Towards Multi-Semantic Image Annotation with Graph Regularized Exclusive Group Lasso

Xiangyu Chen[†‡§] Xiaotong Yuan[§] Shuicheng Yan[§†] Jinhui Tang[‡]
Yong Rui[♯] and Tat-Seng Chua[†‡]

[†] NUS Graduate School for Integrative Sciences and Engineering,
[§] Department of Electrical and Computer Engineering
[‡] School of Computing, National University of Singapore
[♯]Microsoft Advanced Technology Center
{chenxiangyu,eleyuanx,eleyans,tangjh,chuats}@nus.edu.sg,yongrui@microsoft.com

ABSTRACT

To bridge the *semantic gap* between low level feature and human perception, many image classification algorithms have been proposed in the past decade. With the increasing of the demand for image search with complex queries, the explicit comprehensive semantic annotation becomes one of the main challenging problems. However, most of the existing algorithms mainly aim at annotating images with concepts coming from only one semantic view, e.g. cognitive or affective, and naive combination of the outputs from these views shall implicitly force the conditional independence and ignore the correlations among the views. In this paper, to exploit the comprehensive semantic of images, we propose a general framework for harmoniously cooperating the above multiple semantics, and investigating the problem of learning to annotate images with training images labeled in two or more correlated semantic views, such as *fascinating nighttime*, or *exciting cat*. This kind of semantic annotation is more oriented to real world search scenario. Our proposed approach outperforms the baseline algorithms by making the following contributions. 1) Unlike previous methods that annotate images within only one semantic view, our proposed multi-semantic annotation associates each image with labels from multiple semantic views. 2) We develop a multi-task linear discriminative model to learn a linear mapping from features to labels. The tasks are correlated by imposing the exclusive group lasso regularization for competitive feature selection, and the graph Laplacian regularization to deal with insufficient training sample issue. 3) A Nesterov-type smoothing approximation algorithm is presented for efficient optimization of our model. Extensive experiments on NUS-WIDE-Emotive dataset ($56k$ images) with 8×81 emotive cognitive concepts and two benchmark sub datasets from NUS-WIDE well validate the effectiveness of the proposed approach.

MM '11, November 29 - December 01, 2011, Arizona, USA

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing-indexing methods; H.3.3 [**Information Storage and Retrieval**]: Feature Selection

General Terms

Algorithms, Performance, Experimentation

Keywords

Multi-Semantic Image Annotation, Exclusive Group Lasso

1. INTRODUCTION

With the popularity of photo sharing websites, new web images on a large variety of topics have been growing at an exponential rate. At the same time, the contents in images are also enriched and more diverse than ever before. In order to manege this huge amount of variety of images, there is a basic shift from content-based image retrieval to concept-based retrieval techniques. This shift has motivated research on image annotation which prompted a series of challenges in media content processing techniques. The *semantic gap* [11] between high-level semantics and low-level image features is still one of the main challenging problems for image classification and retrieval. However, image semantics can be viewed at two levels: Cognitive level and Affective level [10]. The two view of image semantics are inter-related and can be used to reinforce each other. However, the existing studies in image semantic annotation mainly aim at the assignment of either the cognitive concepts or affective concepts to a new item separately. As a result of this, the combinational semantic concepts cannot be inferred easily. For example, if we want the "fascinating" images with only "nighttime", this kind of annotations are no longer effective because the "fascinating" classifier cannot identify the concept "nighttime". On the other hand, cognitive image annotation also faces the same challenge, and it can only be trained to identify cognitive concepts. This motivates us to harmoniously embed these two or more semantic views into one general framework for annotating the deeper and multi-semantic labels to images. In this paper, we are particularly interested in explicit multi-semantic [1] image annotation. This framework

[1]The *multi-semantic* (or *polysemy*) retrieval has been explored in [16] for multi-modality (visual and textual) based image retrieval, in which a visual object or text word may belong to several concepts. For example, a "horizontal bar"

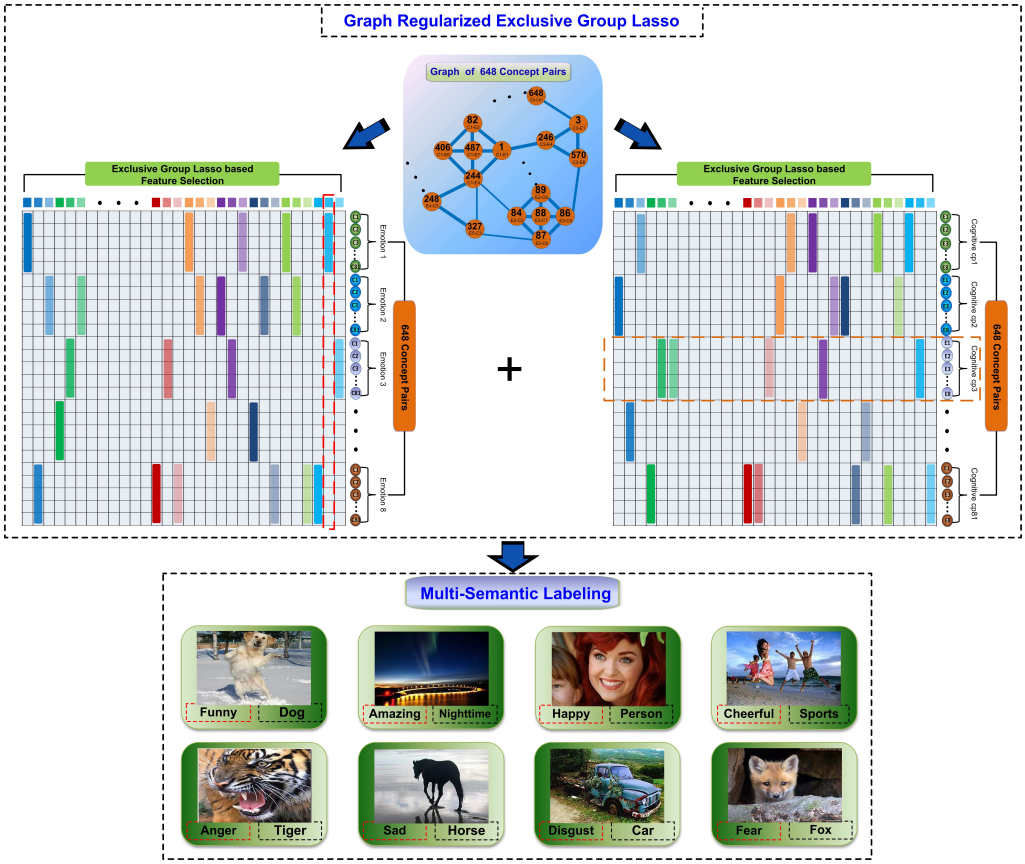

Figure 1: System overview of our proposed MTL scheme for Image Annotation with Multi-Semantic Labeling (IA-MSL). The training data are simultaneously labeled in both cognitive and emotive semantic spaces. The system is trained with a multi-task linear regression model regularized by a graph term (middle of the top part) and an exclusive group lasso term (left and right of the top part). The graph term encourages correlation among tasks while the exclusive group lasso term encourages sparse feature selection across different concepts, i.e., negative relationship among tasks. For better viewing, please see original color pdf file.

not only works well on cognitive and affective views but also can be also applied to other multi-view semantics such as object and scene.

