamma_k - \xi/\tau\}, \forall k \in \{1,\ldots,l\}$.

Update Λ: The last step of the iteration updates the scaled dual variable $\mathbf{\Lambda}$. We update $\mathbf{\Lambda}$ to $\mathbf{\Lambda}^*$ by gradient ascent with step size 1 in scaled ADMM:

$$\mathbf{\Lambda}^* = \mathbf{\Lambda} + \mathbf{W} - \mathbf{F}. \tag{17}$$

Algorithm Discussion: We summarize the optimization procedure for our low-rank least squares exemplar-LDAs in Algorithm 1. We first preprocess input samples by centering the samples with Eq. 4. Then we initialize variables $\mathbf{W},\mathbf{F},\mathbf{\Lambda}$ with uniformly random values between 0 and 1 and start the iterations. In each iteration, we sequentially update $\mathbf{W}$, $\mathbf{F}$, and $\mathbf{\Lambda}$ by Eqautions 14,16, and 17, respectively . The above update steps are repeated until convergence or the maximum number of iterations is reached.

Algorithm 1 Optimization for LRLSE-LDAs

Input: Training data $\mathcal{S}$, and parameters δ, ξ.

1: Center input data by Eq. 4
2: Random initialize $\mathbf{W}, \mathbf{F}, \mathbf{\Lambda}$
3: **repeat**
4: Update $\mathbf{W}$ by closed-form solution in Eq. 14 with current $\mathbf{F}, \mathbf{\Lambda}$
5: Update $\mathbf{F}$ by the SVT method in Eq. 16 with updated $\mathbf{W}$ and current $\mathbf{\Lambda}$
6: Update $\mathbf{\Lambda}$ by Eq. 17 with updated $\mathbf{W}, \mathbf{F}$
7: **until** Convergence or the maximum number of iterations is reached.

Output: The weight matrix $\mathbf{W}$.

2.4 Sub-category discovery and visual recognition

After efficiently solving the LRLSE-LDAs (Eq. 10) using Algorithm 1, we have obtained the weight matrix $\mathbf{W}$, in which each column $\mathbf{w}_i$ is the weight of an exemplar-LDA classifier. We now investigate how to utilize the weights discovered from the LRLSE-LDAs (Eq. 10) to explicitly find interpretable sub-categories, and utilize the low-rank structure from sub-categories for visual recognition.

LRLSE-LDAs exploits the structure of sub-categories implicitly by adding the low-rank regularizer on the exemplar-based methods. To demonstrate the ability of LRLS-ELDA in discovering sub-categories, we use prediction scores of exemplar-LDAs on positive samples to perform clustering. To make prediction scores of exemplar-LDAs comparable to each other, we normalize the score for a test sample $\mathbf{t}$ with the prediction scores of the average negative training samples for each classifier $\mathbf{w}_i$:

$$p(\mathbf{w}_i, \mathbf{t}) = \mathbf{w}_i' \mathbf{t} - \mathbf{w}_i' \boldsymbol{\mu}^-. \tag{18}$$

We then compute the affinity matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ for positive samples. We define each entry $a_{ij} \in \mathbf{A}$ that represents affinity between $\mathbf{s}_i^+$ and $\mathbf{s}_j^+$ as:

$$a_{ij} = \begin{cases} \max\{p(\mathbf{w}_j, \mathbf{s}_i^+) + p(\mathbf{w}_i, \mathbf{s}_j^+),\ 0\}, & i \neq j \\ 0, & i = j. \end{cases} \tag{19}$$

A high prediction score $p(\mathbf{w}_j, \mathbf{s}_i^+)$ means $\mathbf{s}_i^+$ can be recognized by exemplar classifier $\mathbf{w}_j$. High scores of $p(\mathbf{w}_j, \mathbf{s}_i^+)$ and $p(\mathbf{w}_i, \mathbf{s}_j^+)$ lead to high affinity and indicate $\mathbf{s}_i^+$ and $\mathbf{s}_j^+$ are likely from the same sub-category. We perform spectral clustering [32] with this affinity matrix.

