Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence

Vector-Valued Multi-View Semi-Supervised Learning for Multi-Label Image Classification

Yong Luo[†], Dacheng Tao[‡], Chang Xu[†], Dongchen Li[†], and Chao Xu[†]
[†]Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, China
[‡]Faculty of Engineering and Information Technology, University of Technology, Sydney, Sydney, Australia

Abstract

Images are usually associated with multiple labels and comprised of multiple views, due to each image containing several objects (e.g. a pedestrian, bicycle and tree) and multiple visual features (e.g. color, texture and shape). Currently available tools tend to use either labels or features for classification, but both are necessary to describe the image properly. There have been recent successes in using vector-valued functions, which construct matrix-valued kernels, to explore the multi-label structure in the output space. This has motivated us to develop multi-view vector-valued manifold regularization (MV^{3}MR) in order to integrate multiple features. MV^{3}MR exploits the complementary properties of different features, and discovers the intrinsic local geometry of the compact support shared by different features, under the theme of manifold regularization. We validate the effectiveness of the proposed MV^{3}MR methodology for image classification by conducting extensive experiments on two challenge datasets, PASCAL VOC' 07 and MIR Flickr.

Introduction

The contents of a natural image can be usefully summarized by several keywords (or labels). In order to perform image classification directly using binary classification methods (Boutell et al. 2004; Guillaumin, Verbeek, and Schmid 2010), it is necessary to assume that the labels are independent, even though frequently labels appearing in the image are related to each other. Examples are given in Figure 1, in which the left image shows a person riding a bicycle, the middle image shows sea, which usually co-occurs with sky and water, and the right image shows a dog, which is associated with animal. This multi-label dependency (Zhang 2011; Zhou et al. 2012) makes this type of image classification (Boutell et al. 2004; Luo et al. 2013) intrinsically different from simple binary classification.

Moreover, different labels cannot be fully characterized by a single feature. Color information (e.g. color histogram (Van De Weijer and Schmid 2006)), shape cue (encoded in SIFT (Lowe 2004)) and global structure (e.g. GIST (Oliva and Torralba 2001)) can effectively represent

Figure 1: Multi-label image examples from PASCAL VOC' 07 and MIR Flickr datasets.

natural objects (e.g. sky, cloud and plant life), man-made objects (e.g. airplane, car, and TV-monitor), and scenes (e.g. seaside and indoor). However, these parameters cannot simultaneously illustrate all of these concepts effectively. Each type of visual feature encodes a particular property of the image, and characterizes a particular concept (label). This multi-view nature (Sindhwani, Niyogi, and Belkin 2005) distinguishes image classification from single-view tasks, such as texture segmentation and face recognition.

Recently, the vector-valued function (Micchelli and Pontil 2005) has been used to resolve multi-label classification problems (Minh and Sindhwani 2011), and shown to be effective in semantic scene annotation. This method naturally incorporates label-dependencies into the classification model, first by computing the graph Laplacian (Belkin, Niyogi, and Sindhwani 2006) of the output similarity graph, and then using this graph to construct a matrix-valued kernel. This model is superior to most of the existing multi-label learning methods (Chen et al. 2008; Hariharan et al. 2010; Sun, Ji, and Ye 2011), because it naturally considers label correlations and efficiently outputs all the predicted labels at the same time.

Although the vector-valued function is effective for general multi-label classification tasks, it cannot directly handle image classification problems that include images represented by multi-view features. A popular solution is to concatenate all the features into a long vector. This concatenation strategy not only ignores the physical interpretations of different features, but it also addresses the over-fitting problem given limited training samples.

Here we introduce multiple kernel learning to the vector-valued function, and present a multi-view vector-valued manifold regularization (MV^{3}MR) algorithm for handling multi-view features in multi-label image classification. MV^{3}MR associates each view with a particular kernel, assigns a higher weight to the view/kernel carrying more dis-

criminative information, and explores the complementary nature of different views.

In particular, MV^{3}MR combines multi-view information in a large number of unlabeled images in order to discover the intrinsic geometry embedded in the high dimensional ambient space of the compact support of the marginal distribution. The local geometry, approximated by the adjacency graphs induced from multiple kernels of all the corresponding views, is more reliable than that approximated by the adjacency graph induced from a particular kernel of any corresponding view. In this way, MV3MR essentially improves the vector-valued function for multi-label image classification. We carefully designed the MV3MR algorithm so that it determines the set of kernel weights in the learning process of the vector-valued function.

We thoroughly evaluate the proposed MV^{3}MR algorithm on two challenge datasets: PASCAL VOC' 07 (Everingham et al.) and MIR Flickr (Huiskes and Lew 2008). We compare it with a popular MKL algorithm (Rakotomamonjy et al. 2008) and a recently proposed MKL method (Kloft et al. 2011). We also compare MV^{3}MR with competitive multi-label learning algorithms for image classification, namely multi-label compressed sensing (Hsu et al. 2009), canonical correlation analysis (Sun, Ji, and Ye 2011), and vector-valued manifold regularization (Minh and Sindhwani 2011). These algorithms are compared in terms of mean average precision (mAP). The experimental results demonstrate the effectiveness of MV^{3}MR.