In recent years, the semantic-based image annotation has become one of the most important research directions in multimedia community, which focuses on developing automatic annotation algorithms to extract the semantic meaning of images. For cognitive semantics, we usually assign appropriate cognitive concepts to the image for representing and identifying its visual contents. Affective semantics are represented in adjective form and describe the intensities of feelings, moods or sensibility evoked in users when viewing images, such as Amusement, Awe, Contentment, Excitement, Anger, Disgust, Fear and Sad [22, 23]. Due to the huge number of diverse images on the Internet, current image annotation methods based on unitary semantic view are no longer sufficient and efficient. For popular cognitive or affective queries, the returned images can fill many result pages in popular search engines, which may not satisfy the deeper requirement of complex and multi-semantic retrieval. For example, most commercial systems can handle the individual emotional/cognitive words well, like searching only for "cat" or searching only with the word "exciting". But for the case of searching for images with the query "exciting cat", the precision of result will be degraded because most images are only labeled with either affective concepts or cognitive concepts and the desired multi-semantic labeled sample images are really rare due to the lack of mature multi-semantic image annotation technique. Hence a new methodology is needed to explore and label the deeper meanings of images.

Learning to annotate the combinational semantic to images in multi-semantic views is a challenging problem in the real world applications. In this paper, we propose a novel and promising approach, namely, Image Annotation with Multi-Semantic Labeling (IA-MSL), to annotate images simultaneously with labels in two or more correlated views. The key challenge with IA-MSL is the large number of classes involved in training due to the combination of multiple semantic views. Thus, some classes may suffer from the problem of insufficient training samples. A naive solution to avoid this issue is to train the classifiers within each semantic view and then combine the outputs from these semantic views for the ultimate combinational semantic prediction, which however implicitly imposes the conditional independency assumption and ignore the correlations among the

object can belong to high jump or pole vault event. Differently, the term multi-semantic used in this paper emphasizes that an image can be labeled in multiple semantic views.

semantic views. To deal with such an issue and harness the correlations cross semantic views, we propose to formulate IA-MSL as a regularized multi-task discriminative analysis model, where individual tasks are defined as learning linear discriminative models for individual complex semantic concepts. We propose to learn all the tasks in a joint manner by imposing two types of regularization, the graph Laplacian regularization and exclusive group lasso regularization. The graph Laplacian regularization captures the correlation clues to refine concept classifier, especially in cases with insufficient training samples. For each semantic view, since the image features are typically exclusively shared among different concepts in this space, we also exploit a so called exclusive-group-lasso regularizer to capture such negative correlations among category groups (8 emotion groups or 81 cognitive concept groups). Taking the NUS-WIDE-Emotive dataset as an example, in both emotive view with 8 concepts and cognitive view with 81 concepts, it is reasonable to assume that if an image feature is important for one of several concepts, it is less likely for this feature to be also important for the other concepts. Such an exclusive regularization mechanism is empirically shown to be effective to boost the multi-semantic labeling performance

1.1 Our Contributions

The major contributions of this paper are three-fold:

- We propose a novel framework for Image Annotation with Multi-Semantic Labeling (IA-MSL), which exploits high-level semantic of images from two or more semi-orthogonal label views;
- As an implementation of IA-MSL, we develop a multi-task discriminative analysis model to learn a proper linear mapping from features to labels. The proposed model simultaneously considers co-occurrent relationship among tasks through the graph Laplacian regularization, and the negative relationship among tasks in feature sharing.
- A Nesterov-type smoothing approximation algorithm is developed for efficient optimization of the proposed model. Empirical results on real-world large scale datasets validate the efficiency and effectiveness of our approach.

1.2 Related Work

1.2.1 Image Annotation in Unitary Semantic View

Many multi-label/unitary-label annotation algorithms were proposed and well studied to assign labels to each image for a fixed image collection crawled from websites such as Flickr. For this fixed data set, images are assigned with either cognitive concepts or emotive concepts, or other semantic concepts from a unitary semantic view.

For the image annotation in cognitive semantic view, multi-label propagation is a hot and promising research direction. Many methods were proposed to exploit the inter-relations among different labels [19] since some class labels may correlate to each other. For example, Ueda and Saito [31] proposed a generative model for multi-label learning that explicitly incorporates the pairwise correlation between any two class labels. A Bayesian model is introduced in [9] to assign labels through underlying latent representations. Qi *et al.* [28] proposed a unified Correlative Multi-Label (CML) framework to simultaneously classify labels and model correlations between them. Liu *et al.* [21] utilized constrained nonnegative matrix factorization (CNMF) to optimize the consistency between image similarity and label similarity. Zhu *et al.* [40] suggested a maximum entropy model for exploring the label correlation for multi-label learning. In the setting of large-scale multi-label annotation, Chen *et al.* [4] proposed the Kullback-Leibler divergence based multi-label propagation, which encodes the label information of an image as a unit label confidence vector and imposes inter-label constraints and manipulates labels interactively.

For the image annotation in emotive semantic view, most researchers mainly focus on emotive feature analysis and extraction. The popularly adopted methods include Support Vector Machines (SVMs), Random Forest, the C4.5 tree classifier and Naive Bayes classifier method. For example, Machajdik *et al.* [22] investigated methods to extract and combine low-level features that represent the emotional content of an image from psychology and art theory views. For classification, they adopted the Naive Bayes classifier to annotate images with emotive concepts. Wang *et al.* [26] developed and extracted special integrated histogram features and utilize support vector regression for automatically emotional image annotation. In [35], an SVM framework for supervised learning of emotion categories was also adopted with extracting the holistic image features. Besides the specific features, many other works also utilized the generic features. Hayashi *et al.* [13] adopted the RGB color feature and classified the images through neural network. Wu *et al.* [34] adopted SVM method for affective image classification based on general color and texture features.

Different from the above body of efforts on image annotation in unitary semantic view, we propose and investigate the problem of multi-semantic image annotation to meet the requirement of realworld search conditions. For example, the users on the web not only want to look for the images including the "nighttime", but also find these kinds of images with "fascinating" feelings, which express the deeper and higher semantic meanings. The above reviewed images annotation methods may perform well in unitary semantic view for a fixed dataset, but typically hard to be generalized to handle the multi-semantic labeling problem. Our proposed IA-MSL method is a novel solution designed for the latter and has been empirically shown to work well in real world datasets.

1.2.2 Multi-task Learning

Recently, there have been a lot of interests around multi-task learning (MTL), both in theory and practice. The idea behind this paradigm is that, when the tasks to be learned are similar enough or are related in some sense, it may be advantageous to take into account these relations between tasks. Several works have experimentally highlighted the benefit of such a framework [3]. In general, MTL can be addressed through a regularization framework [7]. For example, the joint sparsity regularization favors to learn a common subset of features for all tasks [1][27], while the exclusive sparsity regularization is used in [39] for exclusive feature selection across tasks. Our method follows the regularized MTL framework. In contrast to the existing regularization that is only model parameters dependent, our proposed regularization is characterized by data as well as model parameters, and thus is much more informative.

1.2.3 Group Sparse Inducing Regularization

Learning models regularized by group sparse inducing penalties have been widely studied in both machine learning [36, 38] and signal processing fields [17, 8]. Let $w \in \mathbb{R}^d$ be the n parameters to be regularized. Denote $\mathcal{I} = \{1, ..., d\}$ the variable index and $\mathcal{G} = \{g_i \subseteq \mathcal{I}\}_{i=1}^{l}$ a set of variable index groups. The group formation varies according to the given grouping or hierarchical structure. Denote $\|w_{\mathcal{G}}\|_{p,q} := \sum_{g\in\mathcal{G}} \|w_g\|_p^q$ the $\ell_{p,q}$-norm defined over groups $\mathcal{G}$, where $\|w_g\|_p^q := \left(\sum_{j\in g} |w_j|^p\right)^{q/p}$. The $\ell_{2,1}$-norm regularizer is used in group Lasso [36] which encourages the sparsity on group level. Jacob *et al.* [14] proposed the overlap group Lasso and graph Lasso as variants of group Lasso to handle overlapping groups. Another group sparsity inducing regularizer is the $\ell_{\infty,1}$-norm which is widely used in multi-task learning problems [20, 37]. When $p = 1$, $q = 2$, the $\ell_{1,2}$-norm has recently been studied in the exclusive-Lasso model [39] for the multi-task learning and elitist-Lasso model [18] for audio signal denoising. Unlike the group Lasso regularizer that assumes covariant variables in groups, the exclusive Lasso regularizer models the scenario when variables in the same group compete with each other to be selected in the output.

