Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Low-Rank Multi-View Learning in Matrix Completion for Multi-Label Image Classification

Meng Liu[†], Yong Luo[†§], Dacheng Tao[‡], Chao Xu[†], and Yonggang Wen[§]
[†]Key Laboratory of Machine Perception (MOE), School of EECS, PKU, Beijing 100871, China
[‡]Center for Quantum Computation and Intelligent Systems, UTS, Sydney, NSW 2007, Australia
[§]Division of Networks and Distributed Systems School of Computer Engineering, NTU, 639798, Singapore
{lemolemac, yluo180}@gmail.com, dacheng.tao@uts.edu.au, xuchao@cis.pku.edu.cn, ygwen@ntu.edu.sg

Abstract

Multi-label image classification is of significant interest due to its major role in real-world web image analysis applications such as large-scale image retrieval and browsing. Recently, matrix completion (MC) has been developed to deal with multi-label classification tasks. MC has distinct advantages, such as robustness to missing entries in the feature and label spaces and a natural ability to handle multi-label problems. However, current MC-based multi-label image classification methods only consider data represented by a single-view feature, therefore, do not precisely characterize images that contain several semantic concepts. An intuitive way to utilize multiple features taken from different views is to concatenate the different features into a long vector; however, this concatenation is prone to over-fitting and leads to high time complexity in MC-based image classification. Therefore, we present a novel multi-view learning model for MC-based image classification, called low-rank multi-view matrix completion (lrMMC), which first seeks a low-dimensional common representation of all views by utilizing the proposed low-rank multi-view learning (lrMVL) algorithm. In lrMVL, the common subspace is constrained to be low rank so that it is suitable for MC. In addition, combination weights are learned to explore complementarity between different views. An efficient solver based on fixed-point continuation (FPC) is developed for optimization, and the learned low-rank representation is then incorporated into MC-based image classification. Extensive experimentation on the challenging PASCAL VOC' 07 dataset demonstrates the superiority of lrMMC compared to other multi-label image classification approaches.

Introduction

Multi-label image classification (Ciresan et al. 2011; Ma et al. 2013), in which multiple labels are assigned to a given image, is critical for many web-based image analysis applications. For example, a multi-label image classifier might be used to annotate a newly uploaded image to facilitate retrieval by a text-based search engine.

Over the last decade, a number of multi-label algorithms have been proposed (Boutell et al. 2004; Tsoumakas and Katakis 2007; Hariharan et al. 2010). However, none of these methods are able to handle situations in which some features are missing or parts of training data labels are unknown. In addition, most algorithms are not sufficiently robust to outliers and background noise. To tackle these problems, matrix completion (MC) has recently been introduced as an alternative methodology for multi-label classification (Goldberg et al. 2010; Cabral et al. 2011; Xu, Jin, and Zhou 2013; Yu et al. 2014). In MC-based multi-label classification, a feature-by-item and label-by-item stacked matrix is first concatenated, and then the unknown feature or label entries in the concatenated matrix are completed in accordance with the rank minimization criterion. In this way, the MC-based methods infer the labels of unlabeled data, estimate the values of missing features, and de-noise the observed features.

Although MC-based algorithms have many advantages for general multi-label classification tasks, they cannot directly handle image classification problems (Luo et al. 2014a) when images are represented by multiple views. An intuitive solution is to concatenate multi-view features into a long vector, but this strategy neglects the fact that these views are extracted from different feature spaces with different statistical properties, and as a consequence this approach suffers from an over-fitting problem when the dimension of image features is much larger than the sample size. In addition, feature concatenation often leads to high time complexity in matrix completion and, sometimes, this time cost is intolerable (Cai, Candès, and Shen 2010).

Here, we present a novel multi-view learning model, termed low-rank multi-view matrix completion (lrMMC), which effectively fuses different kinds of features in matrix completion. Specifically, the different views are first projected into a low-dimensional subspace by the proposed low-rank multi-view learning (lrMVL) algorithm; the subspace is forced to be low rank to satisfy the assumption in MC. By simultaneously minimizing the reconstruction errors and the subspace rank, a common representation of all the views is learned. To further explore the complementarity of different views, a weight for the reconstruction of each view is also learned. The optimization problem is efficiently solved using an alternating algorithm, in which the sub-problem of learning the common representation is based on fixed-point continuation (FPC). The learned representation is then utilized as feature data for MC-based multi-label image classification. The main advantages of the proposed lrMMC model

