Multi-Label Learning for Semantic Image Annotation

CHEN XIANGYU

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

NUS GRADUATE SCHOOL FOR INTEGRATIVE

SCIENCES AND ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2013

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Name: CHEN XIANGYU Chen Xiangyu

Date: July 07, 2013

Acknowledgments

This thesis is the result of four years of work. It would have not been possible, or at least not what it looks like now, without the guidance and help of many people. It is now my great pleasure to take this opportunity to thank them.

Foremost, I would like to show my sincere gratitude to my advisor, Prof. Tat-Seng Chua, who has been instrumental in ensuring my academic, professional, financial, and moral well being ever since. He has supported me throughout my research with his patience and knowledge. For the past four years, I have appreciated Prof. Chua's seemingly limitless supply of creative ideas, insight and ground-breaking visions on research problems. He has offered me with invaluable and insightful guidance that directed my research and shaped this dissertation without constraining it. As an exemplary teacher and mentor, his influence has been truly beyond the research aspect of my life.

I also thank my co-advisor, Prof. Shuicheng Yan. I thank him for his patience, encouragement and constructive feedback on my research work, and for his insights and suggestions that helped to shape my research skills. His visionary thoughts and energetic working style have influenced me greatly. During my Ph.D pursuit, Prof. Yan has always been providing insightful suggestion and discerning comments to my research work and paper drafts. His suggestion and guidance have helped to improve my research work.

During my Ph.D pursuit, many lab mates and colleagues have helped me. I like to thank Yantao Zheng, Guangda Li, Bingbing Ni, Richang Hong, Jinhui Tang, Yadong Mu and Xiaotong Yuan for the inspiring brainstorming, valuable suggestion and enlightening feedbacks on my work.

I would like to thank my family, my parents Lixiang and Huanying, and my wife Yue Du. For their selfless care, endless love and unconditional support, my gratitude to them is truly beyond words.

Finally, I would like to thank everybody who was important to the successful realization of thesis, as well as expressing my apology that I could not mention personally one by one. Thank you.

Contents

Summary

With the popularity of photo sharing websites, new web images on a wide variety of topics have been growing at an exponential rate. At the same time, the contents of images are also enriched and more diverse than ever before. This brings about two main challenging problems in semantic image annotation: 1) the semantic space of image dataset is enlarged and may contain two or more semantic spaces; 2) the trend of image corpus is towards large-scale or web-scale setting, which is generally unaffordable for traditional annotation approaches.

To address the first challenging problem, this thesis proposes multi-label learning algorithms for semantic image annotation from two paradigms: multi-label learning on single-semantic space and multi-label learning on multi-semantic space. For the first paradigm, different from most existing works that motivated from label co-occurrence, we propose a novel Label Exclusive Linear Representation (LELR) model for image annotation, which incorporates a new type of context–*label exclusive context*. In the setting of multi-label learning problems, when the number of categories is large, we may expect negative correlations among categories. Given a set of exclusive label groups that describe the negative relationship among class labels, our proposed method enforces exclusive assignment of the labels from each group to a query image. For the second paradigm, we propose a multi-task linear discriminative model for harmoniously integrating multiple semantics, and investigating the problem of learning to annotate images with training images labeled in two or more correlated semantic spaces, such as *fascinating nighttime*, or *exciting cat*. Image semantics can be viewed at two levels: Cognitive level and Affective level. The two spaces of image semantics are inter-related and can be used together to reinforce each other in order to improve the accuracy of concept detection and in particular, to detect complex concepts involving both types of basic concepts.

To address the second challenging problem, this thesis proposes an efficient sparse graph based multi-label learning scheme for large-scale image annotation, whereby both the efficacy and accuracy are further enhanced. In order to annotating large-scale image corpus, we perform the multi-label learning on the so-called hashing-based ℓ_1-graph, which is efficiently derived with Locality Sensitive Hashing approach followed by sparse ℓ_1-graph construction within the individual hashing buckets. Unlike previous large-scale approaches that propagate over individual label independently, our proposed large-scale multi-label propagation (LSMP) scheme encodes the tag information of an image as a unit label confidence vector, which naturally imposes inter-label constraints and manipulates labels interactively. It then utilizes the probabilistic Kullback-Leibler divergence for problem formulation on multi-label propagation.

To demonstrate the advantages and utility of our algorithms, extensive experiments on the challenging real-world benchmarks are provided for each proposed multi-label learning method. We compare each proposed approach to the state-of-the-art methods, as well as offer insights into individual result. The promising performance well validate the effectiveness of the proposed approaches. In the end, some limitations and broad vision for multi-label learning are also discussed.

List of Figures

List of Tables

List of Publications

- Xiangyu Chen, Xiaotong Yuan, Shuicheng Yan, Yong Rui and Tat-Seng Chua. 2011. Towards Multi-Semantic Image Annotation with Graph Regularized Exclusive Group Lasso. In *ACM International Conference on Multimedia.* (Full Paper)

- Xiangyu Chen, Xiaotong Yuan, Shuicheng Yan, and Tat-Seng Chua. 2011. Multi-label Visual Classification with Label Exclusive Context. In *International Conference on Computer Vision.* (Full Paper)

- Xiangyu Chen, Yadong Mu, Shuicheng Yan, and Tat-Seng Chua. 2010. Efficient Large-Scale Image Annotation by Probabilistic Collaborative Multi-Label Propagation. In *ACM International Conference on Multimedia.* (Full Paper)

- Xiangyu Chen, Yadong Mu, Hairong Liu, Yong Rui, Shuicheng Yan and Tat-Seng Chua. 2013. Efficient Large-Scale Image Annotation based on Sparse Induced Graph Construction. Minor Revision on *ACM Transactions on Multimedia Computing, Communications and Applications.*

- Xiangyu Chen, Jin Yuan, Liqiang Nie, Zheng-Jun Zha, Shuicheng Yan and Tat-Seng Chua. 2010. TRECVID 2010 Known-item Search by NUS. *TREC Video Retrieval Evaluation Online Proceedings.*

- Jian Dong, Xiangyu Chen, Tat-Seng Chua and Shuicheng Yan. 2012. Robust Image Annotation via Simultaneous Feature and Sample Outlier Pursuit. Accepted by *ACM Transactions on Multimedia Computing, Communications and Applications.*

- Yadong Mu, Xiangyu Chen, Shuicheng Yan, and Tat-Seng Chua. 2011. Learning Reconfigurable Hashing for Diverse Semantics. In *ACM International Conference on Multimedia Retrieval.* (Oral Paper)

- Yadong Mu, Xiangyu Chen, Xianglong Liu, Tat-Seng Chua, Shuicheng Yan. 2011. Multimedia Semantics-Aware Query-Adaptive Hashing with Bits Reconfigurability, *International Journal of Multimedia Information Retrieval.*

- Yantao Zheng, Shi-Yong Neo, Xiangyu Chen and Tat-Seng Chua. 2009. VisionGo: towards true interactivity. In *ACM International Conference on Image and Video Retrieval.*

Chapter 1

Introduction

1.1 Background

1.1.1 Semantic Image Annotation

For image annotation, the main task is to assign semantic keywords to an image in order to reflect its semantic content. Due to the rapid development of digital photography and the popularity of photo sharing websites, the digital images are increasing in an explosive way. Robust browsing and retrieval of these huge amount of images via semantic keywords is becoming a critical requirement. In the real world, most Internet image search engines efficiently utilize text-based search to satisfy the queries of users, while not exploiting the visual content of images. Utilizing visual content to annotate images with a richer and more relevant set of semantic keywords would allow one to further exploit the fast indexing and retrieval architecture of these search engines, which boosts the search performance at the same time. This makes the problem of annotating images with relevant semantic keywords increasingly important.

In the field of semantic image annotation, one of the main challenges is the well-known “semantic gap” problem, which points to the fact that it is hard to bridge the gap between low level feature and high-level human perception. Humans tend to use high-level semantic concepts (e.g., keywords, text descriptors) to interpret image content and measure their similarity. While the visual features extracted utilizing computer vision techniques are mostly low-level features, such as color, shape, texture, etc. Though a large amount of research has been carried out on designing algorithms to extract effective visual features in the past two decades, these algorithms cannot adequately model image semantics and have many limitations when dealing with broad content image databases [Mojsilovic and Rogowitz, 2001]. Therefore, to satisfy user’s expectations and support query by high-level concepts, a large number of machine learning techniques for bridging the “semantic gap” have been applied along with a great deal of research efforts.