For visual recognition, we adopt an approach inspired by Xu *et al.* for domain generalization [39]. The top K prediction scores from trained exemplar classifiers are fused together to form the final prediction score for a sample $\mathbf{t}$:

$$p(\mathbf{W}, \mathbf{t}) = \frac{1}{K} \sum_{i \in \mathcal{T}(\mathbf{t})} p(\mathbf{w}_i, \mathbf{t}), \tag{20}$$

where $\mathcal{T}(\mathbf{t}) = \{k | 1 < k < n,\ p(\mathbf{w}_k, \mathbf{t})$ is one of the top K predictions for $\mathbf{t}\}$ is the set of indices of selected exemplar classifiers. By selecting the top K exemplar classifiers, we have implicitly assigned the test sample to the same sub-category as the selected K exemplars. The multiclass classification is achieved by a one-versus-all strategy with the fused prediction scores.

Table 1: The statistics of datasets used to evaluate sub-category discovery. This table shows the number of classes, the dimension of features, and the number of samples for each dataset.

Dataset	#classes	#features	#points	Dataset	#classes	#features	#points
Gas Sensor	6	128	13910	Semeion	10	256	1593
Landsat	6	36	4435	MNIST	10	784	60000
Segmentation	7	19	2310	Letter	26	16	20000
Steel Plates	7	27	1941	Isolet	26	617	6238
Digits	10	64	5620	Amazon	50	10000	1500

3 Experiments

We evaluate our low-rank least squares exemplar-LDAs (LRLSE-LDAs) approach for sub-category discovery and visual recognition. We perform LRLSE-LDAs based clustering on various public datasets for sub-category discovery, and perform LRLSE-LDAs based classification on object recognition and human activity recognition datasets for visual recognition.

3.1 Sub-category discovery

We validate our LRLSE-LDAs based clustering introduced in Section. 2 on several publicly available datasets to demonstrate the capacity of discovering interpretable sub-categories in our approach. Following [22], we use the MNIST dataset and nine datasets from the UCI repository: Gas Sensor, Landsat, Segmentation, Steel Plates, Digits, Semeion, Letter, Isolet, and Amazon. We exclude the Wine Quality dataset because the public data provide 11 dimensional feature, which is different from the 12 dimensional feature claimed in [22]. Results are summarized in Tab. 1. The datasets used cover a wide range of types types from images, texts and sensors. The number of samples varies from 1500 to 60000, and the dimensionality of features varies from 16 to 10000.

Following the experimental setting in [22], we randomly split ground truth classes of each dataset into two roughly equal halves, one is regarded as positive and the other as negative. Then both data sets are randomly divided into training and validation subsets. We use the training data to learn the weights of our LRLSE-LDAs by Algorithm 1, and use the binary classification performance on the validation set to tune the hyper-parameter ξ for the low-rank regularizer in LRLSE-LDAs (Eq. 10) among $\{0.01, 0.1, 1, 10, 100\}$ by cross-validation. In all our experiments, we fix the the least squares penalty parameter in exemplar-LDA (Eq. 8) to $\delta = 1$. We adopt the LRLSE-LDAs based clustering method introduced in Section 2.4 to discover sub-categories, using the number of ground truth classes for clustering. The clustering performance can be evaluated by many different metrics [12]. To measure the performance of sub-category discovery, we calculated purity [22] for clustering as follows: First, we remove all the ground truth labels and perform clustering. Then, we find the best one-to-one association between clusters and ground truth labels by the Hungarian algorithm and calculate the percentage of correct assignments. Each experiment is repeated 50 times and the means and standard errors are reported.

We compare our LRLSE-LDAs based clustering method against k-means, latent SVM (LSVM) [15] and discriminant sub-categorization method (Sub-C) [22], which are reported in [22]. We additionally report the results of exemplar-LDAs (E-LDAs), which is a special case of our method when the low-rank parameter ξ is set to zero. We apply the same clustering method introduced in Section 2.4 for E-LDAs baseline and our LRLS-LDA approach.

Table 2: Clustering purity measures for discovering sub-categories. The mean and standard error of 50 runs are reported for different methods on each dataset. Results within one standard error of the maximum value are denoted in boldface.