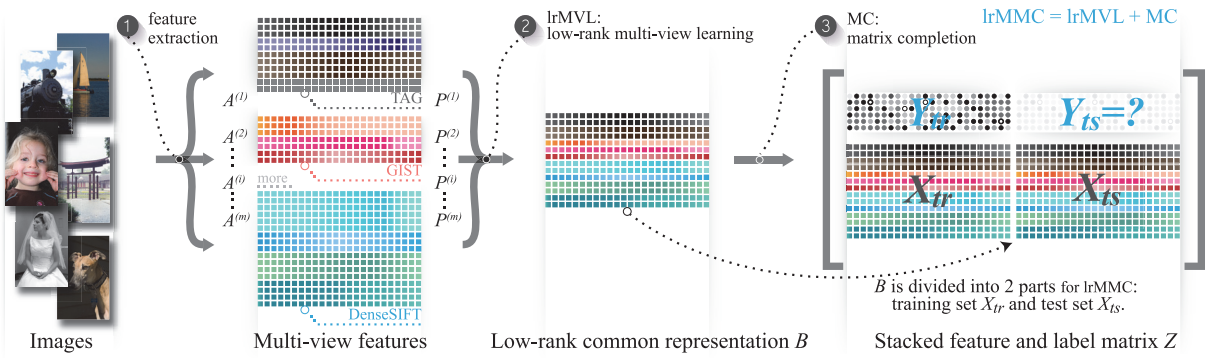

Figure 1: System diagram of the low-rank multi-view matrix completion (lrMMC) model. Features from different views (SIFT, GIST, etc.) are extracted from all images, then lrMVL seeks a low-rank common representation B from multiple views. Finally, the common representation B is divided into two feature sets, training set X_{tr} and test set X_{ts}. The label matrix of X_{tr} is known as Y_{tr}, while the label matrix of Y_{ts} needs to be predicted. Under the assumption that the stacked feature and label matrix Z is low-rank, the unknown label matrix Y_{ts} can be completed by matrix completion (MC).

are that the low-rank property of the original features is well preserved, and the complementarity of the different views is explored.

We use the very challenging PASCAL VOC' 07 dataset (Everingham et al. 2007) to evaluate the lrMVL algorithm. To the best of our knowledge, no other MC-specific multi-view algorithms exist. Therefore, lrMMC is first compared to some MC baselines in which the features from the single view (MC), a concatenation of all views (CMC), and the subspace is learned using a distributed multi-view strategy (DMC) (Long, Philip, and Zhang 2008). To further validate the proposed model, lrMMC is compared to some popular and competitive multi-view image classification approaches, namely SimpleMKL (Rakotomamonjy et al. 2008) and HierSVM (Kludas, Bruno, and Marchand-Maillet 2008) in terms of mean average precision (mAP) (Zhu 2004), Hamming loss (HL) (Schapire and Singer 2000), and ranking loss (RL) (Schapire and Singer 2000). Our experimental results demonstrate that the low-dimensional representation produced by lrMVL is suitable for MC-based image classification, and the performance of the lrMMC model even surpasses that of competitive non-dimensionality reduction multi-view learning algorithms.

Related Work

This section reviews related work on MC using rank minimization and multi-view learning.

Matrix Completion Using Rank Minimization

MC is the process of filling in unknown entries in an uncompleted matrix M. It is theoretically impossible to reconstruct an uncompleted matrix without any hypotheses about the properties of the matrix. Most MC algorithms assume that the matrix to be recovered is low rank (Nie, Huang, and Ding 2012; Ding, Shao, and Fu 2014). In other words, MC aims to find a matrix X that minimizes the difference with the known entries in M so that the rank of X reaches its minimum. This rank minimization approach has received a lot of attention and has become popular due to its success in the Netflix challenge. The rank minimization problem is NP-hard, and is therefore ineffective for most practical applications. However, (Candès and Recht 2009) found that rank(X) and its convex envelope, the nuclear norm $\|X\|_*$, have the same unique solution, and they proved that only a limited number of samples are needed to recover a low-rank matrix, where $\|X\|_*$ is the sum of singular values of X (Fazel 2002). Under this relaxation, several MC methods have been proposed (Candès and Recht 2009; Cai, Candès, and Shen 2010; Keshavan, Montanari, and Oh 2010; Ma, Goldfarb, and Chen 2011). In addition, interior point methods, such as the semi-definite programming (SDP) algorithm, have been successfully applied to this convex optimization problem; however, the off-the-shelf interior point methods can only handle small matrices, and the singular value thresholding (SVT) algorithm was developed to overcome this difficulty (Cai, Candès, and Shen 2010). Although SVT effectively handles large matrices, it may fail because the rank of the uncompleted matrix is very low. The more robust fixed-point continuation (FPC) algorithm (Ma, Goldfarb, and Chen 2011) was therefore proposed, which probably converges to the optimal solution for the unconstrained problem under certain conditions. Empirically, FPC has much better recoverability than SDP, SVT, and other algorithms on the MC problem.

Multi-view Learning

Multi-view learning is an active research topic (Xia et al. 2014; Luo et al. 2014b; Xu, Tao, and Xu 2014). In this paper, *multiple views* mean various descriptions of a given sample. Many methods have been proposed for multi-view classification (Zien and Ong 2007), retrieval (Kludas, Bruno, and Marchand-Maillet 2008), and clustering (Bickel and Scheffer 2004). Depending on the level of fusion being carried out, the multi-view classification methods can be grouped into two major categories: feature-level fusion and classifier-level fusion. This paper will focus on the former. A direct strategy for feature-level fusion is to concatenate the differ-

ent kinds of features into a long vector. However, this often leads to the curse of dimensionality problem and renders it impractical. As a result, many sophisticated techniques have been developed to overcome this problem, and we refer the interested reader to (Xu, Tao, and Xu 2013) for a literature survey of multi-view learning.