Given the set of semantically labelled training images that are represented with low level features, a machine learning algorithm can be trained to utilize the visual feature to perform semantic label matching. Once trained, the algorithm can be used to label new images. There are generally two types of semantic image annotation approaches: single-label learning and multi-label learning for image annotation. In a single-label setting [Shotton et al., 2006], each image will be categorized into one semantic label and only one of the predefined label categories. In other words, only one label will be assigned on each image in this setting. In a multi-label setting [Boutell et al., 2004; Kang, Jin, and Sukthankar, 2006], which is more challenging but much closer to real world applications, each image will be assigned with one or multiple labels from a predefined label set. This thesis focuses on multi-label learning (MLL) for image annotation.

1.1.2 Single-Label Learning for Semantic Image Annotation

For single-label learning algorithms, firstly, low level visual features are extracted from image, and then the features are considered as input to a conventional binary classifier which indicates which concept category it belongs to. Finally, the output of the classifier is the semantic concept which is assigned for image annotation. In a single-label learning setting, once the images are classified into different categories, each image is only annotated with one category concept such as bus, tree, building etc. The common algorithms for single-label learning annotation basically include three types: support vector machines(SVM) [Vapnik, 1995], artificial neural network(ANN) [Frate et al., 2007],and decision tree(DT) [Quinlan, 1986a].

Based on this single-label learning annotation, retrieval of images in the search engine is straightforward by just typing in keywords related to the concept labels . The main advantage of this type of approach is that searching of images is efficient because the search engine needs not to do usual image indexing and expensive on-line matching. However, this type of approach ignores the fact that many images may contain multiple semantic concepts. As a result, many relevant images may be missed from the retrieval list if a user does not search using the exact keyword. One effective way to alleviate this problem is to annotate each image with multiple keywords in order to reflect different semantics contained in the image. This motivates semantic image annotation focusing on multi-label learning for improving the search performance.

1.2 Multi-Label Learning for Semantic Image Annotation

Conventional single-label learning methods for image annotation usually consider an image as an entity associated with only one label in model learning stage. These single-label learning algorithms may sound attractive and straightforward, but they overlook the fact that a real-world image usually contains multiple semantic concepts rather than a single one. In most real-world problems, multiple labels can be assigned to an image. In many online image sharing websites (e.g. Picasa, Flickr, and Yahoo! Gallery), most of the images have more than one tags. For example, an image can be annotated as "road" as well as "car", where the terms "road" and "car" are in different categories. Furthermore, the traditional methods lack a mechanism to rank images according to their similarity to the annotated label. Owing to the great potential of automatically tagging images with related labels, multi-label image annotation is becoming increasingly important and is a more reasonable approach for real-world image annotation, because it assigns an image to several categories and assigns an image to a category with a confidence value which assists in image ranking. This dissertation mainly investigates multi-label learning for semantic image annotation.

The most commonly-used approach for multi-label learning is to divide it into multiple binary classification problems [Chang, K. Goh, and CBSA, 2003; Yan, Tesic, and Smith, 2007], and determine the labels for each test sample by aggregating the classification results from all the classifiers. However, there are three main disadvantages of this type of approach: 1) It assumes each class label independently so that it is not able to utilize the correlation information of labels to boost the performance; 2) It is cannot be employed for annotating images with a large number of classes because each class requires a binary classifier for training; 3) Most binary classification approaches toward multi-label learning suffer severely from the unbalanced data problem [Weiss and Provost, 2003], particularly when the number of classes is large. Given image dataset, once the number of classes is large, the number of negative samples is overwhelmingly larger than the number of positive samples for every class. As a result, most of trained binary classifiers will assign the negative labels to test images. This motivates many researchers to exploit machine learning algorithms for multi-label learning. The detailed related works of multi-label learning will be reviewed in Chapter 2.

Due to the explosive growth of digital technologies, new images on a large variety of topics have been growing at an exponential rate. And the contents in images are enriched and more diverse than ever before. This brings about two main challenges in multi-label learning: (a) the semantic space of image data is enlarged and contains one or more semantic spaces, where there may been multiple semantic spaces included in an image dataset (e.g. cognitive semantic space and emotive semantic space); and (b) the image corpus for annotation is towards to large-scale or web-scale setting, which is generally infeasible for traditional annotation approaches. According to the above mentioned two challenging problems, this thesis focuses on exploiting the semantic multi-label learning from three aspects: (a) multi-label learning on traditional single-semantic space, (b) multi-label learning on multi-semantic space, and (c) multi-label learning in large-scale dataset. For the first challenge, multi-label learning with label exclusive context in single semantic space is first proposed and explored in Chapter 3, then an extension version towards multi-semantic space for multi-label image annotation is proposed and discussed in Chapter 4. For the second challenge, a graph-based semi-supervised multi-label learning approach for large-scale image annotation is exploited in Chapter 5, which is founded on hashing-based l_1 graph construction and Kullback-Leibler divergence based label similarity measurement.

1.2.1 Multi-Label Learning with Label Exclusive Context

Since many words are semantically related, labels in image dataset are usually correlated. This correlation among labels are helpful for predicting labels of test images. For example, the concepts “lake” and “boat” usually appear in the same image. When assigning a label “boat” to a test image, this image may contain the label “lake”. so they are correlated concepts. It is reasonable to make use of such a correlated context of labels for predicting class labels of the query image sample. In the past, many researcher have explored the co-occurrent label context in multi-label learning for image annotation [Zhu et al., 2005; Yu et al., 2005; McCallum, 1999].

In order to further improve the performance of image annotation, we propose a novel Label Exclusive Linear Representation (LELR) method for multi-label image annotation. Unlike the past research efforts based on co-occurrent information of labels, we incorporate a new type of label context named *label exclusive context* into the LELR scheme, which describes the negative relationship among class labels. Given a set of exclusive label groups that describe the negative relationship among class labels, the proposed LELR enforces repulsive assignment of the labels from each group to a test image. Extensive experiments on the challenging real-world benchmarks demonstrate the effectiveness of embedding this new context into multi-label learning scheme.

1.2.2 Multi-Label Learning on Multi-Semantic Space

In order to manege the huge amount and 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 offers a series of challenges in media content processing techniques. The *semantic gap* [Lew et al., 2006] between high-level semantics and low-level image features is still one of the main challenging problems for image classification and retrieval. Moreover, image semantics can be viewed at two levels: Cognitive level and Affective level [Hanjalic, 2006]. The two spaces of image semantics are inter-related and should be used together to reinforce each other in order to improve the accuracy of concept detection and in particular, to detect the complex concepts involving both types of basic concepts.

However, existing studies on image semantic annotation mainly aim at the assignment of either the cognitive concepts or affective concepts to a new item separately. Moreover, they fail to take into consideration the correlation between concepts from different spaces. For example, certain cognitive concepts (such as *snake* and *tiger*) are usually attached with negative emotion, while other concepts (such as *beach* and *sunset*) are associated with positive emotions. As a result, the complex concepts consisting of concepts from different spaces cannot be inferred easily. For detecting these complex concepts, the current learning process requires a huge amount of efforts in extracting different types of cognitive and emotive features and is thus generally unaffordable for large-scale image dataset. Moreover, it is hard to generate concepts from different semantic spaces simultaneously because they require the use of different techniques to be applied to different semantic spaces, and the aggregation of results of individual concepts from different spaces is usually unable to model the meanings of complex query in the real-world search task. This motivates us to harmoniously embed these two or more semantic spaces into one general framework for annotating the deeper and multi-semantic labels to images. In this thesis, we are particularly interested in explicit multi-semantic [1] image annotation under the unified generic visual features. This framework not only works well on cognitive and affective spaces but can also be applied to other multi-space semantics such as object and scene.