2. OUR PROPOSED SCHEME

2.1 Problem Statement

Given a labeled dataset $\{x_i, l_i\}_{i=1}^N$, where $x_i \in R^d$ is the feature vector of the i-th image and l_i is the associated image label. In this study, we assume that l_i is obtained from two or mote different views of labeling. Formally, $l_i = \{l_i^k\}_{k=1}^K$ where $l_i^k \subseteq \mathcal{L}^k$ is the label(s) of image i in the k-th labeling view equipped with label set $\mathcal{L}^k$. It is noteworthy the difference between our multi-semantic labeling classification and the so called multi-label classification. In the latter problem, the labels associated with an image is from a unitary semantic space, e.g., object category. Differently, in our setting, we are interested in the case that the labels associated with the same image are obtained from different semantic views, e.g., object category and emotion. Indeed, for each view k, the label l_i^k can be a multi-label vector in this view. In the following descriptions, for simplicity and clarity purpose, we consider without loss of generality that the labels are obtained from $K = 2$ semantic views. Denote $\mathcal{L} = \mathcal{L}_1 \times \mathcal{L}_2$ the Cartesian products of $\mathcal{L}^1$ and $\mathcal{L}^2$. Let $y_i \in R^{|\mathcal{L}|}$ be the zero-one label matrix indicating whether x_i is jointly labeled as $l^1 \in \mathcal{L}^1$ and $l^2 \in \mathcal{L}^2$. By concatenating the columns of label matrix y_i, we get an $|\mathcal{L}|$ dimensional label vector, which is also denoted by y_i in the rest of this paper. Given the training feature-label set $\{x_i, y_i\}_{i=1}^N$, we are interested in the problem of learning a linear mode $y = Wx$ such that the label of an unseen test sample can be predicted via this model. Naively, one could utilize the following multivariate least squares regression (LSR) model

$$\min_W \left\{ J(W) := \frac{1}{2}\|Y - WX\|^2 \right\}, \tag{1}$$

where $X = [x_1, ..., x_n] \in \mathbb{R}^{d\times n}$ is the feature matrix with each column a training image feature, $Y = [y_1, ..., y_n] \in \mathbb{R}^{|\mathcal{L}|\times n}$ is the label matrix with each column a training image label vector, $W \in \mathbb{R}^{|\mathcal{L}|\times d}$ is the parameter to be estimated. Obviously, the proceeding LSR forms an MTL since the objective J in (1) can be rewritten as:

$$J(W) = \sum_{j=1}^{|\mathcal{L}|} \frac{1}{2}\|Y_j - W_j X\|^2, \tag{2}$$

where $Y_j \in \mathbb{R}^n$ and $W_j \in \mathbb{R}^d$ are the j-th row of Y and W, respectively. In the preceding MTL formulation, we are to learn $|\mathcal{L}|$ different linear regression models (tasks) $Y_j = W_j X$, $j = 1, ..., |\mathcal{L}|$. In this naive formulation, the tasks are learned independently to each other.

For better performance, it is often beneficial to take into account the relationships across tasks by imposing certain regularization to the objective (2). Particular, in the setting of our multi-semantic labeling problem, there are two types of correlations among tasks should be considered.

- **Exclusive feature selection:** In each semantic view, our objective is to differentiate the related categories. Motivated by the exclusive feature sharing prior considered in [39], we may expect a negative correlation among categories, namely, if a visual feature is deemed to be important for one category, it becomes less likely for this feature to be an important feature for the other categories. In order to capture such an exclusive feature selection nature among categories in each semantic view, we propose to utilize an $\ell_{2,1}^2$-norm regularizer analog to the ℓ_1^2-norm regularizer used in the exclusive Lasso model [39].

- **Concepts correlation:** Another important regularization we should explore is the semantic relationship between the combinational concepts in $\mathcal{L}$. This is of particular interest in our work due to the insufficient sample issue severely occurs in multi-semantic annotation. That is, some of the combinational labels in $\mathcal{L}$ are supported by very few or even zero training samples. For example, in our emotion-category dataset, although the category "dog" and the emotion "happy" are supported by plenty of samples, the combinational label ("dog", "happy") is not supported by any sample in the training set. Obviously, for any label j without training samples, $Y_j = 0$, and thus the corresponding W_j will be a zero vector through naive model (2). To handle this issue, one natural way is to propagate the correlation among concepts to their corresponding model parameters. As we will see shortly, the Google similarity distance [6] is a simple and effective choice to describe the correlation among concepts.

Next, we describe in detail the two types of regularization we imposed to the naive MTL model (2).

2.2 An Exclusive Group Lasso Regularizer

In this subsection, we address the regularization of feature exclusive selection across tasks. Let $\mathcal{G}^1$ of size $|\mathcal{L}^1|$ be a group of label index set in $\mathcal{L}$ constructed as follows: each element $g \in \mathcal{G}^1$ is an index set of combinational labels $(l^1, l^2) \in \mathcal{L}$ which share a common $l^1 \in \mathcal{L}^1$. For example, for the category-emotion label views, each group in $\mathcal{G}^1$ is the combination of emotion labels of a certain category. Similarly, we can construct $\mathcal{G}^2$ of size $|\mathcal{L}^2|$ associated with label

set $\mathcal{L}^2$. Let us consider the following regularizer:

$$\Omega(W) := \frac{1}{2}\sum_{i=1}^{d}\left(\|W^i_{\mathcal{G}^1}\|^2_{2,1} + \|W^i_{\mathcal{G}^2}\|^2_{2,1}\right), \quad (3)$$

where $\|W^i_{\mathcal{G}^k}\|^2_{2,1} = \left(\sum_{g\in\mathcal{G}^k}\|W^i_g\|_2\right)^2$, $k = 1, 2$, and $W^i \in \mathbb{R}^{|\mathcal{L}|}$ is the i-th column of W, $W^i_g \in \mathbb{R}^{|\mathcal{L}|}$ is the restriction of vector W^i on the subset g by setting $W^i_j = 0$ for $j \neq g$. For each feature i, the $\ell^2_{2,1}$-norm regularizer $\|W^i_{\mathcal{G}^k}\|^2_{2,1}$ can be viewed as a group Lasso extension of ℓ^2_1 regularizer used in exclusive Lasso [39]. Similar to the analysis in [39], one can confirm that $\|W^i_{\mathcal{G}^k}\|^2_{2,1}$ is sparse inducing and it encourages exclusive selection of features at the level of group $g \in \mathcal{G}^k$. In other words, for each feature i, it tends to assign larger weights to some important groups while assigning small or even zero weights to the other groups.