Low-rank Multi-view Matrix Completion

In this section, we introduce the low-rank multi-view matrix completion (lrMMC) model (see overview in Figure 1). lrMMC has two main parts: the novel low-rank multi-view learning (lrMVL) method and MC-based image classification. The former seeks a low-rank common representation for all views, which is then embedded into the latter. Specifically, we first extract different kinds of features from the images in the dataset, such as SIFT (Lowe 2004) and GIST (Oliva and Torralba 2001). The obtained feature matrix is denoted as $X^{(i)} \in \mathbb{R}^{d_i \times n}, i = 1, 2, \dots m$, where n and m are the number of samples in the dataset and feature views, respectively, and d_i is the dimensionality of the i'th view. Alternatively, $X^{(i)}$ can be preprocessed using dimensionality reduction strategies to derive the pattern matrix $A^{(i)}$, and in this paper we assume $A^{(i)} = X^{(i)}$. The different $A^{(i)}$ is then projected into a low-dimensional common representation B by utilizing the mapping matrix $P^{(i)}$. Each $A^{(i)}$ can be reconstructed using B and $P^{(i)}$. By simultaneously minimizing the total reconstruction errors of all the views and the rank of B, both optimal B and each $P^{(i)}$ can be learned. Finally, the representation B is divided into training and test sets. Labels for the training set are known and labels for the test set can be completed by MC. The details of the technique are given below.

Problem Formulation of lrMVL

Given a multi-view dataset consisting of n samples with m views that are denoted as a set of feature matrices $\mathcal{X} = \{X^{(i)} \in \mathbb{R}^{d_i \times n}\}_{i=1}^m$, we represent their pattern matrices as $\mathcal{A} = \{A^{(i)} \in \mathbb{R}^{k_i \times n}\}_{i=1}^m$. To find a low-dimensional common representation B, the traditional distributed strategy (Long, Philip, and Zhang 2008) is to optimize the following problem:

$$\min_{BB^T=I, \{P^{(i)}\}} \sum_{i=1}^{m} \|P^{(i)}B - A^{(i)}\|_F^2, \tag{1}$$

where $\mathcal{P} = \{P^{(i)} \in \mathbb{R}^{k_i \times k}\}_{i=1}^m$ is a set of mapping matrices. The global optimal solution of this formulation is given by performing eigenvalue decomposition of the matrix $A^T A$, where $A = [A^{(1)}; \cdots; A^{(m)}]$ is a concatenation of the pattern matrices. Although simple and efficient, this multi-view strategy is unsuitable for MC and usually performs less well than the best single-view strategy, as shown in our follow-up experiments. The main reason for this underperformance is that B is forced to be orthogonal, and thus the low-rank assumption in MC is violated. Therefore, we propose to constrain B to be low rank. To further explore the complementarity of different views, we also learn the nonnegative weight θ_i for the i'th view. Thus, the optimization problem becomes

$$\begin{aligned} \min_{B, \{P^{(i)}, \theta_i\}} & \mu\|B\|_* + \sum_{i=1}^{m} \theta_i \|P^{(i)}B - A^{(i)}\|_F^2 + \frac{\gamma}{2}\|\theta\|_2^2 \\ \text{s.t. } & \theta_i \geq 0, \sum \theta_i = 1, i = 1, \dots, m, \end{aligned} \tag{2}$$

where $A^{(i)} \in \mathbb{R}^{k_i \times n}$, $B \in \mathbb{R}^{k \times n}$ and $P^{(i)} \in \mathbb{R}^{k_i \times k}$. Both $\mu \geq 0$ and $\gamma \geq 0$ are trade-off parameters. Since the Frobenius norm is a separable distance function (Long, Philip, and Zhang 2008), by letting $A = [\sqrt{(\theta_1)}A^{(1)}; \dots; \sqrt{(\theta_m)}A^{(m)}] \in \mathbb{R}^{(\sum_i^m k_i) \times n}$ and $P = [\sqrt{(\theta_1)}P^{(1)}; \dots; \sqrt{(\theta_m)}P^{(m)}] \in \mathbb{R}^{(\sum_i^m k_i) \times k}$, we can rewrite (2) in a compact form:

$$\begin{aligned} \min_{B,P,\theta} & \mu\|B\|_* + \|PB - A\|_F^2 + \frac{\gamma}{2}\|\theta\|_2^2 \\ \text{s.t. } & \theta_i \geq 0, \sum \theta_i = 1, i = 1, \dots, m. \end{aligned} \tag{3}$$

Optimization Algorithm of lrMVL

Three variables B, P and θ need to be optimized in (3), and there is currently no direct way to find the global optimal solution. Therefore, we have developed an iterative algorithm to find the optimal solution, which alternately updates these variables and efficiently solves each sub-problem.