1.2.3 Multi-Label Learning in Large-Scale Dataset

The last decade has witnessed a growing interest in image annotation. In many real world scenario cases, we often face the challenging situation that there is no sufficient labeled data whereas large numbers of unlabeled image data may could be far easier to be crawled on the web. And annotating this large-scale unlabeled data often requires the employment of a huge number of experienced human annotators and consuming much time, which directly motivates recent development of large-scale semi-supervised learning (SSL) methods [Zhu, 2006; Subramanya and Bilmes, 2009]. With the small amount of labeled image data, SSL makes itself as an effective annotation technique through working together with other unlabeled data for learning and inference.

For image annotation, a graph is often employed as an effective representation for label propagation in large-scale setting, wherein all images of the entire dataset are expressed as vertices and edges reflecting similarity between the im-

[1] The *multi-semantic* (or *polysemy*) retrieval has been explored in [Kesorn, 2010] 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” object can belong to high jump or pole vault event. Differently, the term multi-semantic used in this chapter emphasizes that an image can be labeled in multiple semantic spaces.

ages. For generative modeling methods, the priori probabilistic assumptions usually play an import role for propagation. Different from this body of generative modeling work, graph-based modelings are especially interested in non-parametric and discriminative local structure discovery with the assumption that the larger the weight of edge connecting vertices, the higher the possibility of sharing the similar labels between the images. And it is also demonstrated that graph-based approaches are usually able to achieve the state-of-the-art performance as compared to other SSL algorithms [Zhu, 2006]. In this thesis, we propose an efficient semi-supervised large-scale multi-label learning approach based on hashing-accelerated ℓ_1-graph construction.

1.3 Thesis Focus and Main Contributions

The overall objective of this thesis is to develop methodologies for multi-label learning image annotation from three aspects: 1) exploiting label exclusive context for multi-label learning on traditional single semantic space; 2) developing multi-task linear discriminative model for multi-label learning on multi-semantic space; and 3) utilizing hashing based sparse ℓ_1-graph construction to exploit multi-label learning annotation in large-scale image dataset. Three major contributions are made in this dissertation.

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

2) Multi-Label Learning on Multi-Semantic Space: To exploit the comprehensive semantic of images, we propose a general framework for harmoniously integrating the above multiple semantics, and investigating the problem of learning to annotate images with training images labeled in two or more correlated semantic spaces. 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 space, our proposed multi-semantic annotation associates each image with labels from multiple semantic spaces. 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 Object&Scene datasets from NUS-WIDE well validate the effectiveness of the proposed approach.

3) Multi-Label Learning in Large-Scale Image Dataset: Motivated by recent development of semi-supervised or active annotation methods, we develop a novel large-scale multi-label learning scheme, whereby both the efficacy and accuracy of large-scale image annotation are further enhanced. Our proposed scheme outperforms the state-of-the-art algorithms by making the following contributions. 1) Unlike previous approaches that propagate over individual label independently, our proposed large-scale multi-label propagation (LSMP) scheme encodes the tag information of an image as a unit label confidence vector, which naturally imposes inter-label constraints and manipulates labels interactively. It then utilizes the probabilistic Kullback-Leibler divergence for problem formulation on multi-label propagation. 2) We perform the multi-label propagation on the so-called hashing-based ℓ_1-graph, which is efficiently derived with Locality Sensitive Hashing approach followed by sparse ℓ_1-graph construction within the individual hashing buckets. 3) An efficient and convergency provable iterative procedure is presented for problem optimization. Extensive experiments on NUS-WIDE dataset (both lite version with $56k$ images and full version with 270k images) well validate the effectiveness and scalability of the proposed approach.

1.4 Organization of the Thesis

The detailed organization of this dissertation is as follows.

Chapter 2 gives a comprehensive review of the related works on single-label learning image annotation, multi-label learning image annotation on single-semantic space, and semi-supervised learning on large-scale dataset.

Chapter 3 presents a label exclusive context based multi-label learning framework for semantic image annotation, which is formulated as an exclusive Lasso (eLasso) model. Extensive evaluations of the framework on the challenging real-world visual classification benchmarks are given.

Chapter 4 further introduces a multi-label learning framework on multi-semantic space, which is a multi-task linear discriminative model to learn a linear mapping from features to labels. Extensive evaluations of the framework on NUS-WIDE-Emotive dataset ($56k$ images) with 8×81 emotive cognitive concepts and Object&Scene datasets from NUS-WIDE are given.

Chapter 5 introduces hashing-based ℓ_1-graph construction for large-scale multi-label image annotation, which utilizes the probabilistic Kullback-Leibler divergence for problem formulation on multi-label learning. Extensive evaluations of the framework on NUS-WIDE dataset (both lite version with $56k$ images and full version with 270k images) are given.

Chapter 6 concludes the thesis with highlight of contributions of this thesis, and discusses future research directions.

Chapter 2

Literature Review

With the proliferation of digital photography, semantic image annotation becomes increasingly important. Image Annotation is typically formulated as a single-label or multi-label learning problem. This chapter serves to introduce the necessary background knowledge and related works of single-label learning, multi-label learning and semi-supervised learning before delving deep into the proposed models of multi-label learning for semantic image annotation.

2.1 Single-Label Learning for Semantic Image Annotation

In semantic image annotation, single-label learning methods usually consider an image as an entity associated with only one label in model learning stage. The common algorithms for single-label learning annotation basically include three types: support vector machines(SVM), artificial neural network(ANN), and decision tree(DT). In the following, we introduce representative works and necessary background knowledge of each of these techniques.

2.1.1 Support Vector Machines

The SVM method comes from the application of statistical learning theory to separating hyperplanes for binary classification problems [Cortes and Vapnik, 1995]. The central idea of SVM is to adjust a discriminating function and find a hyperplane from a training set of image samples to separate the training dataset. In SVM methods, each training sample is represented with a feature vector and a class label. Training a SVM classifier consists in searching for the hyperplane that leaves the largest number of image samples of the same class on the same side, while maximizing the distance of both classes from the hyperplane. SVM is a supervised classifier. And it has been shown with high effectiveness in high dimensional data classifications,especially when the training dataset is small [Vapnik, 1995]. The advantage of SVM over other classifiers is that it can achieve optimal class boundaries by finding the maximum distance between classes. It has been widely employed to solve the classification problems, such as text classification, object detection and image annotation.

Although SVMs are mainly designed for the discrimination of two classes, they can be adapted to multi-class (single-label learning) problems. A multi-class SVM classifier can be obtained by training several classifiers and combining their results. In the training phase, a separate SVM classifier for each concept is trained and each SVM will generate a probability value for a input sample. During the testing phase, the decisions from all classifiers are combined and fused to assign the final class label to a test image. In the past two decades, SVM is successfully applied to image annotation. For example, Chapelle *et al.* [Chapelle, Haffner, and Vapnik, 1999] utilize the above combined SVM framework to train SVM classifiers for 14 semantic concepts. In their work, images are represented with HSV histogram. Each trained classifier is regarded as "one vs. all" classifier. In the testing stage, each SVM classifier generates a probabilistic value. The class with maximum probability is finally considered as the label of the test image. In the work of [Shi et al., 2004a], the authors use SVM to learn the semantic concepts for image regions, where the images are first segmented using k-means algorithms, and 23 SVM classifiers are trained to learn the 23 region level concepts.

2.1.2 Artificial Neural Network

Artificial Neural Networks (ANN) started playing a important role in the field of remote sensing. Since the early nineties, several studies focused on evaluating the performance of ANNs by comparing with traditional statistical methods in remote sensing applications, and in particular in image classification. ANN is a learning network, which learns from training samples and makes decision for a test sample. It consists of multiple layers of interconnected nodes, which are also called perceptrons. Generally, an ANN is also known as multilayer perceptron (MLP).

For image annotation, the first layer of ANN is the input layer which has perceptrons equal to the dimension of input image sample. The number of perceptrons in the output layer is equal to the number of concept classes. The important and open issues are the choice of the number of hidden layers and the number of perceptrons at each hidden layer [Frate et al., 2007]. The numbers of hidden layers and perceptrons are usually selected empirically depending on the practical problems. In an ANN, the connecting edges between perceptrons of different layers are associated with weights. Each perceptron works as a processing element and is governed by an activation function. The activation function generates output based on the weights and the outputs of the perceptrons at the previous layers. For annotating a test image, ANN first learns the edge weights in the process of training, which minimizes the overall learning error. Then each output perceptron generates a confidence measure and the class associated with the maximum measure indicates the decision about the test image.