Dataset	k-means	LSVM [15]	Sub-C [22]	E-LDAs	LRLSE-LDAs
Gas Sensor	46.38 ± 0.69	56.74 ± 1.88	60.82 ± 1.64	67.02 ± 1.69	**70.95 ± 1.82**
Landsat	78.72 ± 2.08	69.37 ± 2.32	76.73 ± 2.38	79.01 ± 1.71	**81.52 ± 1.07**
Segmentation	71.96 ± 1.75	65.89 ± 2.36	74.41 ± 1.85	**80.97 ± 1.87**	**82.00 ± 1.54**
Steel Plates	**53.29 ± 1.51**	**52.64 ± 2.02**	**54.60 ± 1.98**	50.74 ± 1.32	**54.44 ± 1.20**
Digits	76.38 ± 1.72	77.83 ± 1.57	80.15 ± 1.18	88.34 ± 0.89	**90.61 ± 0.88**
Semeion	64.64 ± 1.20	64.32 ± 1.58	66.74 ± 1.43	68.24 ± 1.49	**71.71 ± 1.50**
MNIST	65.38 ± 1.43	63.99 ± 1.36	66.18 ± 1.34	65.80 ± 1.33	**95.20 ± 0.81**
Letter	33.35 ± 0.48	40.27 ± 0.88	44.38 ± 0.74	45.48 ± 0.77	**47.27 ± 0.71**
Isolet	62.15 ± 1.22	61.95 ± 1.22	64.08 ± 1.18	75.23 ± 1.11	**76.71 ± 1.15**
Amazon	24.93 ± 0.32	24.89 ± 0.41	25.08 ± 0.38	**48.38 ± 0.57**	46.90 ± 2.28

The experimental results are summarized in Tab. 2. The table shows that the benefit of LSVM over k-means baseline is uncertain since LSVM improves performance on two datasets, decreases performance on two datasets, and achieves similar performance on the other datasets. Meanwhile, the Sub-C method performs at least as well as LSVM and k-means, outperforms LSVM on eight datasets, and outperforms k-means on six datasets. Our LRLSE-LDAs based clustering method performs best on eight out of ten datasets and ranks second on the two exceptions: Steel Plates and Amazon. Our LRLS-LDA method improves the performance of exemplar-LDA on all the datasets except Amazon, which demonstrates the ability of the low-rank regularizer in Eq. 10 to discover sub-categories. By the low-rank regularizer, we assume weights of exemplar classifiers for samples from the same sub-category lie on a subspace and are similar to each other. While using those classifiers to build the affinity matrix for clustering in Section 2.4, strong connections are achieved between samples from the same sub-category. For the Amazon dataset, the small number of samples (1500 samples from 50 classes) with high dimensional features (10000 dimensions) may be quite different from each other, hence each sample would form a sub-category (which is the assumption of exemplar based methods). Note that the results of LRLSE-LDAs are generated by tuning the low-rank parameter ξ using cross-validation. Our LRLSE-LDAs could achieve similar results as E-LDAs by setting the low-rank parameter as $\xi = 0$.

In addition ,we discuss an observation on parameters. On all ten datasets in Tab. 2, the performance of our LRLSE-LDAs is slightly better or not worse than the results listed in the table when we fix the low-rank parameter $\xi = 0.1$. Since the results are generated by tuning ξ based on cross validation of binary classification, this indicates there may exist a gap between using LRLSE-LDAs for clustering and for classification.

3.2 Visual recognition

We further evaluate the performance of our LRLSE-LDAs approach for visual recognition following the domain generalization experiment in [39]. Domain generalization focuses on cross-domain visual recognition, which assumes the training samples are from several domains and the testing samples are from unseen domains [25, 31, 33]. Recent research of latent domains [19, 23, 39] automatically discovers domains in samples, where each domain can be regarded as a sub-category. Particularly, Xu *et al.* [39] proposed to exploit low-

rank structure by introducing a trace-norm regularizer on the likelihood matrix of exemplar-SVMs. Our approach differs from this work in several ways. First, our trace-norm regularizer on the weight matrix is based on the assumption that the weights of exemplar classifiers lie on a union of subspaces. Moreover, our objective function is convex and could be efficiently solved by ADMM since each subproblem has a closed-form solution, which benefits from using LDA instead of SVM and constraining weight matrix instead of likelihood matrix. We also focus more on the advantage of exploiting low-rank structure for sub-category discovery.