Manifold Regularization and Vector-valued Generalization

First, we briefly introduce the manifold regularization framework (Belkin, Niyogi, and Sindhwani 2006) and its vector-valued generalization (Minh and Sindhwani 2011). Given a set of l labeled examples $\mathcal{D}_l = (x_i, y_i)_{i=1}^{l}$ and a relatively large set of u unlabeled examples $\mathcal{D}_u = (x_i)_{i=l+1}^{N=l+u}$, we consider a non-parametric estimation of a vector-valued function $f : \mathcal{X} \mapsto \mathcal{Y}$, where $\mathcal{Y} = \mathbb{R}^n$ and n is the number of labels. This setting includes $\mathcal{Y} = \mathbb{R}$ as a special case for regression and classification.

Manifold Regularization

Manifold learning has attracted much attention in artificial intelligence (AI) recently (Vu, Carey, and Mahadevan 2012; Suzuki et al. 2012). In manifold regularization, the data manifold is characterized by a nearest neighbor graph $\mathcal{W}$, which explores the geometric structure of the compact support of the marginal distribution. The Laplacian $\mathcal{L}$ of $\mathcal{W}$ and the prediction $\mathbf{f} = [f(x_1), \ldots, f(x_N)]$ are then formulated as a smoothness constraint $\|f\|_I^2 = \mathbf{f}^T \mathcal{L} \mathbf{f}$, where $\mathcal{L} = \mathcal{D} - \mathcal{W}$ and the diagonal matrix $\mathcal{D}$ is given by $\mathcal{D}_{ii} = \sum_{j=1}^{N} \mathcal{W}_{ij}$. The manifold regularization framework minimizes the regularized loss

$$\underset{f \in \mathcal{H}_k}{\operatorname{argmin}} \frac{1}{l} \sum_{i=1}^{l} L(f, x_i, y_i) + \gamma_A \|f\|_k^2 + \gamma_I \|f\|_I^2, \qquad (1)$$

where L is a predefined loss function, k is the standard scalar-valued kernel (i.e. $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$), and $\mathcal{H}_k$ is the associated reproducing kernel Hilbert space (RKHS). Here, γ_A and γ_I are trade-off parameters to control the complexities of f in the ambient space and the compact support of the marginal distribution. The representer theorem (Schölkopf and Smola 2002) ensures that the solution of problem (1) takes the form $f^*(x) = \sum_{i=1}^{N} \alpha_i k(x, x_i)$. Since a pair of close samples means that the corresponding conditional distributions are similar, the manifold regularization $\|f\|_I^2$ helps the function learning.

Vector-Valued Manifold Regularization

In the vector-valued RKHS, where a kernel function K is defined, and the corresponding $\mathcal{Y}$-valued RKHS is denoted by $\mathcal{H}_K$, the optimization problem of the vector-valued manifold regularization (VVMR) (Minh and Sindhwani 2011) is given by

$$\underset{f \in \mathcal{H}_k}{\operatorname{argmin}} \frac{1}{l} \sum_{i=1}^{l} L(f, x_i, y_i) + \gamma_A \|f\|_k^2 + \gamma_I \langle \mathbf{f}, \mathcal{M}\mathbf{f} \rangle_{\mathcal{Y}^{u+l}}, \quad (2)$$

where $\mathcal{Y}^{u+l}$ is the $u+l$-direct product of $\mathcal{Y}$ and the function prediction $\mathbf{f} = (f(x_1), \ldots, f(x_{u+l})) \in \mathcal{Y}^{u+l}$. The matrix $\mathcal{M}$ is a symmetric positive operator that satisfies $\langle \mathbf{y}, \mathcal{M}\mathbf{y} \rangle \geq 0$ for all $\mathbf{y} \in \mathcal{Y}^{u+l}$ and is chosen to be $\mathcal{L} \otimes I_n$. Here, $\mathcal{L}$ is the graph Laplacian, I_n is the $n \times n$ identity matrix, and $\otimes$ denotes the Kronecker (tensor) matrix product. For $\mathcal{Y} = \mathbb{R}^n$, an entry $K(x_i, x_j)$ of the $n \times n$ kernel matrix is defined by

$$K(x_i, x_j) = k(x_i, x_j) \left(\gamma_O \mathcal{L}_{out}^{\dagger} + (1 - \gamma_O) I_n \right), \qquad (3)$$

where $k(\cdot, \cdot)$ is a scalar-valued kernel, and $\gamma_O \in [0, 1]$ is a parameter. Here, $\mathcal{L}_{out}^{\dagger}$ is the pseudo-inverse of the output labels graph Laplacian. The output similarity graph can be estimated by looking each label as a vertex and using the nearest neighbors method. The representation of the j'th label is the j'th column in the label matrix $Y \in \mathbb{R}^{N \times n}$, in which, $Y_{ij} = 1$ if the j'th label is manually assigned to the i'th sample, and -1 otherwise. For the unlabeled samples, $Y_{ij} = 0$. It has been proven in (Minh and Sindhwani 2011) that the solution of the minimization problem (2) takes the form $f^*(x) = \sum_{i=1}^{N} K(x, x_i) a_i$. The vector-valued Laplacian RLS (regularized least squares) estimates vectors $a_i \in \mathcal{Y}, 1 \leq i \leq N$ by solving a Sylvester Equation.

MV^{3}MR: Multi-view Vector-valued Manifold Regularization

In order to handle multi-view multi-label image classification, we generalize VVMR and present the multi-view vector-valued manifold regularization (MV^{3}MR). In contrast to (Guillaumin, Verbeek, and Schmid 2010), which assumes that different views contribute equally to the classification, MV^{3}MR assumes different views contribute differently to the classification and learns the optimal combination coefficients to integrate these different views.