The main advantage of ANN is that the outputs of output layer perceptrons are determined by the previous layers and the connecting edges. Training ANN is not dependent on any other parameter tuning or any assumption about the feature distribution. Many researchers have applied the ANN to image annotation. Frate *et al.* [Frate et al., 2007] use the ANN for satellite image annotation. They utilized a 4-layer ANN to classify pixels of images into four categories: vegetation, asphalt, building, and bare soil. In their experiment, a network of two hidden layers is employed, where each layer consists of 20 neurons. Kim *et al.* [Shi et al., 2004b] utilize the ANN technique to classify images into object and non-object images by 3-layer ANN. They assume that the center 25% of the image significantly characterizes the content of the entire image and use this center part to represent the image. However, the performance of classification will be degraded if the object appears in the other part of the image.

2.1.3 Decision Tree

Decision Tree (DT) learning is a special type of machine learning technique. Many researchers have utilized decision tree (DT) learning to perform image classification. Given a set of training images described by a fixed set of input attributes and a known outcome for each image, a DT is built by recursively dividing the training images into non-overlapping sets, and every time the images are divided, the attribute used for the division is discarded. The procedure continues until all images of a group belonging to the same class or the tree reaches its maximum depth when no attribute remains to separate them [Quinlan, 1986b]. Finally, the above learning process produces a DT which can classify the outcome value based on the given attributes of new images. For annotating a new image, the tree is traversed from the root node to a leaf node using the attribute value of the new image. The decision of the new image is the outcome of the leaf node where the image reaches.

Unlike other classification model whose input-output relationships are difficult to describe, a DT expresses the input-output relationship using human understandable rules (e.g., if-then rules). There are mainly three types of DT algorithms in the literature: ID3 [Quinlan, 1986a], C4.5 [Quinlan, 1993], and CART [Breiman et al., 1993]. Sethi *et al.* [Sethi and Coman, 2001] utilize CART to annotate outdoor images with four classes. They partition each component of HSL colour space into eight intervals and consider each of the 24 intervals as an attribute. As a result, each image in the experiment is represented with 24 attributes. In the work of [Wong and Leung, 2008], acquisition parameters (aperture, exposure time, and focal length, etc.) are used as attributes. Since the attributes are continuous values, they adopt the C4.5 method to classify scenery images into ten semantic concepts. Different from the above mentioned algorithms which can only annotate images globally, Liu *et al.* [Liu, Zhang, and Lu, 2008] utilize DT to annotate regions of segmented images. In order to training a DT, they use weighted average of color and texture features, and develop pre-pruning and post-pruning scheme.

2.2 Multi-Label Learning for Semantic Image Annotation

Generally, image semantics are recognized at two levels: cognitive level and affective level [Hanjalic, 2006]. Many multi-label annotation algorithms are 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. In this section, we will introduce the related works of multi-label learning on single-semantic space from two aspects: multi-label learning on cognitive semantic space and multi-label learning on emotive semantic space.

2.2.1 Multi-Label Learning on Cognitive Semantic Space

Multi-label learning is a hot and promising research direction, especially on cognitive semantic space. In the following of this subsection, multi-label learning means multi-label learning on cognitive semantic space(unless specified otherwise). At the early stage of research on multi-label learning, its literature is primarily geared to text classification or bioinformatics. Therefore, besides giving review on the related works of multi-label learning for semantic image annotation, we also introduce several representative works about text classification methods based on multi-label learning scheme.

Multi-label learning methods can be mainly categorized into two different groups [Tsoumakas and Katakis, 2007]: 1) problem transformation methods, and 2) algorithm adaptation methods. The first group includes methods that are algorithm independent. They transform the multi-label learning task into multiple, independent single-label learning problems and determine the labels for each sample point by aggregating the classification results from all the classifiers. The second group includes methods that employs specific learning algorithms to handle multi-label data directly.

2.2.1.1 Problem Transformation Methods

In this section, we briefly introduce three main problem transformation methods: Binary Relevance Method, Pairwise Classification Method and Label Powerset Method.

1) Binary Relevance Method

In the problem transformation methods, the most well-known method is the binary relevance method (BR) [Godbole and Sarawagi, 2004]. BR converts the multi-label problem into multiple binary problems. Each binary classifier is then utilized to predict the association of a single label. For the classification of a new instance, BR outputs the union of the labels that are positively predicted by the classifiers.

Yan *et al.* [Yan, Tesic, and Smith, 2007] present a BR-based boosting algorithm for multi-label learning. Different from other methods, the binary classifiers are trained on subsets of the samples and attribute spaces. In the learning process, their proposed algorithm reduces the information redundancy in the label space by jointly optimizing the loss functions over all the labels. Ji *et al.* [Ji et al., 2008] introduce a general framework for extracting shared structures in a BR approach. In this framework, a common subspace is assumed to be shared among multiple labels. Although they use an approximation algorithm for the solution to the proposed formulation, the resulting method is computationally expensive. In the work of [Raez, Lopez, and Steinberger, 2004], the authors propose a BR model for solving the class-label imbalance problem. They solve the text categorisation problem by overweighting positive examples in the BR models. In a real-time environment and on large collections, they observe that classification speed can be improved with marginal effect on predictive performance by ignoring rare class labels in text dataset.

For image annotation, Chang *et al.* [Chang, K. Goh, and CBSA, 2003] propose a BR-based soft annotation procedure for providing images with semantical multiple labels. They choose Support Vector Machines (SVMs) and Bayes Point Machines for training binary classifiers. Each classifier assumes the task of determining the confidence score for a semantic label. The annotation starts with labeling a small set of training images, each with one single semantical label. An ensemble of binary classifiers is then trained for predicting label membership for test images. The trained ensemble is applied to each test image to give the image multiple soft labels, and each label is associated with a label membership factor.

Although BR method is conceptually simple and relatively fast, it constructs a decision boundary individually for each label so that this method can not explicitly model label correlations [Yan, Tesic, and Smith, 2007; Godbole and Sarawagi, 2004]. Moreover, due to the typical sparsity of labels in multi-label dataset, each binary classifier is likely to have far more negative examples than positive. The performance of BR is also be affected by class-imbalance [Raez, Lopez, and Steinberger, 2004],

2) Pairwise Classification Method

Another popular transformation method is pairwise classification (PW). The above mentioned BR method is a one-vs-rest paradigm, in which each classifier corresponds to each label in the image dataset. PW is a one-vs-one paradigm where each classifier is associated with each pair of labels [Hullermeier et al., 2008]. As a results, instead of N binary problems for BR (N is the number of labels in the dataset), $M = N(N-1)/2$ binary problems are formed in PW.

Different from BR methods, the classification in PW results in a set of pairwise preferences (which give rise more naturally to a ranking) rather than a label set prediction. PW methods are widely used in ranking schemes. Hullermeier *et al.* [Hullermeier et al., 2008] developed a ranking by pairwise comparison scheme (RPC). The proposed scheme obtains a ranking by counting the votes received by each label. Furnkranz *et al.* [Furnkranz et al., 2008] extend RPC with calibrated label ranking to create a bipartition of relevant and irrelevant labels for multi-label learning. In their proposed scheme, a virtual label partitions a ranking into relevant and irrelevant labels to form a concrete label-set prediction for any test instance.

In order to deal with the large number of classifiers in a PW scheme (quadratic with respect to N), many PW approaches utilize single-label base classifiers to improve scalability. The multi-label pairwise perceptron (MLPP) proposed in [Mencia and Furnkranz, 2008a] trains one perceptron for each possible class-label pair. Although its performance is better than related BR-based perceptron algorithm, it scales quadratically with N rather than linearly. In the work of [Mencia and Furnkranz, 2008b], the authors introduce a modified version of above MLPP which can scale to large label space by using simple perceptrons.

This modified version contains a special adaptation: rather than having to maintain the $(N(N-1))/2$ models normally associated with PW, it instead keeps all examples in memory and builds models dynamically for each prediction.

3) Label Powerset Method

Another fundamental problem transformation method is the label powerset (LP). LP considers each unique set of labels that exists in a multi-label training set as one of the classes of a new single-label classification task. For a new test image, the single-label classifier of LP outputs the most probable class, which is a set of labels.