2.3 A Graph Laplacian Regularizer

We explore in this subsection the semantic relationships between concepts. Suppose that we are given a similarity matrix $P \in \mathbb{R}^{|\mathcal{L}|\times|\mathcal{L}|}$ that stores the pairwise similarity score between concepts. The larger P_{jk} is, the more similar two concepts j and k are, and vice verser. We propose to use the following graph regularizer

$$\Psi(W) := \frac{1}{2}\sum_{j,k=1}^{|\mathcal{L}|} P_{jk} \parallel W_j - W_k \parallel^2. \quad (4)$$

The intuition behind the preceding regularizer is that closely related concepts should have similar regression weights. Different from the $\Omega(W)$ in previous subsection that describes the negative correlation among tasks, the graph regularizer $\Psi(W)$ models the positive correlation among tasks by transferring the weight information among neighboring concepts. Such a mechanism is particularly helpful for robust learning of weights for some combinational concepts only supported by very few or even zero instances in the training set. Denote $L = D - P$ the Laplacian matrix where D is a diagonal matrix whose diagonal entries are the row sums of P. $\mathrm{Tr}(\cdot)$ represents the matrix trace operation. We may equivalently reexpress (4) as the following compact form

$$\Psi(W) = \frac{1}{2}\mathrm{Tr}[W^T L W].$$

Generally speaking, the similarity matrix P can be defined based on any reasonable co-currency measurement such as Google distance [6] and Flickr distance [33]. In our implementation, P is obtained by applying the Normalized Google similarity Distance (NGD) proposed by Cilibrasi and Vitanyi [6]. NGD is simply estimated by exploring the textual information available on the Web. The distance between two concepts is measured by the Google page counts when querying both concept names to the Google search engine. It assumes that the words and phrases acquire meaning from the way they are used in society. Since Google has indexed a vast number of web pages, and the common search term occurs in millions of web pages, this database can somewhat reflect the term distribution in society. Formally, $NGD(x, y)$ between two concepts x and y is defined as

$$NGD(x, y) = \frac{\max\{\ln f(x), \ln f(y)\} - \ln f(x, y)}{\ln N - \min\{\ln f(x), \ln f(y)\}},$$

where $f(x)$, $f(y)$, and $f(x, y)$ in this paper denote the number of images from training data containing concept-pair $x \in \mathcal{L}^1 \times \mathcal{L}^2$ (e.g. emotive-cognitive pair), $y \in \mathcal{L}^1 \times \mathcal{L}^2$, both x and y, respectively. N is the total number of images in training data. We then define $P(x, y) = \exp\{-N(x, y)/\eta\}$ where η is a tunable parameter. The similarity matrix P can also be calculated by other co-occurrent technologies such as Flickr distance [33].

2.4 Graph Regularized Exclusive Group Lasso

Based on the discussion in the previous two subsections, we propose to extend the naive MTL model (1) to the following graph regularized exclusive Lasso MTL:

$$\min_W \left\{ F(W) := \underbrace{J(W) + \lambda\Omega(W)}_{\text{exclusive group Lasso}} + \underbrace{\gamma\Psi(W)}_{\text{graph regularizer}} \right\}. \quad (5)$$

As aforementioned, Equation (5) formulates a regularized MTL with $|\mathcal{L}|$ tasks, each of which learns a linear regression model for certain combinational concept in $\mathcal{L}$. The first two terms in (5) form an exclusive group Lasso objective. The regularizer $\Omega(W)$ encourages the exclusive relationships across tasks. The graph Laplacian regularizer $\Psi(W)$ enforces the semantic correlation among tasks. Through the regularized MTL formulation (5), the parameters W can be learned in a joint manner. It is straightforward to verify that the objective $F(W)$ in (5) is convex but non-smooth since all the three components are convex whereas $\Omega(W)$ is non-smooth. We will develop in the next section an efficient method to optimize problem (5). Once the optimal parameter W^* is obtained, the label vector of a test sample with feature x is given by $y = W^*x$. Such a vector can be used for performance evaluation over testing data.

3. OPTIMIZATION

The non-smooth structure of $\Omega(W)$ makes the optimization of problem (5) a non-trivial task. The general purpose subgradient method as used in [39] is applicable but it typically ignores the structure of problem and suffers from slow rate of convergence. Our idea for optimization is to approximate the original non-smooth objective by a smooth function and then solve the latter by utilizing some off-the-shelf fast algorithms. In this section, we derive a Nesterov's smoothing optimization method [25] to achieve this purpose.

3.1 Smoothing Approximation

It is standard to know that for any vector $p \in \mathbb{R}^n$, its ℓ_2-norm $\|p\|_2$ has a max-structure representation $\|p\|_2 = \max_{\|v\|_2\leq 1}\langle p, v\rangle$. Based on this simple property and the smoothing approximation techniques originally from [25], function $\Omega(W)$ can be approximated by the following smooth function

$$\Omega_\mu(W) = \frac{1}{2}\sum_{i=1}^{d}\left(q^2_{\mathcal{G}^1,\mu}(W^i) + q^2_{\mathcal{G}^2,\mu}(W^i)\right), \quad (6)$$

where

$$q_{\mathcal{G}^k,\mu}(W^i) := \max_{\|V^{i,k}_{\mathcal{G}^k}\|_{2,\infty}\leq 1}\langle W^i, V^{i,k}\rangle - \frac{\mu}{2}\|V^{i,k}\|^2_2. \quad (7)$$

Herein, μ is a parameter to control the approximation accuracy. Formally, we have the following result on approximation accuracy of Ω_μ towards Ω:

PROPOSITION 1. *Assume that* $\|W^i\|_2 \leq R$. *Then* $\Omega_\mu(W)$ *is a* μ*-accurate approximation to* $\Omega(W)$, *that is*

$$\Omega_\mu(W) \leq \Omega(W) \leq \Omega_\mu(W) + C\mu, \tag{8}$$

where $C \equiv \sqrt{2}dR\left(|\mathcal{L}^1|^2 + |\mathcal{L}^2|^2\right)/2$.

The proof is given in Appendix A. Proposition 1 shows that for fixed $\mu > 0$, the function Ω_μ can be seen as a uniform smooth approximation of function Ω.

For a fixed W^i, denote $V^{i,k}(W^i)$ the unique minimizer of (7) for $k = 1, 2$, respectively. It is easy to check that for $k = 1, 2$, $\forall g \in \mathcal{G}^k$,

$$V_g^{i,k}(W^i) = \frac{W_g^i}{\max\left\{\mu, \|W_g^i\|_2\right\}}.$$

The following result states that Ω_μ is differentiable and its gradient can be analytically calculated:

THEOREM 1. *Function* $\Omega_\mu(W)$ *is well defined, convex and continuously differentiable with gradient*

$$\nabla\Omega_\mu(W) = \left[\nabla\Omega_\mu(W^1), ..., \nabla\Omega_\mu(W^d)\right], \tag{9}$$

where for $i = 1, ..., d$,

$$\nabla\Omega_\mu(W^i) = q_{\mathcal{G}^1,\mu}(W^i)V^{i,1}(W^i) + q_{\mathcal{G}^2,\mu}(W^i)V^{i,2}(W^i). \tag{10}$$

Moreover, $\nabla\Omega_\mu(W)$ *is Lipschitz continuous with the constant*

$$L_\mu = \left(\frac{2\sqrt{2}R}{\mu} + |\mathcal{L}^1|^2 + |\mathcal{L}^2|^2\right)d. \tag{11}$$

The proof is given in Appendix B.