Update for P. By initializing B with a random matrix and $\theta_i = \frac{1}{m}, i = 1, \dots, m$, we can rewrite (3) with respect to P as

$$\min_P \|PB - A\|_F^2. \tag{4}$$

Let $L(P)$ denote the objective function in Equation (4). By taking the derivative of $L(P)$ with respect to P and setting it to zero, we obtain

$$P = AB^T(BB^T)^{-1}. \tag{5}$$

However, BB^T is a semi-definite matrix since B is low rank. Thus BB^T is irreversible, and this problem can be dealt with the regularization or pseudoinverse tricks. We choose the regularization strategy, and then the solution becomes

$$P^* = AB^T(BB^T + \eta I)^{-1}, \tag{6}$$

where η is a small regularization parameter and I is the identity matrix. Each ${P^{(i)}}^*$ can be obtained by dividing the i'th block of P^* with $\sqrt{(\theta_i)}$.

Update for B. With the obtained P^* and fixed θ, the subproblem for optimizing (3) with respect to B is given by

$$\min_B \mu\|B\|_* + \|PB - A\|_F^2. \tag{7}$$

We introduce the FPC algorithm (Ma, Goldfarb, and Chen 2011) to obtain the solution of B. The FPC algorithm is comprised of a series of gradient updates $h(\cdot) = I(\cdot) - \tau g(\cdot)$ and shrinkage, where τ is the step size. The gradient descent is given by

$$h(B) = B - 2\tau P^T(PB - A), \tag{8}$$

where B in this inner iteration is initialized with the rank-1 approximation of B in the last iteration of the outer iteration.

The shrinkage operator $S_v(\cdot) = max(0, \cdot - v)$ is applied to the singular values of the result of (8) to ensure that the rank of B is minimized. The gradient update and shrinkage steps are alternated until convergence, which is guaranteed since $h(B)$ is contractive with a properly chosen step size (Ma, Goldfarb, and Chen 2011).

Theorem 1. *Provided the step size* $\tau \in (0, \frac{1}{\lambda_{max}(P^T P)})$, *then* $h(B)$ *is a contraction.*

Proof. Assume that the step size $\tau \in (0, 1/\lambda_{max}(P^T P))$, then $-1 < \lambda_i(I - 2\tau P^T P) \leq 1$, where $\lambda_i(I - 2\tau P^T P)$ is the i'th eigenvalue of $I - 2\tau P^T P$. Hence,

$$\begin{aligned} \|h(B_1) - h(B_2)\|_F &= \|(B_1 - B_2)(I - 2\tau P^T P)\|_F \\ &\leq \|B_1 - B_2\|_F \|I - 2\tau P^T P\|_2 \\ &\leq \|B_1 - B_2\|_F \end{aligned} \tag{9}$$

then $h(B)$ is contractive. □

Update for θ. With the obtained P^* and B^*, we can rewrite (3) with respect to θ as:

$$\begin{aligned} &\min_{\theta} \theta^T q + \frac{\gamma}{2}\|\theta\|_2^2 \\ &\text{s.t. } \theta_i \geq 0, \sum \theta_i = 1, i = 1, \ldots, m, \end{aligned} \tag{10}$$

where $q = [q_1, \ldots, q_m]^T$ with each $q_i = \|P^{(i)^*} B^* - A^{(i)}\|_F^2$. We adopt the coordinate descent algorithm to solve (10). Therefore, in each iteration of the descent procedure, only two elements θ_i and θ_j are selected to be updated; the others are fixed. By using the Lagrangian of problem (10) and considering the sum to one constraint, we obtain the following updating rule:

$$\begin{cases} \theta_i^* = \dfrac{\gamma(\theta_i + \theta_j) + (q_j - q_i)}{2\gamma} \\ \theta_j^* = \theta_i + \theta_j - \theta_i^* \end{cases}. \tag{11}$$

The obtained θ_i^* or θ_j^* may violate the constraint $\theta_i \geq 0$. Thus we set $\theta_i^* = 0$ if $\gamma(\theta_i + \theta_j) + (q_j - q_i) < 0$, and vice versa for θ_j^*.

The learning procedure of the low-rank multi-view learning method is summarized in Algorithm 1. The stopping criterion for the algorithm is the difference between the objective values of two consecutive steps. The lrMVL algorithm is guaranteed to converge to the local optimum of (3), since each of the aforementioned sub-problems is convex.

Transduction with Matrix Completion

After obtaining the low-rank common representation B from multiple views, it is embedded into matrix completion for image classification, thereby achieving the low-rank multi-view matrix completion (lrMMC) model. Specifically, we divide B into two parts: training set B_{tr} and test set B_{ts}. The label matrix Y_{tr} of B_{tr} is already known, and the label matrix Y_{ts} needs to be predicted. By replacing the feature matrix X in (Cabral et al. 2011) with B, we have

$$Z = \begin{bmatrix} Y \\ B \\ \mathbf{1}^T \end{bmatrix} = \begin{bmatrix} Y_{tr} & Y_{ts} \\ B_{tr} & B_{ts} \\ \multicolumn{2}{c}{\mathbf{1}^T} \end{bmatrix}, \tag{12}$$

Algorithm 1 The learning procedure of low-rank multi-view learning (lrMVL) method.