Boutell *et al.* [Boutell et al., 2004] presented a LP-based multi-label learning scheme for scene classification, which uses multi-label training model and specific testing criteria. They use a Support Vector Machine (SVM) as a classifier. A new training strategy named cross training is utilized to build SVM classifiers. In the last they extend the SVM classifier to multi-label scene classification. However, the above proposed scheme may lead to data sets with a large number of classes and few examples per class.

Due to the fact that the scalability of LP is quite poor (quadratically with the number of distinct label sets), LP can not be widely used as an off-the-shelf method. Tsoumakas *et al.* [Tsoumakas and Katakis, 2007] develop an efficient approach RAkEL (RAndom K-labEL subsets), which constructs an ensemble of LP classifiers. In the proposed RAkEL scheme, each LP classifier is trained based on a different small random subset of the set of labels. A ranking of the labels is produced by averaging the zero-one predictions of each model per considered label. In addition, a thresholding scheme is also exploited to produce a classification.

This method has become one of the most well known in the multi-label literature.

In the above mentioned LP methods, label correlations is directly utilized in the process of learning. However, different from the binary models of BR and PW, LP model only classify new samples with label sets it has already seen in the training set. The computation and implementation of LP is also complex because it requires as many class labels in the single-label transformation as there are distinct label sets in the training data (usually, these labels are likely to be very sparse with respect to training samples.).

2.2.1.2 Algorithm Adaptation Methods

The algorithm adaptation methods utilize specific learning algorithms to solve multi-label learning problem. Many research works in recent belong to the second group such as Label Ranking, Co-occurrent Context, Label Propagation. In the following, we introduce several representative multi-label learning methods of the second group.

1) Label Ranking

Another group of algorithm adaptation approaches toward multi-label learning is label ranking. These algorithms first learn a ranking function of class labels from the labeled samples in training dataset, and then utilize the function to sort the class labels for test samples.

In the area of Bioinformatics and Text Mining, Elisseeff *et al.* [Elisseeff and Weston, 2002] present a label ranking scheme based on a large margin ranking system that shares a lot of common properties with SVMs. They develop a linear model that minimizes the Ranking Loss while having a large margin. In the work of [Schapire and Singer, 2000], Adaboost.MH and Adaboost.MR are two extensions of AdaBoost [Freund and Schapire, 1997] for multi-label learning. In the first extension, the goal of the learning algorithm is to predict all and only all of the correct labels. Thus, the learned classifier is evaluated in terms of its ability to predict a good approximation of the set of labels associated with a given document. In the second extension, the goal is to design a classifier that ranks the labels so that the correct labels will receive the highest ranks.

For image annotation, In [Liu et al., 2009], Liu *et al.* propose to rank the image tags according to their relevance with respect to the associated images by tag similarity and image similarity in a random walk model. To estimate the tag relevance to the images, they first get the initial tag relevance scores based on probability density estimation, and then apply a random walk on a tag similarity graph to refine the scores. Since all the tags have been ranked according to their relevance to the image, for each uploaded image, they find the K nearest neighbors based on low-level visual features. Finally, the top ranked tags of the K neighboring images are collected and recommended to the user.

Motivated by the fact that the existing user-provided image tags in public photo sharing websites are imprecise and incomplete, Zhu *et al.* [Zhu, Yan, and Ma, 2010] propose a tag ranking and refinement scheme in form of convex optimization which comprehensively considers the tag characteristics from the points of view of low-rank, error sparsity, content consistency and tag correlation. They use a matrix to represent the image-tag relationship. In their work, the tag refinement problem is formulated as a decomposition of the user-provided tag matrix D into a low-rank refined matrix A and a sparse error matrix E, namely $D = A + E$. Compared with existing works, the low-rank and error sparsity are firstly integrated into the optimization procedure for multi-label learning. With the assistance of constraints of content consistency and tag correlation, the proposed approach is capable of correcting imprecise tags and enriching the incomplete ones.

Compared to single-label learning approaches, one advantage of label ranking approaches is that they are superior at dealing with large numbers of classes since only a ranking function is learned to compare the relevance of individual class labels with respect to the test samples. Another advantage is that these approaches do not have the issue of unbalanced data because no binary decision has to be made regarding class labels. However, these label ranking approaches do not explicitly explore the correlated information of labels, which is similar to the single-label learning approaches.

2) Label Co-occurrent Context

The essential difference between single-label learning and multi-label learning is that labels in single-label learning are assumed to be mutually exclusive while labels in multi-label learning are often assumed to be correlated. In the context of multi-label learning, there has been many works focusing on exploiting label correlation (label co-occurrent context) for semantic image annotation.

Zhu *et al.* [Zhu et al., 2005] suggest a maximum entropy model for multi-label learning classification. They explore correlations among categories with maximum entropy method and derive a classification algorithm for multi-labelled documents. The experimental results validate that multi-labelled classification is beneficial and helpful in the model considering the correlation between labels, especially when the correlation is relatively strong.

In the work of [Ueda and Saito, 2002], a probabilistic generative model for multi-label learning is proposed to explicitly incorporates the pairwise correlation between any two class labels. They assume that multi-labeled text has a mixture of characteristic words appearing in single-labeled text that belong to each category of the multi-categories. By employing SVM and Bag-of-Word (BOW) representation, two types of probabilistic generative models for multi-labeled text called parametric mixture models (PMM1, PMM2) are presented in this work.

Yu *et al.* [Yu et al., 2005] introduce a multi-label informed latent semantic indexing (MLSI) algorithm which preserves the information of inputs and meanwhile captures the correlations between the multiple outputs. They assume a linear model between the input features and the output class labels, and a regression model is proposed to find the appropriate linear combination weights. The label correlation is explored by imposing a common prior for the combination weights on different classes. They use this method as a preprocessing step and achieve encouraging results on the multi-label text classification problems.

Griffiths *et al.* [Griffiths and Ghahramani, 2005] propose a Bayesian model to assign labels through underlying latent representations. They define priors over infinite combinatorial structures from nonparametric Bayesian statistics, and use it to develop methods for unsupervised learning in which each object is represented by a sparse subset of an unbounded number of features. And these features can be binary, take on multiple discrete values, or have continuous weights. In addition, they assume a probability distribution over equivalence classes of binary matrices with a finite number of rows and an unbounded number of columns, which is suitable for use as a prior in probabilistic models that represent objects using a potentially infinite array of features.

Despite above mentioned efforts in exploiting the co-occurrent information of labels, most of the research is limited to pairwise correlation between class labels. In order to improving the performance, there are several algorithms for multi-label learning that assume a particular structure among the class labels.

Cai *et al.* [Cai and Hofmann, 2004] present a hierarchical classification method that generalizes Support Vector Machine learning, which directly incorporates prior knowledge about class relationships. They assume a hierarchical structure among the class labels, and the introduced method is based on discriminant functions that are structured in a way that mirrors the class hierarchy. The method can work with arbitrary, not necessarily singly connected taxonomies and can deal with task-specific loss functions.

Rousu *et al.* [Rousu et al., 2004] also proposed a hierarchical structure among the class labels to explore multi-label learning problem. They present a variation of the maximum margin multi-label learning framework, which is suited to the hierarchical classification task and allows efficient implementation via gradient-based methods. And the classification hierarchy is represented as a Markov network equipped with an exponential family defined on the edges.

McCallum [McCallum, 1999] assumes that class labels can be divided into a small number of disjoint clusters. They describes a probabilistic generative model that can represents the correlations between class labels, in which the multiple classes that comprise a document are represented by a mixture model. Since the labeled training data indicates which classes were responsible for generating a document, and it does not indicate which class was responsible for generating each word. The authors use EM to fill in this missing value by learning both the distribution over mixture weights and the word distribution in each class's mixture component.

Kang *et al.* [Kang, Jin, and Sukthankar, 2006] proposed a correlated label propagation framework for multi-label learning, which explicitly exploits high-order correlation between labels. Different from previous approaches to multi-label learning that either treat class label independently or only take into count the pairwise correlations, the proposed algorithm exploits label correlations of any order. Most existing approaches only consider the propagation of a single class label between training examples and test examples. Motivated by this, the proposed framework takes into account the simultaneous propagation of multiple labels. They formulate the proposed framework as a linear programming problem with an exponential number of constraints, which cannot be practically solved using standard techniques. To solve this problem exactly and efficiently, an algorithm based on properties of submodular functions is introduced.