Given a small number of labeled samples and a relatively large number of unlabeled samples, MV^{3}MR first computes an output similarity graph by using the label information of the labeled samples. The Laplacian of the label graph

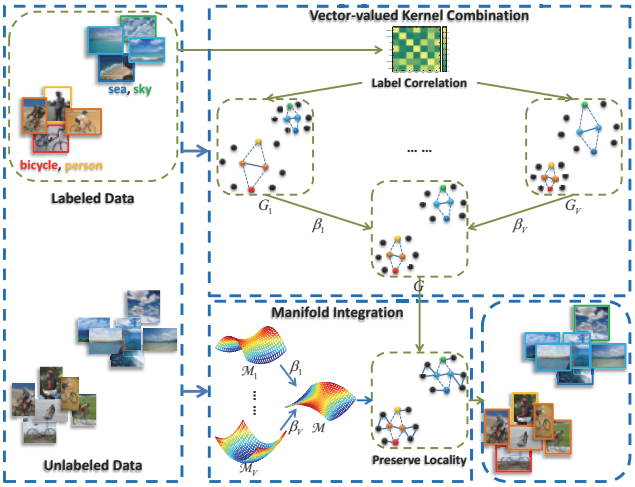

Figure 2: A summary diagram of the proposed MV^{3}MR algorithm. The given labels are used to construct an output similarity graph, which encodes the label correlations. Features from different views of the labeled and unlabeled data are used to construct different Gram matrices (with label correlations incorporated) $G_v, v = 1, \ldots, V$, as well as the different graph Laplacians $\mathcal{M}_v, v = 1, \ldots, V$. We learn the weight β_v by combining both G_v and $\mathcal{M}_v$. The combined Gram matrix G is used for classification while preserving locality on the integrated manifold M.

is incorporated in the scalar-valued Gram matrix G_v^k over labeled and unlabeled data in order to enforce label correlations on each view, so that the vector-valued Gram matrices $G_v = G_v^k \otimes Q, v = 1, \ldots, V$ can be obtained, where $Q = \gamma_O \mathcal{L}_{out}^{\dagger} + (1-\gamma_O) I_n$. Meanwhile, we also compute the vector-valued graph Laplacians $\mathcal{M}_v, v = 1, \ldots, V$ by using the features of the input data from different views. MV^{3}MR then learns the combination coefficient β_v for combining both G_v and $\mathcal{M}_v$ by the use of alternating optimization. Finally, the combined Gram matrix G, together with the regularization on the combined manifold M, is used for classification. Figure 2 summarizes the above procedure. Technical details are given below.

Rationality

Let V be the number of views and v be the view index. On the feature space of each view, we define the corresponding positive definite scalar-valued kernel k_v, which is associated with an RKHS $\mathcal{H}_{k_v}$. It follows from the functional framework (Rakotomamonjy et al. 2008) that there exists an RKHS $\mathcal{H}_k$ associated with the kernel $k(x, x') = \sum_{v=1}^{V} \beta_v k_v(x, x')$, and any function in $\mathcal{H}_k$ is a sum of functions belonging to $\mathcal{H}_{k_v}$. The vector-valued kernel $K(x, x') = k(x, x') \otimes Q = \sum_{v=1}^{V} \beta_v K_v(x, x')$, where we have used the bilinearity of the Kronecker product. Each $K_v(x, x') = k_v(x, x') \otimes Q$ corresponds to an RKHS, as described for the RKHS for vector-valued functions (Minh and Sindhwani 2011). Thus, the kernel K is associated with an RKHS $\mathcal{H}_K$. This functional framework motivates the MV^{3}MR algorithm.

Problem Formulation

In the multi-view setting and theme of manifold regularization, we will learn the vector-valued function f by linearly combining the kernels and graphs from different views. The optimization problem is given by

$$\begin{aligned} \underset{f \in \mathcal{H}_k}{\operatorname{argmin}} \frac{1}{l} \sum_{i=1}^{l} L(f, x_i, y_i) + \gamma_A \|f\|_k^2 \\ + \gamma_I \langle \mathbf{f}, \mathcal{M}\mathbf{f} \rangle_{\mathcal{Y}^{u+l}} + \gamma_B \|\beta\|_2^2, \\ \text{s.t.} \sum_v \beta_v = 1, \beta_v \geq 0, v = 1, \ldots, V \end{aligned} \tag{4}$$

where $\beta = [\beta_1, \ldots, \beta_V]^T$, γ_A, γ_I and γ_B are positive trade-off parameters. The decision function takes the form $f(x) + b = \sum_v f^v(x) + b$ and belongs to an RKHS $\mathcal{H}_K$ associated with the kernel $K(x, x') = \sum_v \beta_v K_v(x, x')$. In addition, $M = \sum_v \beta_v \mathcal{M}_v$ with each $\mathcal{M}_v$ being a vector-valued graph Laplacian constructed on $\mathcal{H}_{K_v}$. Since $\mathcal{H}_K$ is an RKHS, according to (Minh and Sindhwani 2011), the Representer Theorem (Schölkopf and Smola 2002) follows for a fixed set of $\{\beta_v\}$.

Theorem 1 *For a fixed set of $\{\beta_v\}$, the minimizer of problem (4) admits an expansion*

$$f^*(x) = \sum_{i=1}^{N} K(x, x_i) a_i, \tag{5}$$

where $a_i \in \mathcal{Y}, 1 \leq i \leq N = u + l$ are vectors to be estimated, and $K(x, x_i) = \sum_{v=1}^{V} \beta_v K_v(x, x_i)$.