Input: Multi-view feature matrices $A = [A^{(1)}; \ldots; A^{(m)}]$ and the dimension k of B;
Output: low-rank common representation B;

```
initialize B with a random matrix and θ_i = 1/m, i = 1, ..., m;
while globalRelError > δ do
    update P using Equation (6).
    for μ = μ_1 > μ_2 > ... > μ_s = μ̄ do
        while fpcRelError > ε do
            Gradient: M = B − 2τP^T(PB − A);
            update B with shrinking:
            M = UΣV^T; B = US_{τμ}(Σ)V^T;
        end while
    end for
    update θ using Equation (11).
end while
```

where $B_{tr} \in \mathbb{R}^{k \times n_{tr}}$, $B_{ts} \in \mathbb{R}^{k \times n_{ts}}$, $Y_{tr} \in \mathbb{R}^{c \times n_{tr}}$ and $Y_{ts} \in \mathbb{R}^{c \times n_{ts}}$. Y_{ts} can be obtained by solving MC-1 problem as presented in (Goldberg et al. 2010, Cabral et al. 2011), which predicts unknown labels from the stacked feature and label matrix.

Complexity Analysis

The complexity of lrMMC has two parts: the first is for the developed lrMVL and the other is for MC-1 (Goldberg et al. 2010; Cabral et al. 2011). For lrMVL (see Algorithm 1), the calculation of P (step 3) consists of some matrix multiplication and inversion. This leads to a complexity of $O(n^2(2k + \sum_{i=1}^m k_i + n))$. From step 6 to step 8, there is a singular value decomposition of M, as well as some matrix multiplications to obtain M and B. Usually, the SVD computation cost of a $(k \times n)$ matrix is taken as $O(nk^2 + n^3)$, and thus the complexity of step 6 to step 8 is given by $O(n^3 + k^3 + n^2 k + nk^2)$. Suppose the number of inner iterations (step 5 to step 9) is t, then the time complexity of updating B is $O(st(n^3 + k^3 + n^2 k + nk^2))$, where s is the number of elements in the μ sequence. The time cost for calculating θ (step 11) can be omitted since it is usually much smaller than the update of P and B.

It can be seen that the time cost of lrMVL is dominated by the calculation of B, and thus the time complexity of lrMVL is $O(Tst(n^3 + k^3 + n^2 k + nk^2))$, where T is the number of outer iterations (step 2 to step 12). In practice, t is frequently smaller than 5; s and T are smaller than 10. Therefore, the proposed lrMVL algorithm is quite efficient. Similarly, the time complexity for MC-1 is $t's'(O(n^3 + (k+c)^3 + n^2(k + c) + n(k + c)^2))$, where c is the number of classes, t' is the number of iterations, and s' is the number of elements in the μ sequence for MC-1. The time complexity of the lrMMC model is a sum of these two parts.

Experiments

Here, we present an experimental evaluation of the performance of lrMMC on PASCAL VOC' 07 dataset. We first

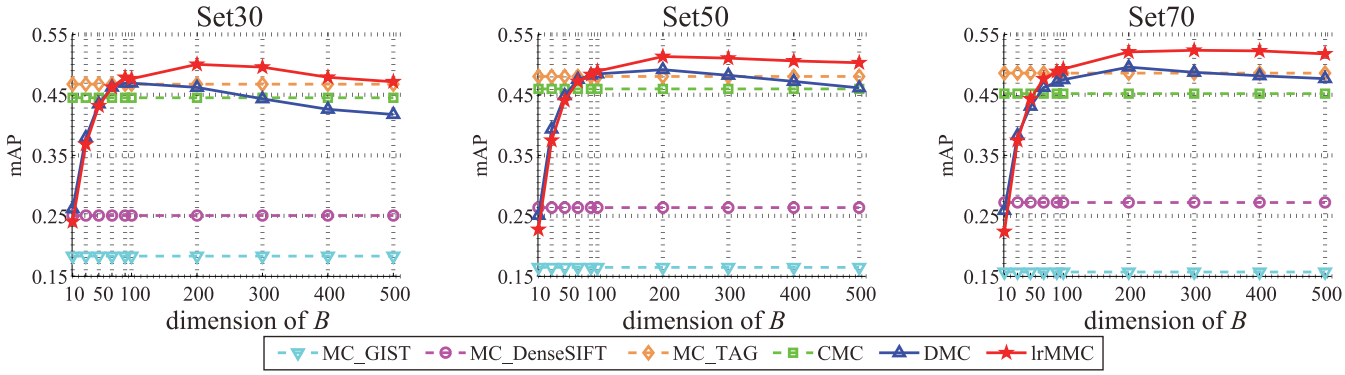

Figure 2: Classification results of the compared MC-based methods in terms of mAP. The dimensions are fixed for the single-view method and CMC, and thus their mAP scores remain the same. The solid curves show how the dimension of B affects DMC and lrMMC. The standard deviations are around 0.005, and thus are invisible.