Different from this body of work motivated from Co-occurrent Label Context, we exploit in Chapter 3 a novel type of context named *label exclusive context* for multi-label learning, which describes the negative relationship among class labels.

3) Label Propagation

As a graph-based approach, Label Propagation belongs to the semi-supervised learning and has been proved to be an effective for both text categorization and image annotation [Kang, Jin, and Sukthankar, 2006; Cao, Luo, and Huang, 2008; Wang and Zhang, 2006; Zhou, Scholkopf, and Hofmann, 2005]. The central idea of label propagation is to first construct a graph in which each node represents a data point and each edge is assigned a weight (e.g. the similarity between data points), then propagate the class labels of labeled data to unlabeled data based on the constructed graph in order to make predictions. A number of machine learning algorithms have been exploited for label propagation, including the approaches based on Green functions [Zhou, Scholkopf, and Hofmann, 2005], harmonic functions [Zhu, Ghahramani, and Lafferty, 2003], Gaussian processes [Chu and Ghahramani, 2005]

Based on Green functions, Zhou *et al.* [Zhou, Scholkopf, and Hofmann, 2005] propose a general regularization framework on directed graphs for label propagation. Given a directed graph in which some of the nodes are labeled, the authors exploits the link structure of the graph to infer the labels of the remaining unlabeled nodes. For the regularization framework, the objective functions are defined over nodes of a directed graph that forces the classification function to change slowly on densely linked subgraphs.

Gaussian fields and harmonic functions-based methods are motivated by assuming that the label assignments should be smooth over the entire graph. The work in [Zhu, Ghahramani, and Lafferty, 2003] is one of most popular approaches in label propagation. Zhu *et al.*develop an label propagation approach to semi-supervised learning based on a Gaussian random field and harmonic functions, where the propagation is performed on a weighted graph representing labeled and unlabeled data. In their proposed scheme, firstly, labeled and unlabeled data are represented as vertices in a weighted graph with edge weights encoding the similarity between instances. Then the learning problem is formulated in terms of a Gaussian random field on this graph. This work concentrates on the use of only the mean of the field, which is characterized in terms of harmonic functions and spectral graph theory. The fully probabilistic framework is closely related to

Gaussian process classification

In [Chu and Ghahramani, 2005], the authors propose a probabilistic kernel approach to label propagation for preference learning over instances or labels. The formulation of learning label preference is based on Gaussian processes. A new likelihood function is proposed to capture the preference relations in the Bayesian framework. In addition, this approach remains linear with the size of the training instances, rather than growing quadratically, which provides a general Bayesian framework for model adaptation and probabilistic prediction.

Based on a linear neighborhood model, Wang *et al.* [Wang and Zhang, 2006] present a semi-supervised learning approach called Linear Neighborhood Propagation(LNP), which assumes that each data point can be linearly reconstructed from its neighborhood. The authors exploit the structure of the whole dataset through synthesizing the linear neighborhood around each data object. The LNP algorithm first approximates the whole graph by a series of overlapped linear neighborhood patches. The edge weights in each patch is solved by a standard quadratic programming procedure. Then all the edge weights will be aggregated together to form the weight matrix of the whole graph for annotating. Finally, LNP propagates the labels from the labeled points to the whole dataset using these linear neighborhoods with sufficient smoothness.

Label propagation is an important technique in machine learning. It has shown the promising performance in label learning problem. Despite the various motivations behind the above mentioned approaches utilizing label propagation, most of them are based on the prior assumption of consistency: nearby data points or data points on the same structure are likely to have the same class label. Based on the same assumption, we develop a multi-label propagation scheme for large-scale image annotation in Chapter 5, which based on efficient sparse graph construction.

2.2.2 Multi-Label Learning on Emotive Semantic Space

Unlike research in cognitive semantic space, most researchers in the field of emotive semantic image classification focus on effective and special visual feature extraction in order to explore the emotion contained in images. The popularly adopted methods for emotive semantic annotation include Support Vector Machines (SVMs), Neural Network, Random Forest, the C4.5 tree classifier and Naive Bayes classifier method. A summary of below introduced representative works in multi-label learning on emotive semantic space is provided in Table 2.1

Wang *et al.* [Wang, Yu, and Jiang, 2006] extract special integrated histogram features and utilize support vector regression for automatically emotional image annotation. The authors analyze the emotional space and propose a scheme to annotate the image emotion semantic automatically and realize emotional image retrieval. Based on psychological experiments measuring evoked feelings by art paintings, they first identify an orthogonal three-dimension emotional factor space of image through 12 pairs of emotional words. Then, the following three specific image features are designed for each emotional factor to predict it. They are luminance-warm-cool fuzzy histogram, saturation-warm-cool fuzzy histogram integrated with color contrast and luminance contrast integrated with edge sharpness. The values of emotional factors can be predicted from the image features automatically by using support vector machine of regression (SVR). Finally, an emotion-based image retrieval system is designed and implemented, in which the users can perform retrieval using semantic words.

In the work of [Yanulevskaya et al., 2008], an SVM framework for supervised learning of emotion categories is first adopted with extracting the holistic image features. Then, the authors develop an emotion categorization system trained by ground truth from psychology studies. The training data contains emotional valences scored by human subjects on the International Affective Picture System (IAPS). The extracted specific features (Gabor features, the Wiccest features, or both) for each ground truth image are used to train a classifier to distinguish between the various emotional valences. For test the performance, the emotion classification system is applied to a collection of masterpieces. Although the results are preliminary, they demonstrate the potential of machines to elicit realistic emotions as can be derived from visual scenes.

Machajdik *et al.* [Machajdik and Hanbury, 2010] investigate methods to extract and combine low-level features that represent the emotional content of an image from psychology and art theory views. Firstly, the authors exploit theoretical and empirical concepts from psychology and art theory to extract image features that are specific to the domain of artworks with emotional expression. Then, they concentrate on determining the affect of still images and studying the extracted specific feature for the task of affective image classification. For classification, they adopte the Naive Bayes classifier to annotate images with emotive concepts.

Besides focusing on extracting specific features for affective image annotation, many other works also employ the generic features to classify emotional images.

Hayashi *et al.* [Hayashi and Hagiwara, 1998] adopt the RGB color feature and classified the affective images through neural network. They developed a im-

Table 2.1: A list of the representative works in multi-label learning on emotive semantic space.

Work	**Extracted Features**	**Type of Feature**	**Algorithm**
[Hayashi and Hagiwara, 1998]	RGB color feature	generic feature	Neural Network
[Wu, Zhou, and Wang, 2005]	shape, color, and texture feature	generic feature	SVM
[Wang, Yu, and Jiang, 2006]	luminance-warm-cool fuzzy histogram, saturation-warm-cool fuzzy histogram integrated with color contrast and luminance contrast integrated with edge sharpness	specific feature	SVR
[Yanulevskaya et al., 2008]	Gabor features, Wiccest features	specific feature	SVM
[Machajdik and Hanbury, 2010]	combined specific low-level features learned from psychology and art theory	specific feature	Naive Bayes

age query system by impression words named the IQI system that automatically estimates impression words from various kinds of images. For the image characteristics, the system utilizes the RGB color projection distributions in the vertical and the horizontal axis, which correlate closely with the impression of an image and a spectrum of frequency domain which expresses density of the image. In addition, extracting these generic image feature is much simpler than other extraction methods of specific features. The authors prove that the proposed image characteristics are effective for the reduction of the units in the input layer and are robust for shift of the image.

In the work of [Wu, Zhou, and Wang, 2005], a method to classify and retrieve affective images is proposed, which utilizes SVM for affective image classification based on general features (shape, color, and texture features). Firstly, several adjective words are collected which are usually used by people when observing images, such as the word of lovely and beautiful etc. Then, users are invited to evaluate the images in the training set. In addition, affective space of image is constructed according to the thought of “dimensions”. Visual features of images are extracted and visual feature space of image is constructed. Finally, the mapping between affective space and visual feature space is calculated by SVMs.