By using the least squares error, i.e. $L(f, x_i, y_i) = \|f(x_i) - y_i\|^2$, we can rewrite (4) as

$$\begin{aligned} \underset{\mathbf{a}, \beta}{\operatorname{argmin}} \frac{1}{l} \|J_{nl}^{Nn} G\mathbf{a} - \mathbf{y}\|^2 + \gamma_A \mathbf{a}^T G \mathbf{a} \\ + \gamma_I \mathbf{a}^T G \mathcal{M} G \mathbf{a} + \gamma_B \|\beta\|_2^2, \\ \text{s.t.} \sum_v \beta_v = 1, \beta_v \geq 0, v = 1, \ldots, V \end{aligned} \tag{6}$$

where $J_{nl}^{Nn} \in \mathbb{R}^{Nn \times Nn}$ is a diagonal matrix with the first nl elements 1, and the rest 0, $\mathbf{a} = \{a_1, \ldots, a_{u+l}\} \in \mathbb{R}^{n(u+l)}$ and $\mathbf{y} = \{y_1, \ldots, y_{u+l}\} \in \mathbb{R}^{n(u+l)}$ are both column vectors, each with $a_i, y_i \in \mathbb{R}^n$, and $y_{l+1} = \ldots = y_{u+l} = 0$. Here, $G = \sum_{v=1}^{V} \beta_v G_v$ is the combined vector-valued Gram matrix over the labeled and unlabeled samples defined on kernel K, $\mathcal{M} = \sum_{v=1}^{V} \beta_v \mathcal{M}_v$ is the integrated vector-valued graph Laplacian.

Optimization

We have two variables, $\mathbf{a}$ and β, to be optimized in (6). In this formulation, there is a cubic term with respect to the variable β, $\sum_{i=1}^{V} \sum_{j=1}^{V} \beta_i \beta_j (\mathbf{a}^T G_i (\sum_{k=1}^{V} \beta_k \mathcal{M}_k) G_j \mathbf{a})$, which is inconvenient for optimization. We therefore introduce the classical alternating direction method of multipliers (ADMM) (Yang and Zhang 2011) method in order to solve this problem. In (6), we replace the graph combination weights $\{\beta_k\}$ with the auxiliary variables $\{\theta_k\}$. Then

the problem can be reformulated as:

$$\begin{aligned}\underset{\mathbf{a},\beta,\theta}{\operatorname{argmin}} &\frac{1}{l}\|J_{nl}^{Nn}G\mathbf{a}-\mathbf{y}\|^2+\gamma_A\mathbf{a}^TG\mathbf{a}\\ &+\gamma_I\mathbf{a}^TG\mathcal{M}G\mathbf{a}+\gamma_B\|\beta\|_2^2,\\ \text{s.t.}&\sum_v\beta_v=1,\beta_v\geq 0,v=1,\ldots,V;\beta=\theta.\end{aligned}\tag{7}$$

Here, G take the form as in (6), while $\mathcal{M}=\sum_{v=1}^V\theta_v\mathcal{M}_v$. Let $\mathcal{W}(\mathbf{a},\beta,\theta)$ be the objective of (7), we use the augmented Lagrangian method (ALM) (Yang and Zhang 2011) to take the constraint $\beta=\theta$ into consideration, and solve the problem (7) by minimizing the following augmented Lagrangian function:

$$\begin{aligned}L^A(\mathbf{a},\beta,\theta;\lambda)=\mathcal{W}(\mathbf{a},\beta,\theta)+\lambda^T(\beta-\theta)+\frac{\mu}{2}\|\beta-\theta\|_2^2,\\ \text{s.t.}\sum_v\beta_v=1,\beta_v\geq 0,v=1,\ldots,V\end{aligned}\tag{8}$$

where λ is a vector of Lagrange multipliers, and $\mu\geq 0$ is a penalty parameter. According to the ALM algorithm, we can solve $L^A(\mathbf{a},\beta,\theta;\lambda)$ for $\mathbf{a}$, β and θ jointly with fixed λ, and then update λ by keeping $\mathbf{a}$, β, θ fixed. The optimization of $\mathbf{a}$, β and θ can also be done separately due to the separable structure of the objective function. We present the steps in Algorithm 1, and the details of optimizing $\mathbf{a}$, β and θ are given as follows:

• **Update for** $\mathbf{a}$: By initializing $\beta_v=\theta_v=\frac{1}{V},v=1,\ldots,V$, and let $G=\sum_{v=1}^V\beta_vG_v$, $\mathcal{M}=\sum_{v=1}^V\theta_v\mathcal{M}_v$, we rewrite (8) with respect to $\mathbf{a}$ as

$$L^A(\mathbf{a})=\underset{\mathbf{a}}{\operatorname{argmin}}\,\mathbf{a}^T(GJG+l\gamma_AG+l\gamma_IG\mathcal{M}G)\mathbf{a}-2\mathbf{a}^TGJ\mathbf{y},\tag{9}$$

where we have ignored the constant term $\mathbf{y}^T\mathbf{y}$. Note that the matrix-valued Gram matrix $G=G^k\otimes Q$, where $G^k\in\mathbb{R}^{N\times N}$ is a scalar-valued Gram matrix and $Q=(\gamma_OL_{out}^\dagger+(1-\gamma_O)I_n)$. Similar as presented in (Minh and Sindhwani 2011), the optimal $\mathbf{a}$ can be obtained by solving an equivalent Sylvester equation,

$$-\frac{1}{l\gamma_A}(J_l^NG^k+l\gamma_I\mathcal{L}G^k)AQ-A+\frac{1}{l\gamma_A}Y=0,\tag{10}$$

where $\mathbf{a}=\text{vec}(A^T)$, and J_l^N is a diagonal matrix where the first l entries 1, and the others 0.