investigate the performance of lrMMC with respect to the dimension of the common representation by comparing it with some MC-based strategies. Then the average performances of the compared methods using their best dimensions are presented. Finally, to further verify the effectiveness of the proposed model, we compare lrMMC with some competitive non-dimensionality reduction multi-view algorithms, namely SimpleMKL (Rakotomamonjy et al. 2008) and HierSVM (Kludas, Bruno, and Marchand-Maillet 2008), in terms of mean average precision (mAP) (Zhu 2004), Hamming loss (HL) (Schapire and Singer 2000), and ranking loss (RL) (Schapire and Singer 2000). Before all of these evaluations, we present the dataset and features we used, as well as our experimental settings.

Dataset, Features, and the Evaluation Criteria

The PASCAL VOC' 07 (VOC for short) dataset (Everingham et al. 2007) consists of 9,963 images in 20 object classes. We use the features from (Guillaumin, Verbeek, and Schmid 2010), which provides several different image representations and tags. For this paper, we chose three representative feature views: the local SIFT (Lowe 2004), global GIST (Oliva and Torralba 2001), and TAG, which is the textual information. The dimensions of SIFT, GIST, and TAG are 1,000, 512, and 804, respectively.

Three popular evaluation criteria for multi-label classification are used: average precision (AP), Hamming loss (HL), and ranking loss (RL). AP is the ranking performance computed for each label, and the mean value for all the labels, i.e., mAP, is reported. HL and RL are used to evaluate the label set predictions for each instance. All three criteria are widely used to evaluate the performance of multi-label classification (Zhang and Zhou 2007).

The positive and negative samples are quite unbalanced in the VOC dataset. Thus, for each object class, 30, 50, and 70 positive samples, and the same number of negative samples, are randomly selected to form three labeled sets of different size: Set30, Set50, and Set70. The standard VOC test set (Everingham et al. 2007) is used for testing, and 20% of the 4,952 test images are randomly selected for validation. The best-performing parameters on the validation set are used for the final test. All the experiments are run ten times, and both the mean and standard deviation are reported.

Comparisons with the MC-based Methods

The experimental setup of the compared methods is as follows:

MC: uses the single-view features, where the methods for different views are denoted as MC_GIST, MC_DenseSIFT, and MC_TAG, respectively. The MC-based transduction (also multi-label classification here) is performed by adopting the MC-1 algorithm, as presented in (Goldberg et al. 2010; Cabral et al. 2011). The candidate set for choosing λ is $\{10^i | i = -4, \ldots, 2\}$. The parameter μ is initialized as $\mu_0 = 0.25\sigma_1$, where σ_1 is the largest singular value of Z_0, and decreases with a factor of 0.25 in the continuation steps until $\mu = 10^{-12}$.

CMC: uses the concatenation of the normalized features of all views in MC. Parameters μ and λ are the same as in MC.

DMC: seeks the common representation B using the distributed strategy, as presented in (Long, Philip, and Zhang 2008). The dimension k of B is chosen from the set $\{10, 30, 50, 70, 90, 100, 200, 300, 400, 500\}$.

lrMMC: induces the common representation B by the proposed lrMVL algorithm. The parameter μ in lrMVL is determined as in MC-1, and the parameter γ is tuned over the set $\{10^i | i = -4, \ldots, 3\}$. The algorithm stops the iteration when the difference of the objective function is smaller than 10^{-3}.

The performance of DMC and lrMMC in terms of mAP with respect to the dimension k of the common representation, as well as the other methods with fixed dimensions, are shown in Figure 2. It can be seen that: 1) the TAG features perform the best, followed by SIFT and GIST. The concatenated method (CMC) is inferior to the best single-view method due to over-fitting, illustrating the benefit of the proposed multi-view learning algorithm; 2) when k increases, the performance of both DMC and lrMMC improves sharply at first but then declines. This is because a rapidly increasing

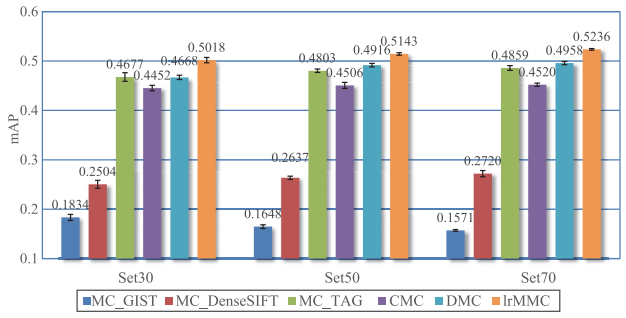

Figure 3: Performances of the compared MC-based methods on the VOC dataset with labeled sets of different size.

amount of information contained in the original features is first explored, but, when the dimension is large, the learned variables in MC may be not reliable due to the limited number of labeled samples; 3) the optimal dimension of lrMMC is larger than DMC, and the superiority of lrMMC over DMC becomes larger with increasing k. This is because the learned B tends not to be low rank when the dimension is too small; 4) DMC sometimes performs worse than MC_TAG, while the proposed lrMMC is consistently better than the other methods when $k \geq 100$. This benefit is significant at its optimal dimension.