2.2.3 Summary

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 above mentioned multi-label learning algorithms aim at annotating images with concepts coming from only one semantic view, e.g. cognitive or affective. They ignore the fact that a real world image dataset may consists of two or more semantic subspace. And the two or multiple spaces of image semantics are inter-related and can be used to reinforce each other for improving the annotation performance. Therefore, different from the above body of efforts on multi-label learning in unitary semantic space (either cognitive or emotive space), we propose and investigate the problem of multi-semantic image annotation in Chapter 4 to meet the requirement of real-world search conditions, which harmoniously embeds both cognitive semantic learning and affective semantic learning into one general framework for annotating the multi-semantic labels to images.

2.3 Semi-Supervised Learning in Large-Scale Dataset

For many applications like image annotation, especially in large-scale setting, annotating training data is often very time-consuming and tedious. This directly motivates a large number of recent endeavors for exploring semi-supervised learning for annotation when the dataset scales up to large-scale setting.

Recently many researchers focus on embedding the graph Laplacian regularizers into *transductive support vector machines* (TSVMs) on large-scale data. Collobert *et al.* [Collobert et al., 2006] apply *concave-convex procedure* (CCCP) to TSVMs when scaling up to large-scale setting, which shows that the improvements in SVM scalability can also be applied to TSVMs. In this work, CCCP is employed to iteratively optimizes non-convex cost functions that can be expressed as the sum of a convex function and a concave function. The optimization is carried out iteratively by solving a sequence of convex problems obtained by linearly approximating the concave function in the vicinity of the solution of the previous convex problem. The proposed large-scale transductive SVMs method is guaranteed to find a local minimum and has no difficult parameters to tune. This makes the proposed method an efficient approach for implementing Transductive SVM with an empirical scaling of $(L+U)^2$, which involves training a sequence of typically $1-10$ conventional convex SVM optimization problems (where L and U are the numbers of labeled and unlabeled examples.).

Sindhwani *et al.* [Sindhwani and Keerthi, 2006] provide an efficient and scalable implementation of TSVM for linear classification problems involving large and sparse datasets, which enhances the training speed of TSVM. In this work, the authors explore and present a family of semi-supervised linear support vector classifiers based on the finite Newton technique. These SVM algorithms are designed to handle partially-labeled sparse datasets with possibly very large number of examples and features. In this family of semi-supervised SVMs inspired from Deterministic Annealing (DA) techniques, the global minimizer is parametrically tracked. The proposed approach alleviates the problem of local minima in the

TSVM optimization procedure which results in better solutions on some problems. A computationally attractive training algorithm is also presented, which involves a sequence of alternating convex optimizations. Therefore, as proved by the experiments, these algorithms can be valuable and helpful in applied scenarios where sparse classification problems arise frequently, labeled data is scarce and plenty of unlabeled data is easily available.

Karlen *et al.* [Karlen et al., 2008] solve a large-scale transduction problem ($650,000$ samples), in which the regularizer of TSVM is trained by stochastic gradient descent. In this work, the authors introduce a large scale nonlinear method that elegantly combines the two main regularization principles for discriminative semi-supervised learning: transduction and manifold-based regularization. After They train the proposed system using stochastic gradient descent and choose linear or multi-layer architectures rather than kernel methods. This results in faster training and testing times than other TSVM algorithms. Since it also allows semi-supervised learning to be performed online, the proposed method can be scale to large-scale online computing.

To deal with much more larger dataset ($900,000$ labeled and unlabeled samples), Tsang *et al.* [Tsang and Kwok, 2006] adopt a sparsified manifold regularizer and formulates as a center-constrained minimum enclosing ball problem, which derives the sparse solutions with both low time and space complexities by utilizing *core vector machine* (CVM). In this work, the authors exploited two issues associated with the Laplacian SVM: 1) How to obtain a sparse solution for fast testing? 2) How to handle data sets with millions of unlabeled examples? For the first issue, a sparsified manifold regularizer based on the ϵ-insensitive loss is proposed. For the second one, manifold regularization is incorporated into the CVM. In the

Table 2.2: A list of the representative works of semi-supervised learning in large-scale dataset.

Work	Size of Dataset	Type of Annotation	Algorithm
[Collobert et al., 2006]	70,000	single-label learning	TSVMs
[Sindhwani and Keerthi, 2006]	80,000	single-label learning	TSVMs
[Tsang and Kwok, 2006]	900,000	single-label learning	CVM
[Karlen et al., 2008]	650,000	single-label learning	TSVMs
[Subramanya and Bilmes, 2009]	120 million	single-label learning	based on KLD

end, the above proposed solutions to the two issues make the resultant algorithm have low time and space complexities. In addition, a sparse solution can also be recovered by avoiding the underlying matrix inversion in the original LapSVM.

The work in [Subramanya and Bilmes, 2009] solves a problem on a 120 million node graph, in which the graph-regularized transductive learning formulation is based on minimizing a Kullback-Leibler divergence (KLD) based loss. At the beginning of the work, the authors provide theoretical analysis and give certain theoretical properties of the proposed graph-regularized transductive learning objective function. They prove that AM on the proposed KLD based objective converges to the true optimum, and provide a test for convergence. Then, in order to handling large-scale data, they propose a graph node ordering algorithm that is cache cognizant and leads to a linear speedup in parallel computations. This ensures that the algorithm can scale to large-scale datasets.

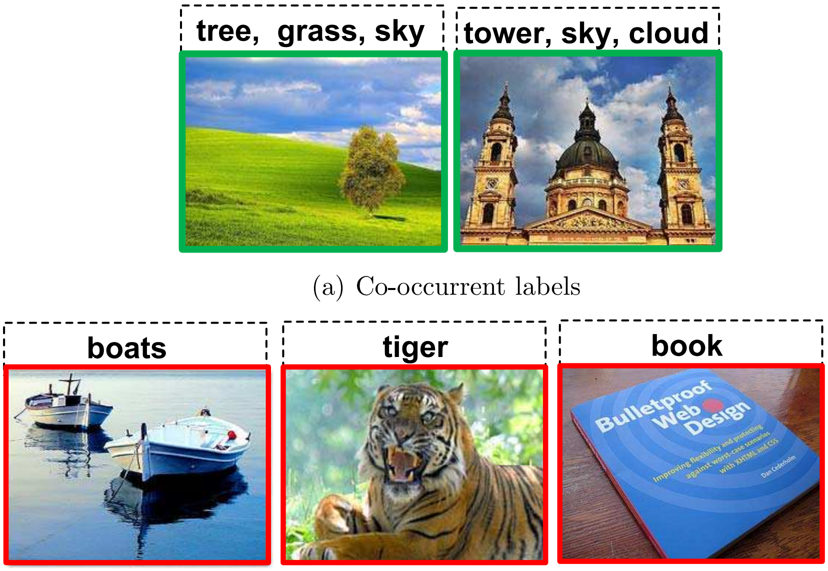
A summary of above mentioned representative works of semi-supervised learning in large-scale dataset is provided in Table 2.2. From Table 2.2, we observe that most of existing large-scale semi-supervised learning approaches focus on single-label annotation. However, a real world images usually has multiple semantics and requires multi-label annotation. The challenging situation for image annotation is that there is no sufficient labeled images whereas large numbers of unlabeled image data may could be far easier to crawled on the web, and annotating this large-scale unlabeled images with multiple labels is generally unaffordable for traditional large-scale semi-supervised learning methods. Different from above mentioned works on large-scale single-label learning, in this thesis, we are particularly interested in efficient multi-label learning for image annotation in large-scale setting, which will be introduced in Chapter 5.