Two datasets are used in our experiments for visual recognition, one is the Office-Caltech dataset [18, 34] for object recognition, and the other is the IXMAS dataset [37] for action recognition, following [19, 39]. The Office-Caltech dataset contains images from four domains: Amazon (A), Caltech-256 (C), images captured by digital SLR camera (D) ,and web camera (W). We extract DeCAF$_6$ features [10] using the pre-trained Convolutional Neural Networks for images in Office-Caltech. The IXMAS dataset contains the videos from eleven actions captured by five cameras (Cam 0, Cam 1, ..., Cam 4) from different viewpoints, and each action is performed three times by twelve actors. To exclude the irregularly performed actions as suggested by [19, 39], we keep the first five actions (check watch, cross arms, scratch head, sit down, get up) performed by six actors (*Alba, Andreas, Daniel, Hedlena, Julien, Nicolas*). Dense trajectories features [35] are extracted to form a bag-of-words representation for each video sequence. For the two datasets, we mix samples from several ground truth domains/views for training and testing, and perform cross-domain visual recognition. The classification performance is measured by multi-class recognition accuracy.

We compare our LRLSE-LDAs approach against the discriminant sub-categorization (Sub-C) approach [22] and the low-rank exemplar-SVMs (LRE-SVMs) approach [39], which are reported in [39]. We additionally report the results of SVM, exemplar-SVMs (E-SVMs), and exemplar-LDAs (E-LDAs) as baselines. The parameter in our LRLSE-LDAs is empirically fixed as $\xi = 0.1$ for all the experiments. In cross-domain recognition, we do not use cross validation since the distributions of training and testing samples are different. The experimental results are summarized in Tab. 3. In this experiment, methods that consider sub-categories outperform SVM methods that do not consider sub-categories. Our LRLSE-LDAs outperforms Sub-C in five out of six cases, and is generally better than E-SVMs and E-LDAs without exploiting the low-rank structure. SVM based exemplar classifiers perform slightly better than LDA based exemplar classifiers. Comparing with LRE-SVMs, which exploits low-rank structure from likelihood matrices to build up exemplar-SVMs, LRLSE-LDAs performs slightly worse. However, our LRLSE-LDAs has a convex formulation that could be solved efficiently. Our LRLSE-LDAs usually converges in less than 20 iterations, while LRE-SVMs usually stops when the maximum iteration number is achieved. In all the six cases, the training of LRLSE-LDAs only takes seconds while LRE-SVMs takes hours, as shown in Tab. 3. The time is evaluated with an Intel Core i7 2.9GHz CPU.

4 Conclusion

We proposed a novel approach for discriminative sub-categorization. Our approach trains the exemplar classifiers for atomic sub-categories to address the intra-variance of a category, and simultaneously exploits low rank structure to preserve information while discovering sub-categories. The problem is formulated as a convex objective function that is efficiently solved using ADMM. We conduct comprehensive experiments on various datasets to validate the effectiveness and efficiency of our method for sub-category discovery and visual recognition.

Table 3: Recognition accuracies (%) of different methods for cross-domain visual recognition. The first- and second-best results for each dataset are denoted in boldface and with underline, respectively. Note that our LRLSE-LDAs is much faster than LRE-SVMs without losing much performance.

Training		A,C	D,W	C,D,W	Cam 0,1	Cam 2,3,4	Cam 0,1,2,3
Testing		D,W	A,C	A	Cam 2,3,4	Cam 0,1	Cam 4
SVM		82.68	76.06	90.73	71.70	63.83	56.61
Sub-C [22]		82.61	78.65	90.75	<u>78.11</u>	76.90	64.04
E-SVMs		82.73	80.85	91.47	76.86	68.04	<u>72.98</u>
LRE-SVMs [39]		<u>84.59</u>	**81.17**	**91.87**	**79.96**	**80.15**	**74.97**
E-LDAs		82.56	80.30	91.24	76.82	65.56	67.08
LRLSE-LDAs		**84.99**	<u>80.92</u>	<u>91.86</u>	77.39	<u>77.78</u>	68.19
Train time (seconds)	LRE-SVMs	1.51e+4	2.19e+3	1.09e+4	2.06e+3	1.95e+3	3.56e+3
	LRLSE-LDAs	24.21	7.67	15.44	49.67	45.53	36.52

References

[1] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, 2002.

[2] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. *NIPS*, 2007.