3.2 Smooth Minimization via APG

Based on the results in the previous subsection, we now propose to solve the following smooth optimization problem as an approximation to the non-smooth problem (5):

$$\min_W \{F_\mu(W) := J(W) + \lambda\Omega_\mu(W) + \gamma\Psi(W)\}. \tag{12}$$

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

$$\nabla F_\mu(w) = (WX - Y)X^T + \lambda\nabla\Omega_\mu(W) + \gamma LW.$$

Therefore, we can apply any first-order methods, e.g., proximal gradient descent [24] and BFGS [15], to optimize the smooth objective (12). In our implementation, for simplicity and efficiency, we employ the Accelerated Proximal Gradient method [30] to optimize the smoothed problem (12). The algorithm is formally described in Algorithm 1. For a fixed μ, it is shown that APG has $\mathcal{O}(1/t^2)$ asymptotical convergence rate bound, where t is the time instance. If we describe convergence in terms of the number of iterations needed to reach an ϵ solution, i.e., $|F_\mu(w) - \min F_\mu| \leq \epsilon$, then by choosing $\mu \approx \epsilon$ the rate of convergence is $O(1/\epsilon)$. It is noteworthy that the convergent complexity of Algorithm 1 depends on constant $1/L_\mu$ which is dominated by the factor μ when it is small. To further accelerate Algorithm 1 for extremely small μ, one may apply the continuation technique as suggested in [2].

Algorithm 1 Smooth minimization for Problem (5)

Input: $X \in \mathbb{R}^{d\times n}$, $Y \in \mathbb{R}^{|\mathcal{L}|\times d}$, $\mathcal{G}^1$, $\mathcal{G}^2$, λ, γ, μ.
Output: $W^t \in \mathbb{R}^{|\mathcal{L}|\times d}$
Initialization: Initialize W_0, V_0 and let $\alpha_0 \leftarrow 1$, $t \leftarrow 0$.
repeat
  $U_t = (1 - \alpha_t)W_t + \alpha_t V_t$,
  Calculate $\nabla\Omega_\mu(U_t)$ according to (9), (10), and L_μ according to (11).
  $V_{t+1} = V_t - \frac{1}{\alpha_t L_\mu}\left(-(Y - WX)X^T + \lambda\nabla\Omega_\mu(U_t) + \gamma LW\right)$,
  $W_{t+1} = (1 - \alpha_t)W_t + \alpha_t V_{t+1}$,
  $\alpha_{t+1} = \frac{2}{t+1}$, $t \leftarrow t + 1$.
until Converges

4. EXPERIMENTS

To validate the effectiveness of IA-MSL, we conduct extensive experiments on two large scale image datasets: NUS-WIDE-Emotive; NUS-WIDE-Object&Scene [5]. The NUS-WIDE-Emotive set contains two types of semantic labels: cognitive concept category with 81 tags and emotion category with 8 affective tags. The underlying image diversity and complexity make it a good test bed for multi-semantic image annotation experiments. The publicly available NUS-WIDE-Object&Scene is a subset of NUS-WIDE [5] obtained after noisy tag removal. It is also annotated in two sematic views: the scenes category with 33 tags and objects category with 31 tags, which is also suitable for our test. Moreover, since unitary semantic is a special case of multi-semantic, we also compare our proposed algorithm with existing methods on NUS-WIDE-Emotive with individual cognitive semantic and emotive semantic, separately. We report quantitative results on both datasets, with an emphasis on the comparison with the state-of-the-art related algorithms in terms of annotation accuracy.

4.1 Datasets

NUS-WIDE-Emotive dataset is an emotion version of the publicly available NUS-WIDE-LITE [5] database consisting of 55,615 images. Two kinds of semantic labels are associated to each image: an 81-D label vector indicating its relationship to 81 cognitive object categories and an 8-D label vector indicating its relationship to 8 affective semantic concepts(tightly related to tags yet relatively high-level). For cognitive semantic, the 81-D object category label vector for each image is immediately available from NUS-WIDE. For emotive semantic concepts, we adopt the similar categories as studied in [22, 23]: Amusement, Awe, Contentment, Excitement, Anger, Disgust, Fear, Sad to represent 8 different types of positive and negative emotions. Each emotive concept covers several similar emotions as show in Tabel 1. To label the emotive ground truth on this dataset, the images were peer rated in a web-survey where the participants could select the best fitting emotional category from the eight categories. 10 human subjects with almost equal gender distribution and with ages ranging from 23 to 30 years old have helped to achieve the annotation task. For each image the category with the most votes was selected as the ground truth. Images with inconclusive human votes were removed from the set. For our experiment, We randomly select half of the images for training and the rest for testing. On image features, we use a 1134-D feature as a concatenation of 225-D blockwise color moments, 128-D wavelet texture,

Table 1: A list of detailed emotions the eight emotive categories cover.

Name	Similar Words
Amusement	fun, delight, playful, entertainment
Awe	amazing, wonder, admiration, fascinating
Contentment	happy, calm, relaxed, satisfaction
Excitement	joy, cheerful, lively, exhilaration
Disgust	yucky, repellent, revolting, distasteful
Fear	dread, horror, concern, creepy, terrible
Sad	sorrow, melancholy, misery, unhappiness,
Anger	fury, rage, wrath, choler, offense

75-D edge direction histogram, 64-D color histogram, 144-D color correlogram and 500-D bag of visual words [5].

NUS-WIDE-Object&Scene [5] are two subsets from NUS-WIDE. In this paper, we select 50,000 images from these two datasets. It consists of two kinds label categories: 31 concepts for object category and 33 concepts for scene category. Each image is assigned with a 31-D object label vector and a 33-D scene label vector. For evaluation, we construct a training set of size 25,000 whilst the rest are used for testing. The same 1134-D feature as used for the previous dataset is also applied here.

4.2 Baselines and Evaluation Criteria

We systematically compare our proposed IA-MSL with six baseline algorithms as listed in Table 2. Amongst them,

- The *support vector machines* (SVM) is a baseline for binary-class classification problem. Here we use its multi-class version by adopting the conventional one-vs-all strategy.
- The Naive Multi-task Learning (NMTL) refers to the independent MTL regression model (2).
- The SVM-E and NMTL-E are two *enrichment* methods of SVM and NMTL, respectively. By saying enrichment of a classifier from two semantic spaces $\mathcal{L}^1$ and $\mathcal{L}^2$, we mean to train two such classifiers (with confidence label vector output) in $\mathcal{L}^1$ and $\mathcal{L}^2$ separately, and then obtain multi-semantic confidence vector y of test sample x using the following strategy

$$y = y^1 \otimes y^2, \tag{13}$$

  where $y^1 \in \mathbb{R}^{|\mathcal{L}^1|}$ and $y^2 \in \mathbb{R}^{|\mathcal{L}^2|}$ are the label confidence vectors of x from semantic space $\mathcal{L}^1$ and $\mathcal{L}^2$, respectively, and $\otimes$ denotes the Kronecker product. In such a scheme, we made the semantic space independent assumption, i.e., $P(l^1, l^2 \mid x) = P(l^1 \mid x)P(l^2 \mid x)$, $\forall l^1 \in \mathcal{L}^1, l^2 \in \mathcal{L}^2$.
- The Multi-task Learning with Graph Laplacian (MTLG) and Multi-task Learning with Exclusive Lasso (MTLE) are two special cases of the regularized MTL framework (5), by setting $\lambda = 0$ and $\gamma = 0$, respectively.

In order to further study the performance in unitary semantic space, we also compare IA-MSL with several state-of-the-art annotation algorithms as listed in in Table 3, on each semantic space of NUS-WIDE-Emotive.

Many measurements can be used to evaluate multi-label image annotation performance for concepts propagated to the unlabeled images, e.g., ROC curve, precision recall curve, Average Precision (AP), and so on. In this work, we adopt one of the most widely used criteria, AUC (area under ROC curve) [12], for annotation accuracy evaluation on each category, and Mean AUC (MAUC) for average performance evaluation on the entire dataset. All experiments are conducted on a common desktop PC equipped with Intel dual-core CPU (frequency: 3.0 GHz) and 32G bytes physical memory.