• **Update for** β: With the obtained $\mathbf{a}^*$, the sub-problem for optimizing $L^A(\mathbf{a},\beta,\theta;\lambda)$ with respect to β can be given by

$$\begin{aligned}\underset{\beta}{\operatorname{argmin}}\,&\beta^T\left(H+(l\gamma_B+\frac{\mu}{2})I_V\right)\beta-\beta^Th,\\ \text{s.t.}&\sum_v\beta_v=1,\beta_v\geq 0,v=1,\ldots,V\end{aligned}\tag{11}$$

where we have defined $H_{ij}=(\mathbf{a}^*)^TG_i(J_{nl}^{Nn}+l\gamma_I\mathcal{M})G_j\mathbf{a}^*$ and $h=\{h_1,\ldots,h_V\}$ with each $h_i=2(\mathbf{a}^*)^TG_iJ_{nl}^{Nn}\mathbf{y}-l\gamma_A(\mathbf{a}^*)^TG_i\mathbf{a}^*+\mu\theta_i-\lambda_i$[1]. We adopt the coordinate descent algorithm to solve (11). In each iteration round during the coordinate descent procedure, two elements β_i and β_j are selected to be updated, while the others are fixed. By using the Lagrangian of problem (11) and considering that $\beta_i+\beta_j$ will not change due to the constraint $\sum_{v=1}^V\beta_v=1$, we have the following solution for updating β_i and β_j:

$$\begin{cases}\beta_i^*=\dfrac{(2l\gamma_B+\mu)(\theta_i+\theta_j)+(h_i-h_j)+2t_{ij}}{2(H_{ii}-H_{ij}-H_{ji}+H_{jj})+2(2l\gamma_B+\mu)},\\ \beta_j^*=\beta_i+\beta_j-\beta_i^*,\end{cases}\tag{12}$$

where $t_{ij}=(H_{ii}-H_{ij}-H_{ji}+H_{jj})\beta_i-\sum_k(H_{ik}-H_{jk})\beta_k$. The obtained β_i^* or β_j^* may violate the constraint $\beta_v\geq 0$. Thus, if $(2l\gamma_B+\mu)(\beta_i+\beta_j)+(h_i-h_j)+2t_{ij}\leq 0$, we set $\beta_i^*=0$, and if $(2l\gamma_B+\mu)(\beta_i+\beta_j)+(h_j-h_i)+2t_{ji}\leq 0$, we have $\beta_j^*=0$.

• **Update for** θ: To optimize $L^A(\mathbf{a},\beta,\theta;\lambda)$ with respect to θ, we have the following sub-problem:

$$\begin{aligned}\underset{\theta}{\operatorname{argmin}}\,&\theta^Ts+\frac{\mu}{2}\theta^T\theta,\\ \text{s.t.}&\sum_v\theta_v=1,\theta_v\geq 0,v=1,\ldots,V\end{aligned}\tag{13}$$

where $s=[s_1,\ldots,s_V]^T$ with each $s_v=\gamma_I(\mathbf{a}^*)^TG\mathcal{M}_vG\mathbf{a}^*-\mu\beta_v-\lambda_v$[2]. Similarly, the solution of (13) can be obtained by using the coordinate descent, and the criteria for updating θ_i and θ_j in an iteration round is given by

$$\begin{cases}\theta_i^*=0,\theta_j^*=\theta_i+\theta_j,\text{if }\mu(\theta_i+\theta_j)+(s_j-s_i)\leq 0,\\ \theta_j^*=0,\theta_i^*=\theta_i+\theta_j,\text{if }\mu(\theta_i+\theta_j)+(s_i-s_j)\leq 0,\\ \theta_i^*=\dfrac{\mu(\theta_i+\theta_j)+(s_j-s_i)}{2\mu},\theta_j^*=\theta_i+\theta_j-\theta_i^*,\text{else.}\end{cases}\tag{14}$$

Algorithm 1 The ADMM optimization procedure of the proposed MV^{3}LRLS algorithm

Input: Labeled data $D_l^v=\{(x_i^v,y_i)_{i=1}^l\}$ and unlabeled data $D_u^v=\{(x_i^v)_{i=l+1}^N\}$ form different views, $v=1,\ldots,V$ is the view index.
Algorithm parameters: γ_A, γ_I, γ_B and μ
Output: $N\times n$ matrix A, and the view combination coefficients $\{\beta_v\},v=1,\ldots,V$.