[3] Yusuf Aytar and Andrew Zisserman. Enhancing exemplar-SVMs using part level transfer regularization. In *BMVC*, 2012.

[4] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[5] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[6] Jianhui Chen, Jiayu Zhou, and Jieping Ye. Integrating low-rank and group-sparse structures for robust multi-task learning. In *SIGKDD*, 2011.

[7] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. NEIL: Extracting visual knowledge from web data. In *ICCV*, 2013.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[9] Santosh K. Divvala, Alexei A. Efros, and Martial Hebert. How important are 'deformable parts' in the deformable parts model? In *Parts and Attributes Workshop, ECCV*, 2012.

[10] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014.

[11] Jian Dong, Wei Xia, Qiang Chen, Jianshi Feng, Zhongyang Huang, and Shuicheng Yan. Subcategory-aware object classification. In *CVPR*, 2013.

[12] Kun Duan, David J Crandall, and Dhruv Batra. Multimodal learning in loosely-organized web images. In *CVPR*, 2014.

[13] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

[14] Ehsan Elhamifar and René Vidal. Sparse subspace clustering. In *CVPR*, 2009.

[15] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.

[16] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

[17] Tom Goldstein, Gavin Taylor, Kawika Barabin, and Kent Sayre. Unwrapping ADMM: efficient distributed computing via transpose reduction. *CoRR*, abs/1504.02147, 2015.

[18] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012.

[19] Boqing Gong, Kristen Grauman, and Fei Sha. Reshaping visual datasets for domain adaptation. In *NIPS*, 2013.

[20] Zaid Harchaoui, Matthijs Douze, Mattis Paulin, Miroslav Dudik, and Jérôme Malick. Large-scale image classification with trace-norm regularization. In *CVPR*, 2012.

[21] Bharath Hariharan, Jitendra Malik, and Deva Ramanan. Discriminative decorrelation for clustering and classification. In *ECCV*. 2012.

[22] Minh Hoai and Andrew Zisserman. Discriminative sub-categorization. In *CVPR*, 2013.

[23] Judy Hoffman, Brian Kulis, Trevor Darrell, and Kate Saenko. Discovering latent domains for multisource domain adaptation. In *ECCV*. 2012.

[24] Armand Joulin, Jean Ponce, and Francis R Bach. Efficient optimization for discriminative latent class models. In *NIPS*, 2010.

[25] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *ECCV*. 2012.

[26] Cheng-Hao Kuo and Ramakant Nevatia. Robust multi-view car detection using unsupervised sub-categorization. In *WACV*, 2009.

[27] Tian Lan, Michalis Raptis, Leonid Sigal, and Greg Mori. From subcategories to visual composites: A multi-level framework for object detection. In *ICCV*, 2013.

[28] Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In *ICML*, 2010.

[29] Tomasz Malisiewicz, Abhinav Gupta, and Alexei A Efros. Ensemble of exemplar-SVMs for object detection and beyond. In *ICCV*, 2011.

[30] Sebastian Mika. Kernel Fisher discriminants. *PhD thesis*, 2002.

[31] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, 2013.

[32] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. In *NIPS*, 2002.

[33] Li Niu, Wen Li, and Dong Xu. Visual recognition by learning from web data: A weakly supervised domain generalization approach. In *CVPR*, 2015.

[34] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*. 2010.

[35] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, 103(1):60–79, 2013.

[36] Xin-Jing Wang, Zheng Xu, Lei Zhang, Ce Liu, and Yong Rui. Towards indexing representative images on the web. In *ACM MM*, 2012.

[37] Daniel Weinland, Edmond Boyer, and Remi Ronfard. Action recognition from arbitrary views using 3d exemplars. In *ICCV*, 2007.

[38] Linli Xu, James Neufeld, Bryce Larson, and Dale Schuurmans. Maximum margin clustering. In *NIPS*, 2004.

[39] Zheng Xu, Wen Li, Li Niu, and Dong Xu. Exploiting low-rank structure from latent domains for domain generalization. In *ECCV*, 2014.

[40] Jieping Ye. Least squares linear discriminant analysis. In *ICML*, 2007.

[41] Chun-Nam John Yu and Thorsten Joachims. Learning structural SVMs with latent variables. In *ICML*, 2009.