Table 2: The baseline algorithms.

Name	Methods
SVM	Support Vector Machine
SVM-E	The enrichment of SVM from individual spaces.
NMTL	Naive MTL as in (2)
NMTL-E	The enrichment of N-SVM from individual spaces.
MTLG	Regularized MTL with only graph Laplacian
MTLE	Regularized MTL with only exclusive group Lasso

Table 3: The baseline algorithms for comparison in individual semantic spaces of NUS-WIDE-Emotive.

Name	Methods
SVM	Support Vector Machine
LNP	Linear Neighborhood Propagation [32]
EGSSC	Entropic Graph Semi-Supervised Classification [29]
LSMP	Large-scale Multi-label Propagation [4]

4.3 Experiment-I: NUS-WIDE-Emotive

On NUS-WIDE-Emotive, we category all labels into 648 (8 emotions $\times$ 81 objects) combination classes. The ground truth of 648 labels is derived by simple Cartesian product of 8 emotive labels and 81 cognitive labels. Some of these 648 multi-semantic labels suffer from the issue of insufficient training samples, which is not rare in real world retrieval scenario. In such a multi-semantic setting, we compare IA-MSL with six baselines listed in Table 2. Table 4 lists the quantitative results. Note that for each of the 8 emotive classes, its AUC is obtained by averaging over the 81 AUCs associated with this emotion but for different object categories. The AUCs for 81 object categories are calculated similarly but omitted from this conference submission due to space limit. From these results we are able to make the following observations:

- IA-MSL solution simultaneously outperforms the competing methods in MAUC and AUCs on all of the 8 emotive classes.
- On comparison between IA-MSL and NMTL, since both utilize the same features, the improvement of the former over the latter is supposed to stem from the fact that IA-MSL explicitly encodes exclusive group lasso and graph Laplacian regularizer in discriminative analysis. As simplified versions of IA-MSL, MTLG and MTLE are both superior to NMTL but inferior to IA-MSL.
- It is interesting to note that the enrichment methods SVM-E and NMTL-E outperform SVM and NMTL, respectively. This is not surprising since both SVM and NMTL suffer from the insufficient training sample problem in multi-semantic spaces, while SVM-E and NMTL-E bypasses this problem by training and

Table 4: The MAUCs of different image annotation algorithms on the NUS-WIDE-Emotive for 648 Concepts.

Methods	SVM	SVM-E	NMTL	NMTL-E	MTLG	MTLE	IA-MSL
Amusement	55.7	57.9	60.0	61.2	65.7	66.1	**71.1**
Excitement	54.2	56.2	64.4	65.2	68.1	71.2	**75.4**
Awe	56.8	57.9	64.7	64.9	65.0	67.8	**69.7**
Contentment	67.0	68.9	75.1	76.4	76.4	80.9	**83.7**
Disgust	30.2	31.3	35.4	36.0	34.1	35.1	**37.0**
Anger	59.1	60.7	67.2	68.1	68.3	72.0	**77.2**
Fear	54.2	55.7	59.7	60.0	61.5	64.3	**68.9**
Sad	61.2	62.3	67.4	67.8	68.1	70.8	**73.6**
MAUC %	54.8	56.1	62.0	63.1	65.1	66.1	**69.6**

Table 5: The AUCs and MAUC of different image annotation algorithms on the NUS-WIDE-Emotive for 8 Emotive Categories.

AUC %	SVM	NMTL	MTLG	MTLE	IA-MSL
Amusement	73.0	76.0	77.9	77.9	**78.1**
Excitement	34.8	64.6	66.9	66.9	**67.2**
Awe	28.5	70.0	71.2	71.2	**72.2**
Contentment	33.2	65.2	67.1	67.0	**68.2**
Disgust	25.1	68.7	73.3	73.3	**75.8**
Anger	32.1	64.9	67.3	67.2	**69.8**
Fear	30.2	68.6	71.2	71.1	**72.7**
Sad	26.1	73.5	36.9	74.5	**75.6**
MAUC %	36.1	67.8	70.1	71.1	**73.7**

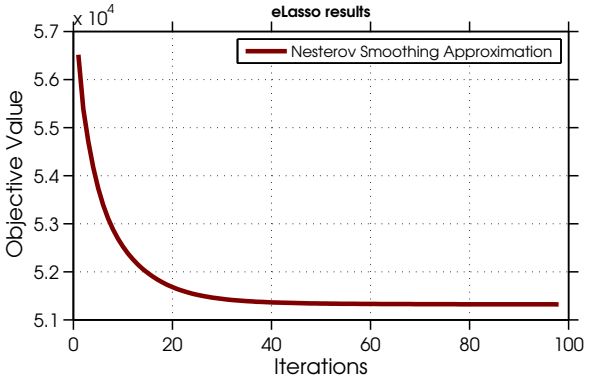

Figure 2: Convergence curve of IA-MSL on NUS-WIDE-EMOTIVE dataset.

testing in unitary space, and then fusing the results in individual spaces as final output.

To show the convergence performance of the proposed smoothing approximation optimization scheme developed in Section 3, we illustrate in Figure 2 the objective value convergence curve on NUS-WIDE-Emotive. It can be observed that the algorithm converges fast in less than 100 iterates. As a first-order information, the smoothing approximation method used in IA-MSL scales well w.r.t. the sample size N and feature dimensionality d. In our practice, a typical training time on this dataset is about 512 seconds. The per query time of IA-MSL is 0.05 second.

By setting the semantic space number $K = 1$, IA-MSL is immediately applicable to unitary semantic image annotation. We have also compared IA-MSL with baselines in Table 2. Table 5 lists the results for 8 emotive classes. Table 6 lists the corresponding results for 81 cognitive object categories. To make the table compacter, we sort the 81 concepts according to the descend order of training sample number and evenly divide them into 8 groups. The AUCs in Table 6 are obtained by averaging over each of these 8 concept groups. From the results in both tables we can see that

Table 6: The MAUCs of different image annotation algorithms on the NUS-WIDE-Emotive for 81 object concepts.

Methods	SVM	NMTL	MTLG	MTLE	IA-MSL
Group1	71.1	75.4	78.8	80.3	**86.4**
Group2	57.0	74.5	78.1	79.6	**85.7**
Group3	53.7	76.2	79.1	80.4	**86.4**
Group4	54.3	79.1	82.3	83.8	**89.9**
Group5	40.1	72.4	74.8	76.3	**84.3**
Group6	35.0	75.0	78.3	79.9	**86.3**
Group7	25.1	75.6	79.1	80.6	**86.8**
Group8	9.1	72.6	76.0	77.5	**83.4**
MAUC %	42.7	75.1	78.5	80.2	**86.1**

Table 7: The unitary semantic annotation results on NUS-WIDE-LITE.

Methods	SVM	LNP	EGSSC	LSMP	IA-MVL
MAUC	38.5	74.5	75.0	78.3	81.5

IA-MSL also outperforms the baselines for unitary semantic annotation. Moreover, we also compare IA-MSL with several representative unitary semantic image annotation algorithms on NUA-WIDE-LITE as listed in Table 7. It can be seen that our method outperforms the state-of-the-arts methods.

One direct application of IA-MSL is real world image retrieval with multi-semantic query words. On NUS-WIDE-Emotive, by inputting the emotive-cognitive query word "Amusement Dog", the returned top 6 ranked images by IA-MSL, NMTL and SVM are shown in Figure 3. From this example we can see that IA-MSL is more accurate than the other two for multi-semantic image retrieval.