1: Construct the scalar kernel G_v^k and graph Laplacian $\mathcal{L}_v$ for each view, set $\beta_v=\theta_v=1/V,v=1,\ldots,V$; calculate $G^k=\sum_{v=1}^V\beta_vG_v^k$ and $\mathcal{L}=\sum_{v=1}^V\theta_vG_v^k$.
2: **Iterate**
3: Solve for A with the computed G^k and $\mathcal{L}$ through (10), where $\mathbf{a}^{k+1}=\text{vec}(A^T)$;
4: Solve $\beta^{k+1}=\text{argmin}_\beta L^A(\mathbf{a}^{k+1},\beta,\theta^k;\lambda^k)$ using (12) and updated G^k;
5: Solve $\theta^{k+1}=\text{argmin}_\theta L^A(\mathbf{a}^{k+1},\beta^{k+1},\theta;\lambda^k)$ using (14) and updated $\mathcal{L}$;
6: $\lambda^{k+1}=\lambda^k+\mu(\beta^{k+1}-\theta^{k+1})$.
7: **Until convergence**

[1]By using the basic Kronecker product properties, we can reformulate H and h_i with respect to G_i^k and $\mathcal{L}$ as $H_{ij}=(\text{vec}(QA^TG_i^k))^T(\text{vec}(QA^TG_j^kJ_l^N)+l\gamma_I\text{vec}(QA^TG_j^k\mathcal{L}))$ and $h_i=2\mathbf{y}^T\text{vec}(QA^TG_i^kJ_l^N)-l\gamma_A\mathbf{a}^T\text{vec}(QA^TG_i^k)+\mu\theta_i-\lambda_i$.

[2]By the use of basic Kronecker product properties, s_v can be rewritten as $s_v=\gamma_I(\text{vec}(QA^TG^k))^T\text{vec}(QA^TG^k\mathcal{L}_v)-\mu\beta_v-\lambda_v$.

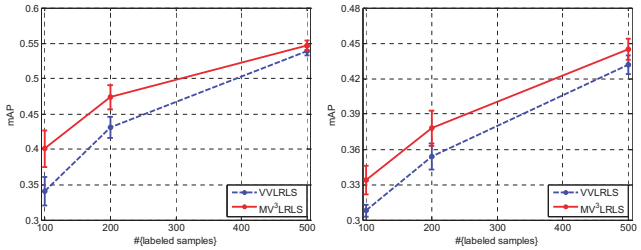

Figure 3: The mAP performance enhancements by learning the weights (β) for different views: (Top: PASCAL VOC 07; Bottom: MIR Flickr).

We summarize the learning procedure of the multi-view vector-valued Laplacian regularized least squares (MV^{3}LRS) method in Algorithm 1. The stopping criterion for terminating the algorithm can be the difference of the objective value between two consecutive steps. Alternatively, we can stop the iterations when the variation of β is smaller than a pre-defined threshold. Our implementation is based on the difference of the objective value. The complexity of MV^{3}LRS is $O(kN^3)$, where k is the number of iterations in Algorithm 1 and N is the number of training samples.

Experimental Evaluation

We validate the effectiveness of MV3MR on two challenge datasets, PASCAL VOC' 07 (VOC) (Everingham et al.) and MIR Flickr (MIR) (Huiskes and Lew 2008). The VOC dataset contains 10,000 images labeled with 20 categories, while the MIR dataset contains 25,000 images labeled with 38 categories. For the VOC dataset (Everingham et al.), we use the standard train/test partition (Everingham et al.), which splits 9,963 images into a training set of 5,011 images and a test set of 4,952 images. For the MIR dataset (Huiskes and Lew 2008), images are randomly split into equally sized training and test sets. For both datasets, we randomly select 20% of the test images for validation. The parameters of all the algorithms compared in our experiments are tuned using the validation set. From the training examples, 10 random choices of $l \in \{100, 200, 500\}$ labeled samples are used in our experiments.

We use several visual views and the tag feature according to (Guillaumin, Verbeek, and Schmid 2010). The visual views include SIFT features (Lowe 2004), local hue histograms (Van De Weijer and Schmid 2006), global GIST descriptors (Oliva and Torralba 2001) and some color histograms (RGB, HSV and LAB). The local descriptors (SIFT and hue) are computed densely on the multi-scale grid and quantized using k-means, which will result in a visual word histogram for each image. Therefore, we have 7 different representations in total.

We pre-compute kernel for each view and normalize it to unit trace. For the visual representations, the kernel is defined by $K(x_i, x_j) = \exp\left(-\lambda^{-1} d(x_i, x_j)\right)$, where $d(x_i, x_j)$ denotes the distance between x_i and x_j. Following the example of (Guillaumin, Verbeek, and Schmid 2010), we choose L1 distance for the color histogram representations (RGB, HSV and LAB), L2 for GIST, and χ^2 for the visual word histograms (SIFT and hue). For the tag features, a linear kernel $K(x_i, x_j) = x_i^T x_j$ is constructed.

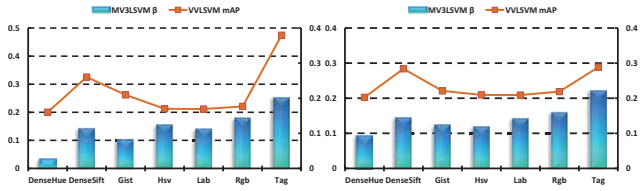

Figure 4: The view combination weight β learned by MV^{3}LRLS, as well as the mAP of using VVMR for each view; Left: PASCAL VOC' 07; Right: MIR Flickr.

Following (Guillaumin, Verbeek, and Schmid 2010), the average precision (AP) (Zhu 2004) is utilized to evaluate the ranking performance under each label. Usually, the mean value over all labels, i.e. mAP is reported.