The mAP values of MC, CMC, DMC, and lrMMC under their optimal parameters are shown in Figure 3. It is clear that lrMMC achieves the best performance with all settings. In particular, there is a 3.5%, 2.27%, and 2.78% improvement over DMC for Set30, Set50, and Set70, respectively.

Comparison with Non-dimensionality Reduction Multi-view Algorithms

The final experiment compares lrMMC with two non-dimensionality reduction multi-view algorithms, HierSVM and SimpleMKL. Their details are as follows:

HierSVM: learns SVM classifiers for each view separately, and then fuses the results by using an additional SVM classifier. This algorithm is named after its hierarchical SVM structure. The values of the tradeoff parameter C for each SVM classifier are optimized on the set $\{10^i | i = -1, \ldots, 6\}$.

SimpleMKL: deals with the multiple kernel learning task by solving a standard SVM optimization problem. SimpleMKL first constructs a kernel for each view and then learns a linear combination of the different kernels and a classifier based on the combined kernel. The penalty factor C of SimpleMKL is tuned on the same set as in HierSVM.

The performances of these methods and lrMMC on the VOC dataset are reported in Table 1. It can be seen that: 1) the mAP scores of lrMMC are consistently higher than HierSVM and SimpleMKL for all three different labeled sets; 2) under the HL criterion, SimpleMKL only performs well on Set30, while lrMMC performs better when more labeled samples are available; the HL performance of HierSVM is always unsatisfactory; 3) in contrast, HierSVM is promising in terms of RL, especially on Set30, while lrMMC outperforms it with more labeled samples; the performance of SimpleMKL is poor. In summary, lrMMC is superior to HierSVM and SimpleMKL in most cases, and the performance of HierSVM and SimpleMKL varies significantly under different criteria. It should be noted that both HierSVM and SimpleMKL are non-dimensionality reduction methods and SimpleMKL is carried out in the kernel space, whereas lrMMC is a dimensionality reduction approach aided by a linear MC technique.

Methods	Set30	Set50	Set70
	mAP ↑ versus #labeled samples		
HierSVM	0.401 ± 0.004	0.421 ± 0.005	0.432 ± 0.006
SimpleMKL	0.497 ± 0.006	0.514 ± 0.006	0.519 ± 0.001
lrMMC	$\mathbf{0.502 \pm 0.005}$	$\mathbf{0.515 \pm 0.003}$	$\mathbf{0.524 \pm 0.002}$
	HL ↓ versus #labeled samples		
HierSVM	0.067 ± 0.001	0.066 ± 0.001	0.066 ± 0.001
SimpleMKL	$\mathbf{0.059 \pm 0.001}$	0.056 ± 0.001	0.055 ± 0.001
lrMMC	0.069 ± 0.001	$\mathbf{0.052 \pm 0.001}$	$\mathbf{0.050 \pm 0.000}$
	RL ↓ versus #labeled samples		
HierSVM	$\mathbf{0.124 \pm 0.002}$	0.116 ± 0.001	0.113 ± 0.001
SimpleMKL	0.173 ± 0.009	0.167 ± 0.012	0.158 ± 0.004
lrMMC	0.145 ± 0.005	$\mathbf{0.116 \pm 0.001}$	$\mathbf{0.110 \pm 0.001}$

Table 1: A comparison of non-dimensionality reduction multi-view algorithms on the VOC dataset using different evaluation criteria. ↑ indicates *the larger the better*; ↓ indicates *the smaller the better*.

Conclusions

In this paper, we present a low-rank multi-view matrix completion (lrMMC) model for multi-label image classification. We first developed a low-rank multi-view learning (lrMVL) algorithm to seek a common low-dimensional representation from features of multiple views; the learned representation is then embedded into MC for classification. Our model has the advantages of being able to preserve the low-rank property of the natural features and explore the complementary properties of different views. We also developed an efficient solver for optimization based on fixed-point continuation (FPC). Experiments on the challenging PASCAL VOC '07 dataset show that our proposed lrMMC not only outperforms single-view or multi-view MC-based strategies but also other competitive multi-view approaches.

Acknowledgements

This research was partially supported by grants from NBRPC 2011CB302400, NSFC 61375026, 2015BAF15B00, JCYJ 20120614152136201, and Australian Research Council Projects FT-130101457 and DP-140102164.

References

Bickel, S., and Scheffer, T. 2004. Multi-view clustering. In *IEEE International Conference on Data Mining*, 19–26.

Boutell, M. R.; Luo, J.; Shen, X.; and Brown, C. M. 2004. Learning multi-label scene classification. *Pattern Recognition* 37(9):1757–1771.