4.4 Experiment-II: NUS-WIDE-Object &Scene

On this dataset, we category all labels into three setting: 33 scene classes, 31 object classes and 1023 (33 scene × 31 concepts) combination classes. The ground truth of 1023 labels is also derived by Cartesian product of 33 scene labels and 31 object labels. Again, some of these 1023 multi-semantic labels suffer from the issue of insufficient training samples. We compare IA-MSL with six baseline algorithms as shown in Table 2. Table 8 lists the quantitative results. To make the results more compactly shown, we sort the 1033 concepts according to the descent order of training sample number and evenly divide them into 5 groups. The AUCs in Table 8 are obtained by averaging over each of these 5 concept groups. As can be observed that IA-MSL outperforms the competing methods in MAUC and AUCs on all the 5 concept groups. It is noteworthy that on Group 5, all the involved comparing algorithms return AUC 0. This is unsurprising since Group 5 is composed by those concepts with very few or even zero training samples, and thus all the

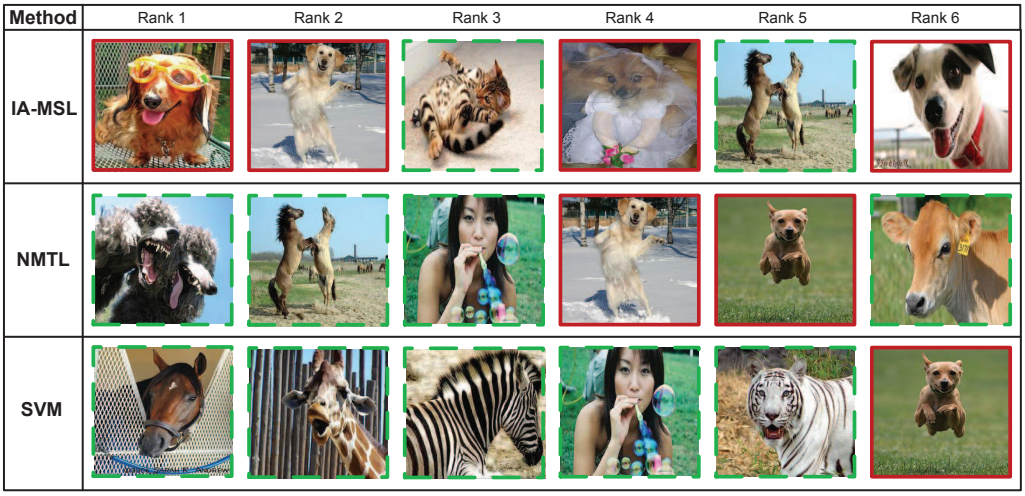

Figure 3: Some exemplar results of query search and ranking by IA-MSL (top row), NMTL (middle row) and SVM (bottom row) on NUS-WIDE-Emotive with the query: "Amusement Dog". The red border indicates correct result while the green one incorrect.

Table 9: The MAUCs of different image annotation algorithms on the NUS-WIDE-Object&Scene for 31 object concepts.

Methods	SVM	NMTL	MTLG	MTLE	IA-MSL
Group1	71.2	72.6	77.8	79.9	**87.4**
Group2	58.9	71.6	76.5	78.6	**84.1**
Group3	40.4	75.1	80.1	82.2	**88.7**
Group4	21.3	75.3	80.3	82.4	**87.9**
Group5	10.1	74.8	79.8	81.3	**87.0**
MAUC %	44.5	73.8	78.9	81.0	**87.5**

Table 10: The MAUCs of different image annotation algorithms on the NUS-WIDE-Object&Scene for 33 scene concepts.

Methods	SVM	NMTL	MTLG	MTLE	IA-MSL
Group1	70.0	72.4	78.8	81.5	**87.0**
Group2	57.1	59.6	64.5	67.2	**83.7**
Group3	39.8	73.8	79.3	82.0	**88.1**
Group4	19.1	72.5	78.9	81.1	**87.6**
Group5	9.0	72.1	77.6	80.3	**86.8**
MAUC %	43.3	72.3	77.8	80.5	**87.1**

algorithms fail including. A typical running time for training on this dataset is about 470 seconds seconds.. The per query time of IA-MSL is 0.08 second.

Specially, in the setting of unitary semantic image annotation, we have also compared IA-MSL with the algorithms listed in Table 3. Table 9 and Table 10 list the corresponding results for 31 objects and 33 scenes, respectively. For the same purpose of making the results compactly shown, we sort both the 31 objects and 33 scenes according to the descent order of training sample number and evenly divide each of them into 5 groups. The AUCs are obtained by averaging over each of these 5 groups. From the results in Table 9 & 10 we observe again that IA-MSL also outperforms the baselines for unitary semantic annotation.

5. CONCLUSION

In this paper, we proposes the IA-MSL method to explore multi-semantic meaning of images based on two or more semi-orthogonal label views from multi-semantic. We formulate this challenging problem as a multi-task discriminative analysis model, where individual tasks are defined by learning of linear discriminative model for individual complex semantic concepts. We consider all the tasks in a joint manner by imposing two types of regularization, the graph Laplacian regularization and exclusive group lasso regularization. A Nesterov-type smoothing approximation method is developed for model optimization. The proposed algorithm is experimented on two image benchmarks built for multi-semantic annotation. We validate the superiority of IA-MSL in terms of both accuracy and efficacy. In future, we can attach a few sub-categories to each category of the above 8 Emotive Categories to expand our search range towards real world search scenario.

Appendix

A. PROOF OF PROPOSITION 1

PROOF. Since $0 \in \{V^i : \|V^i\|_{2,\infty} \leq 1\}$, by (7) we get that for $k = 1, 2$:

$$0 \leq q_{\mathcal{G}^k,\mu}(W^i) \leq \max_{\|V^{i,k}_{\mathcal{G}^k}\|_{2,\infty}\leq 1} \langle W^i, V^{i,k}\rangle = \|W^i_{\mathcal{G}^k}\|_2. \quad \text{(A.1)}$$

Therefore by definition of Ω in (3) we get the validity of the first inequality in (8). Since $\|V^i_{\mathcal{G}^k}\|_{2,\infty} \leq 1$,

$$q_{\mathcal{G}^k,\mu}(W^i) \geq \max_{\|V^{i,k}_{\mathcal{G}^k}\|_{2,\infty}\leq 1} \langle W^i, V^{i,k}\rangle - \frac{\mu}{2} = \|W^i_{\mathcal{G}^k}\|_{2,1} - \frac{\mu|\mathcal{L}^k|^2}{2}. \quad \text{(A.2)}$$

Combining (A.1) and (A.2) we get

$$\left| q_{\mathcal{G}^k,\mu}(W^i) - \|W^i_{\mathcal{G}^k}\|_{2,1} \right| \leq \frac{|\mathcal{L}^k|^2\mu}{2},$$

Table 8: The MAUCs of different image annotation algorithms on the NUS-WIDE-Object&Scene for 1023 Concepts.