Performance Enhancement with Multi-view Learning

It has been shown in (Minh and Sindhwani 2011) that VVMR performs well for transductive semi-supervised multi-label classification and can provide a high-quality out-of-sample generalization (Strange and Zwiggelaar 2011). The proposed MV^{3}MR framework is a multi-view generalization of VVMR that incorporates the advantages from MKL for handling multi-view data. Therefore, we first evaluate the effectiveness of learning the view combination weights using the proposed multi-view learning algorithm for transductive semi-supervised multi-label classification. The experimental setup of the two compared methods is given as follows:

• **VVLRLS**: the vector-valued Laplacian RLS presented in (Minh and Sindhwani 2011). The parameters γ_A and γ_I in (2) are both optimized over the set $\{10^i | i = -8, -7, \ldots, -2, -1\}$. We set the parameter γ_O in (3) to 1.0 since it has been demonstrated empirically in (Minh and Sindhwani 2011) that with a larger γ_O, the performance will usually be better. The mean of the multiple Gram matrices and input graph Laplacians are pre-computed for experiments. The number of nearest neighbors for constructing the input and output graph Laplacians are tuned on the sets $\{10, 20, \ldots, 100\}$ and $\{2, 4, \ldots, 20\}$, respectively.

• **MV^{3}LRLS**: a least squares implementation of the proposed MV3MR framework as presented in Algorithm 1. We tune the parameters γ_A and γ_I as in VVLRLS and γ_O is set to 1.0. The additional parameters γ_B and μ are optimized over $\{10^i | i = -8, -7, \ldots, -2, -1\}$. The number of nearest neighbors for constructing the input and output graph Laplacians are optimized as in VVLRLS.

The experimental results on the two datasets are shown in Figure 3. We can see that learning the combination weights using our algorithm is always superior to simply using the uniform weights for different views. We also find that when the number of labeled samples increases, the improvement becomes smaller. This is because the multi-view learning actually helps to approximate the underlying data distribution. This approximation can be steadily improved with the increase of the number of labeled samples, and thus the significance of the multi-view learning to the approximation

Table 1: MAP Performance Evaluation on the Two Datasets

	VOC mAP vs. #{labeled samples}			MIR mAP vs. #{labeled samples}		
Methods	100	200	500	100	200	500
MLCS	0.332±0.017	0.412±0.016	0.525±0.007	0.289±0.010	0.342±0.011	0.424±0.010
KLS_CCA	0.347±0.019	0.432±0.014	0.536±0.007	0.321±0.009	0.369±0.017	**0.445±0.009**
MV^{3}LRLS	**0.401±0.026**	**0.474±0.017**	**0.547±0.007**	**0.334±0.012**	**0.378±0.015**	**0.445±0.009**
SimpleMKL	0.381±0.024	0.453±0.020	0.538±0.011	0.321±0.014	0.365±0.017	**0.444±0.011**
LpMKL	0.391±0.024	0.462±0.012	0.540±0.006	0.327±0.010	0.367±0.014	0.436±0.008

gradually decreases.

Analyses of the Multi-view Learning

In the following, we present empirical analyses of the multi-view learning procedure. In Figure 4, we select $l = 100$ and present the view combination coefficients β learned by MV^{3}LRLS, together with the mAP by using VVLRLS for each view. From the results, we find that the tendency of the kernel and graph weights are both consistent with the corresponding mAP in general, i.e. the views with a higher classification performance tend to be assigned larger weights, taking the DenseSIFT visual view (the 2nd view) and the tag (the last view), for example. However, a larger weight may sometimes be assigned to a less discriminative view; for example, the weight of Hsv (the 4th view) is larger than the weight of DenseSIFT (the 2nd view). This is mainly because the coefficient $\mathbf{a}$ is not optimal for every single view, in which only G_v and $\mathcal{M}_v$ are utilized. The learned $\mathbf{a}$ minimizes the optimization problem (6) by using the combined Gram matrix G and integrated graph Laplacian $\mathcal{M}$, which means that the learned vector-valued function is smooth along the combined RKHS and the integrated manifold. In this way, the proposed algorithm effectively exploits the complementary properties of different views.

Comparisons with Multi-label and Multi-kernel Learning Algorithms

Our last set of experiments compares MV^{3}LRLS with several competitive multi-label methods, as well as some well-known and competitive MKL algorithms, in predicting the unknown labels of the unlabeled data. We specifically compare MV^{3}LRLS with the following methods on the challenging VOC and MIR datasets:

• **MLCS (Hsu et al. 2009)**: a multi-label compressed sensing algorithm that taking advantage of the sparsity of the labels. We choose the label compression ratio to be 1.0 since the number of the labels n is not very large here. Mean of the multiple kernels from different views is pre-computed for experiments.

• **KLS_CCA (Sun, Ji, and Ye 2011)**: a least-squares formulation of the kernelized canonical correlation analysis for multi-label classification. The ridge parameter is chose from the candidate set $\{0, 10^i | i = -3, -2, \ldots, 2, 3\}$. Mean of the multiple kernels is pre-computed to run the algorithm.

• **SimpleMKL (Rakotomamonjy et al. 2008)**: a popular SVM-based multiple kernel learning algorithm that determines the combination of multiple kernels by a reduced gradient descent algorithm. The penalty factor C is tuned on the set $\{10^i | i = -1, 0, \ldots, 7, 8\}$. We apply SimpleMKL to multi-label classification by learning a binary classifier for each label.