Cabral, R. S.; De la Torre, F.; Costeira, J. P.; and Bernardino, A. 2011. Matrix completion for multi-label image classification. In *Advances in Neural Information Processing Systems*, 190–198.

Cai, J.; Candès, E. J.; and Shen, Z. 2010. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization* 20(4):1956–1982.

Candès, E. J., and Recht, B. 2009. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics* 9(6):717–772.

Ciresan, D. C.; Meier, U.; Masci, J.; Maria Gambardella, L.; and Schmidhuber, J. 2011. Flexible, high performance convolutional neural networks for image classification. In *International Joint Conference on Artificial Intelligence*, 1237–1242.

Ding, Z.; Shao, M.; and Fu, Y. 2014. Latent low-rank transfer subspace learning for missing modality recognition. In *AAAI Conference on Artificial Intelligence*, 1192–1198.

Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2007. The PASCAL visual object classes challenge 2007 (voc2007) results.

Fazel, M. 2002. *Matrix Rank Minimization with Applications*. Ph.D. Dissertation, Stanford University.

Goldberg, A. B.; Zhu, X.; Recht, B.; Xu, J.; and Nowak, R. D. 2010. Transduction with matrix completion: Three birds with one stone. In *Advances in Neural Information Processing Systems*, 757–765.

Guillaumin, M.; Verbeek, J.; and Schmid, C. 2010. Multimodal semi-supervised learning for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 902–909.

Hariharan, B.; Zelnik, M. L.; Varma, M.; and Vishwanathan, S. 2010. Large scale max-margin multi-label classification with priors. In *International Conference on Machine Learning*, 423–430.

Keshavan, R. H.; Montanari, A.; and Oh, S. 2010. Matrix completion from a few entries. *IEEE Transactions on Information Theory* 56(6):2980–2998.

Kludas, J.; Bruno, E.; and Marchand-Maillet, S. 2008. Information fusion in multimedia information retrieval. In *Adaptive Multimedia Retrieval: Retrieval, User, and Semantics*. Springer. 147–159.

Long, B.; Philip, S. Y.; and Zhang, Z. M. 2008. A general model for multiple view unsupervised learning. In *SIAM International Conference on Data Mining*, 822–833.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2):91–110.

Luo, Y.; Liu, T.; Tao, D.; and Xu, C. 2014a. Decomposition-based transfer distance metric learning for image classification. *IEEE Transactions on Image Processing* 23(9):3789–3801.

Luo, Y.; Tang, J.; Yan, J.; Xu, C.; and Chen, Z. 2014b. Pre-trained multi-view word embedding using two-side neural network. In *AAAI Conference on Artificial Intelligence*, 1982–1988.

Ma, Z.; Yang, Y.; Nie, F.; and Nicu, S. 2013. Thinking of images as what they are: Compound matrix regression for image classification. In *International Joint Conference on Artificial Intelligence*, 1530–1536.

Ma, S.; Goldfarb, D.; and Chen, L. 2011. Fixed point and bregman iterative methods for matrix rank minimization. *Mathematical Programming* 128(1-2):321–353.

Nie, F.; Huang, H.; and Ding, C. H. 2012. Low-rank matrix recovery via efficient schatten p-norm minimization. In *AAAI Conference on Artificial Intelligence*, 655–661.

Oliva, A., and Torralba, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision* 42(3):145–175.

Rakotomamonjy, A.; Bach, F.; Canu, S.; Grandvalet, Y.; et al. 2008. SimpleMKL. *Journal of Machine Learning Research* 9:2491–2521.

Schapire, R. E., and Singer, Y. 2000. Boostexter: A boosting-based system for text categorization. *Machine Learning* 39(2-3):135–168.

Tsoumakas, G., and Katakis, I. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining* 3(3):1–13.

Xia, R.; Pan, Y.; Du, L.; and Yin, J. 2014. Robust multi-view spectral clustering via low-rank and sparse decomposition. In *AAAI Conference on Artificial Intelligence*, 2149–2155.

Xu, M.; Jin, R.; and Zhou, Z. 2013. Speedup matrix completion with side information: Application to multi-label learning. In *Advances in Neural Information Processing Systems*, 2301–2309.

Xu, C.; Tao, D.; and Xu, C. 2013. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*.

Xu, C.; Tao, D.; and Xu, C. 2014. Large-margin multi-view information bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(8):1559–1572.

Yu, H.; Jain, P.; Kar, P.; and Dhillon, I. 2014. Large-scale multi-label learning with missing labels. In *International Conference on Machine Learning*, 593–601.

Zhang, M., and Zhou, Z. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition* 40(7):2038–2048.

Zhu, M. 2004. Recall, precision, and average precision. Technical report, Department of Statistics and Actuarial Science, University of Waterloo. Working Paper 2004-09.

Zien, A., and Ong, C. S. 2007. Multiclass multiple kernel learning. In *International Conference on Machine Learning*, 1191–1198.