Methods	SVM	SVM-E	NMTL	NMTL-E	MTLG	MTLE	IA-MSL
Group1	61.3	62.5	79.8	81.2	82.5	84.6	**86.7**
Group2	50.0	51.9	65.8	67.2	71.3	72.4	**78.7**
Group3	41.2	42.1	50.5	52.1	55.0	56.5	**75.8**
Group4	5.3	5.5	5.6	6.1	6.2	7.3	**13.0**
Group5	0	0	0	0	0	0	**0**
MAUC %	38.0	40.2	47.2	48.6	51.0	52.5	**61.3**

Thus

$$\begin{aligned} & \left| q^2_{\mathcal{G}^k,\mu}(W^i) - \|W^i_{\mathcal{G}^k}\|^2_{2,1} \right| \\ = & \left| q_{\mathcal{G}^k,\mu}(W^i) - \|W^i_{\mathcal{G}^k}\|_{2,1} \right| \cdot \left| q_{\mathcal{G}^k,\mu}(W^i) + \|W^i_{\mathcal{G}^k}\|_{2,1} \right| \\ \leq & \frac{|\mathcal{L}^k|^2\mu}{2} 2\|W^i_{\mathcal{G}^k}\|_{2,1} \leq \sqrt{2}\mu|\mathcal{L}^1|\|W^i\|_2 \leq \sqrt{2}\mu|\mathcal{L}^k|^2 R, \end{aligned}$$

which implies that

$$q^2_{\mathcal{G}^k,\mu}(W^i) \geq \|W^i_{\mathcal{G}^k}\|^2_{2,1} - \sqrt{2}\mu|\mathcal{L}^k|^2 R.$$

By summarizing both sides of the preceding inequality for $k = 1, 2$ over $i = 1, ..., d$, we get the validity of the second inequality in (8). □

B. PROOF OF THEOREM 1

PROOF. Fix an $i \in \{1, ..., d\}$. Analog to the standard analysis and results (see, e.g. [25, Theorem 1]) we can derive that $q_{\mathcal{G}^k,\mu}(W^i)$, $k = 1, 2$, is well defined and continuously differentiable with gradients given by

$$\nabla q_{\mathcal{G}^k,\mu}(W^i) = V^{i,k}(W^i),$$

which is Lipschitz continuous with constant

$$L^i_{k,\mu} = \frac{1}{\mu}. \tag{B.1}$$

By chain rule of derivative we get that for $k = 1, 2$,

$$\frac{1}{2}\nabla q^2_{\mathcal{G}^k,\mu}(W^i) = q_{\mathcal{G}^k,\mu}(W^i)V^{i,k}(W^i),$$

which proves the (10), and consequently (9).

To prove the Lipschitz continency of $\nabla\Omega_\mu(W)$, one may first confirm the Lipschitz continuousness of $\frac{1}{2}\nabla q^2_{\mathcal{G}^k,\mu}(W^i)$, $k = 1, 2$,

$$\begin{aligned} & \|q_{\mathcal{G}^k,\mu}(W^i)\nabla q_{\mathcal{G}^k,\mu}(W^i) - q_{\mathcal{G}^k,\mu}(U^i)\nabla q_{\mathcal{G}^k,\mu}(U^i)\|_2 \\ = & \|q_{\mathcal{G}^k,\mu}(W^i)\nabla q_{\mathcal{G}^k,\mu}(W^i) - q_{\mathcal{G}^k,\mu}(W^i)\nabla q_{\mathcal{G}^k,\mu}(U^i) \\ & + q_{\mathcal{G}^k,\mu}(W^i)\nabla q_{\mathcal{G}^k,\mu}(U^i) - q_{\mathcal{G}^k,\mu}(U^i)\nabla q_{\mathcal{G}^k,\mu}(U^i)\|_2 \\ \leq & |q_{\mathcal{G}^k,\mu}(W^i)| \cdot \|\nabla q_{\mathcal{G}^k,\mu}(W^i) - \nabla q_{\mathcal{G}^k,\mu}(U^i)\|_2 \\ & + \|\nabla q_{\mathcal{G}^k,\mu}(U^i)\|_2 \cdot |q_{\mathcal{G}^k,\mu}(W^i) - q_{\mathcal{G}^k,\mu}(U^i)| \\ \leq & \left(\frac{\sqrt{2}R}{\mu} + |\mathcal{L}^k|^2 \right) \|W^i - U^i\|_2 \end{aligned} \tag{B.2}$$

where the last equality follows the basic facts: (i) constant in (B.1), (ii) $|q_{\mathcal{G}^k,\mu}(W^i)| \leq \|W^i_{\mathcal{G}^k}\|_{2,1} \leq \sqrt{2}R$, (iii) $\|\nabla q_{\mathcal{G}^k,\mu}(U^i)\|_2 = \|V^{i,k}(U^i)\|_2 \leq |\mathcal{L}^k|$, and (iv) $|q_{\mathcal{G}^k,\mu}(W^i) - q_{\mathcal{G}^k,\mu}(U^i)| \leq \|\mathcal{L}^k\|\|W^i - U^i\|_2$ (due to the boundness of $\nabla q_{g,\mu}$ in (iii)). By combining (6) and (B.2) we establish the validity of (11). □

C. REFERENCES

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73 (3):243–272, 2008.

[2] S. Becker, J. Bobin, and E. Candés. Nesta: a fast and accurate first-order method for sparse recovery. *SIAM J. on Imaging Sciences, submitted*, 2009.

[3] R. Caruana. Multi-task learning. *Machine Learning*, 28:41–75, 1997.

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

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

[6] R. Cilibrasi and P. M. B. Vitanyi. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19:370–383, 2007.

[7] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *SIGKDD*, 2004.

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

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

[10] A. Hanjalic. Extracting moods from pictures and sounds: towards truly personalized tv. *Signal Processing Magazine*, 23(2):90–100, 2006.

[11] A. Hanjalic. M. s. lew and n. sebe and c. djeraba and r. jain. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):1–19, 2006.

[12] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29ÍC36, 1982.

[13] T. Hayashi and M. Hagiwara. Image query by impression words-the iqi system. *IEEE Transactions on Consumer Electronics*, 44:347–352, 1998.

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

[15] N. Jorge and W. S. J. *Numerical Optimization*. Springer-Verlag, 2006.

[16] K. Kesorn. Multi-model multi-semantic imagee retrieval. 2010.

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

[18] M. Kowalski and B. Torreesani. Sparsity and persistence: mixed norms provide simple signals models with dependent coefficient. *Signal, Image and Video Processing*, doi:10.1007/s11760-008-0076-1, 2008.

[19] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H. jiang Zhang. Tag ranking. In *WWW*, 2009.

[20] H. Liu, M. Palatucci, and J. Zhang. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *ICML*, pages 649–656, 2009.

[21] Y. Liu, R. Jin, and L. Yang. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In *AAAI*, 2006.

[22] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM MM*, 2010.

[23] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior Research Methods*, 37(4):626–630, 2005.

[24] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer, 2004.

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

[26] W. ning Wang, Y. lin Yu, and S. ming Jiang. Image retrieval by emotional semantics: A study of emotional space and feature extraction. In *IEEE Int. Conf. on Systems, Man and Cybernetics*, 2006.

[27] G. Obozinski, B. Taskar, and M. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Journal of Statistics and Computing*, 20:231–252, 2009.

[28] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang. Correlative multi-label video annotation. In *MM*, 2007.

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

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

[31] N. Ueda and K. Saito. Parametric mixture models for multilabeled text. In *NIPS*, 2002.

[32] F. Wang and C. Zhang. Label propagation through linear neighborhoods. In *ICML*, 2006.

[33] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In *MM*, 2008.

[34] Q. Wu, C. Zhou, and C. Wang. Content-based affective image classification and retrieval using support vector machines. *Affective Computing and Intelligent Interaction*, 37(84):239–247, 2005.

[35] V. Yanulevskaya, J. C. van Gemert, K. Roth, A. K. Herbold, N. Sebe, and J. M. Geusebroek. Emotional valence categorization using holistic image features. In *IEEE Int. Conf. on Image Processing*, 2008.

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

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

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

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

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