• **LpMKL (Kloft et al. 2011)**: a recently proposed MKL algorithm, which extend MKL to l_p-norm with $p \geq 1$. The penalty factor C is tuned on the set $\{10^i | i = -1, 0, \ldots, 7, 8\}$ and we choose the norm p from the set $\{1, 8/7, 4/3, 2, 4, 8, 16, \infty\}$.

The performance of the compared methods on the VOC dataset and MIR dataset are reported in Table 1. From the results, we firstly observe that the performance keeps improving with the increasing number of labeled samples. Secondly, the performance of the simpleMKL algorithm, which learns the kernel weights for SVM, can be inferior to the multi-label algorithms with the mean kernel in many cases. MV^{3}LRLS is superior to multi-view (SimpleMKL and LpMKL) and multi-label algorithms in general, and consistently outperforms other methods in terms of mAP. In particular, in comparison with SimpleMKL, we obtain a 5.2%, 4.6% and 1.7% mAP improvement on VOC when using 100, 200 and 500 labeled samples, respectively. The level of improvement drops when more labeled samples are available, for the same reason described in our first set of experiments.

Conclusion and Discussion

Most of the existing work on multi-label image classification use only single feature representation, and the multiple feature methods usually assume that a single label is assigned to an image. However, an image is usually associated with multiple labels and different kinds of features are necessary to describe the image properly. Therefore, we have developed multi-view vector-valued manifold regularization (MV^{3}MR) for multi-label image classification in which images are naturally characterized by multiple views. MV^{3}MR combines different kinds of features in the learning process of the vector-valued function for multi-label classification. We also derived a least squares formulation of MV^{3}MR, which results in MV^{3}LRLS. The new algorithm effectively exploits the label correlations and learns the view weights to integrate the consistency and complementary properties of different views. Intensive experiments on two challenge datasets PASCAL VOC' 07 and MIR Flickr show that MV^{3}LRLS outperforms the traditional multi-label algorithms as well as some well-known multiple kernel learning methods. Furthermore, our method provides a strategy

for learning from multiple views in multi-label classification and can be extended to other multi-label algorithms.

Acknowledgments

This work is partially supported by NBRPC 2011CB302400, NSFC 60975014, 61121002, JCYJ20120614152136201, NSFB 4102024, and Australian Research Council Discovery Project with number DP-120103730.

References

Belkin, M.; Niyogi, P.; and Sindhwani, V. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *The Journal of Machine Learning Research* 7:2399–2434.

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

Chen, G.; Song, Y.; Wang, F.; and Zhang, C. 2008. Semi-supervised multi-label learning by solving a sylvester equation. In *SIAM international conference on Data Mining*, 410–419.

Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

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

Hsu, D.; Kakade, S.; Langford, J.; and Zhang, T. 2009. Multi-label prediction via compressed sensing. In *Advances in Neural Information Processing Systems*, 772–780.

Huiskes, M. J., and Lew, M. S. 2008. The MIR flickr retrieval evaluation. In *ACM international conference on Multimedia Information Retrieval*, 39–43.

Kloft, M.; Brefeld, U.; Sonnenburg, S.; and Zien, A. 2011. Lp-norm multiple kernel learning. *The Journal of Machine Learning Research* 12:953–997.

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

Luo, Y.; Tao, D.; Geng, B.; Xu, C.; and Maybank, S. 2013. Manifold regularized multitask learning for semi-supervised multilabel image classification. *IEEE Transaction on Image Processing* 22(2):523–536.

Micchelli, C., and Pontil, M. 2005. On learning vector-valued functions. *Neural Computation* 17(1):177–204.

Minh, H., and Sindhwani, V. 2011. Vector-valued manifold regularization. In *International Conference on Machine Learning*, 57–64.

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.; and Grandvalet, Y. 2008. SimpleMKL. *Journal of Machine Learning Research* 9:2491–2521.

Schölkopf, B., and Smola, A. 2002. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press.

Sindhwani, V.; Niyogi, P.; and Belkin, M. 2005. A co-regularization approach to semi-supervised learning with multiple views. In *ICML Workshop on Learning with Multiple Views*, 74–79.

Strange, H., and Zwiggelaar, R. 2011. A generalised solution to the out-of-sample extension problem in manifold learning. In *AAAI Conference on Artificial Intelligence*.

Sun, L.; Ji, S.; and Ye, J. 2011. Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(1):194–200.

Suzuki, I.; Hara, K.; Shimbo, M.; Matsumoto, Y.; and Saerens, M. 2012. Investigating the effectiveness of laplacian-based kernels in hub reduction. In *AAAI Conference on Artificial Intelligence*.

Van De Weijer, J., and Schmid, C. 2006. Coloring local feature extraction. *European Conference on Computer Vision* 334–348.

Vu, H.; Carey, C.; and Mahadevan, S. 2012. Manifold warping: Manifold alignment over time. In *AAAI Conference on Artificial Intelligence*.

Yang, J., and Zhang, Y. 2011. Alternating direction algorithms for \ell_1-problems in compressive sensing. *SIAM journal on scientific computing* 33(1):250–278.

Zhang, M. 2011. Lift: Multi-label learning with label-specific features. In *International Joint Conference on Artificial Intelligence*, 1609–1614.

Zhou, Z.; Zhang, M.; Huang, S.; and Li, Y. 2012. Multi-instance multi-label learning. *Artificial Intelligence* 176(1):2291–2320.

Zhu, M. 2004. Recall, precision and average precision. Technical report, University of Waterloo.