Chapter 3

Multi-Label Learning with Label Exclusive Context

3.1 Introduction

Multi-label learning aims to solve the problem where each image sample can be assigned with multiple class labels simultaneously. As in a fixed data set, many concepts are semantically related, and hence the class labels may correlate to each other. As shown in Figure 3.1(a), the objects {"tree", "grass", "sky"} or {"tower", "sky", "cloud"} are frequently contained in the same image and thus form two groups of co-occurrent labels. As reviewed in Chapter 2, this kind of co-occurrent context is widely employed by many researchers for image annotation. However, the performance of multi-label learning could further be improved. Different from this body of work motivated from label co-occurrence, we exploit in this chapter a complementary type of context, the *label exclusive context*, that describes the negative relationship among class labels. For example, as shown in Figure 3.1(b),



(a) Co-occurrent labels

(b) Exclusive labels

Figure 3.1: Two types of label context in real-scene images. The label co-occurrent context as in (a) describes the *positive* correlation among labels. The label exclusive context as in (b) describes the *negative* correlation among labels. In this chapter, we will novelly incorporate the label exclusive context with linear representation for visual classification.

the objects {"boats", "tiger", "book"} seldom simultaneously appear in a real-scene image. We call such a kind of negatively correlated labels as exclusive labels. The label exclusive context has recently been explored in [Desai, Ramanan, and Fowlkes, 2009; Choi et al., 2010] for the object detection tasks. Here we are particularly interested in the effect of label exclusive context in the setup of multi-label image classification. Given a multi-label query image and several groups of exclusive labels learned from the training images, it is reasonable to expect that the labels in each group should be exclusively assigned to the predicted label vector. This motivates us to develop a visual classification framework with which label exclusive context may be naturally incorporated.

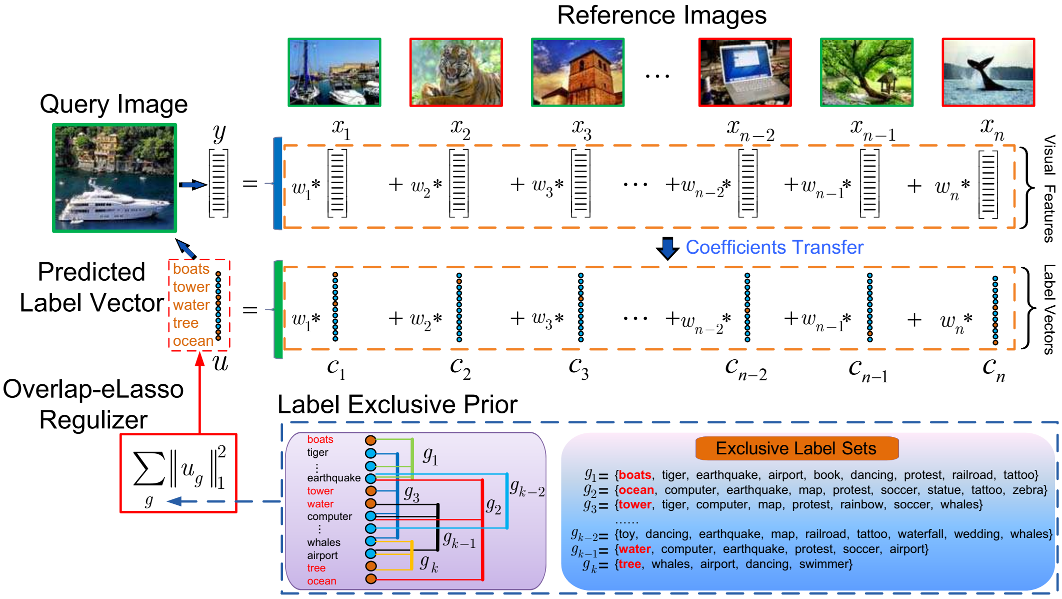

Reference Images
Query Image
x_1 x_2 x_3 … x_{n-2} x_{n-1} x_n
y
= w_1* $+ w_2*$ $+ w_3*$ $\cdots +w_{n-2}*$ $+w_{n-1}*$ $+ w_n*$
Visual Features
Coefficients Transfer
Predicted
Label Vector
boats
tower
water
tree
ocean
u
= w_1* $+ w_2*$ $+ w_3*$ $\cdots +w_{n-2}*$ $+w_{n-1}*$ $+ w_n*$
Label Vectors
c_1 c_2 c_3 c_{n-2} c_{n-1} c_n
Overlap-eLasso
Regulizer
$\sum_g \|u_g\|_1^2$
Label Exclusive Prior
boats
tiger
earthquake
tower
water
computer
whales
airport
tree
ocean
g_1 g_3 g_2 g_{k-2} g_{k-1} g_k
Exclusive Label Sets
g_1= {boats, tiger, earthquake, airport, book, dancing, protest, railroad, tattoo}
g_2= {ocean, computer, earthquake, map, protest, soccer, statue, tattoo, zebra}
g_3= {tower, tiger, computer, map, protest, rainbow, soccer, whales}
……
g_{k-2}= {toy, dancing, earthquake, map, railroad, tattoo, waterfall, wedding, whales}
g_{k-1}= {water, computer, earthquake, protest, soccer, airport}
g_k= {tree, whales, airport, dancing, swimmer}

Figure 3.2: Flowchart of linear representation with exclusive label context.

3.1.1 Scheme Overview

The major contribution of this work is a label exclusive context regularized linear representation and classification method. It is notoriously hard to impose repulsive forces between labels. Desai *et al.* [Desai, Ramanan, and Fowlkes, 2009] utilize a greedy algorithm to learn and impose repulsive forces for non-maxima suppression for the object detection task. In this chapter,, the problem of image classification is formulated as an exclusive Lasso [Zhou, Jin, and Hoi, 2010] model which is a recent advance in sparse learning. Figure 3.2 depicts the working mechanism of our method. For a given query image feature y, we seek a linear representation coefficient vector w that best reconstructs y from reference image features X. The predicted label vector u of the query image is the linear combination of the reference image label vectors (zero-one vector indicating multi-label) $C = [c_1, ..., c_n]$ using the same coefficient vector w (to be estimated), i.e., $u = Cw$. Given a set of exclusive label groups $\mathcal{G}$, we expect that at most one label inside each exclusive label set $g \in \mathcal{G}$ will be non-zero in the predicted label vector u. The problem can be cast as an exclusive Lasso with group overlaps and affine transformation. To optimize such an variant of eLasso, we develop a Nesterov-type smoothing approximation [Nesterov, 2005] method to convert the non-smooth problem to a smooth problem and then solve it using the Accelerated Proximal Gradient method [Tseng, 2008]. Moreover, in our application, the exclusive label groups are automatically learned using the dense subgraph searching method [Liu and Yan, 2010]. Empirical studies on the challenging visual classification tasks validate the effectiveness of our label exclusive linear representation and classification method. Flowchart of linear representation with exclusive label context is shown in Figure 3.2. In this system, we have a dictionary of reference images $X = [x_1, ..., x_n]$ with labels $C = [c_1, ..., c_n]$, and a collection of predefined or learned exclusive label sets $\mathcal{G}$. Given a query image y, our method tends to exclusively select labels inside each label set $g \in \mathcal{G}$ to appear in the predicted label vector $u = Cw$ where w (to be learned) best reconstructs the query image, i.e., $y \approx Xw$. This model can be cast as an exclusive Lasso problem with group overlaps and affine transformation. For better viewing, please see original color pdf file.

3.1.2 Related Work

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

3.1.2.1 Sparse Linear Representation for Classification

Linear representation with sparse inducing regularizer has enjoyed considerable popularity in recent multi-class visual recognition applications [Wright et al., 2009; Yan and Wang, 2009; Yuan and Yan, 2010]. Given a query image feature and a dictionary of reference features, the objective of sparse linear representation is to select a small set of reference images to reconstruct the query image. Such a sparse representation scheme is typically free of model training and robust to sparse noise. In this work, we show that the label exclusive context can be elegantly integrated into linear representation to boost classification performance.

3.1.2.2 Group Sparse Inducing Regularization

Learning models regularized by group sparse inducing penalties have been widely studied in both machine learning [Yuan and Lin, 2006; Zhao, Rocha, and Yu, 2009] and signal processing fields [Kowalski, 2009; Fornasier and Rauhut, 2008]. Let $w \in \mathbb{R}^n$ be the n parameters to be regularized. Denote $\mathcal{I} = \{1, ..., n\}$ the variable index and $\mathcal{G} = \{g_i \subseteq \mathcal{I}\}_{i=1}^{l}$ a set of variable index groups. The group formation varies according to the given grouping or hierarchical structure. Denote $\|w_{\mathcal{G}}\|_{p,q} := \sum_{g \in \mathcal{G}} \|w_g\|_p^q$ the $\ell_{p,q}$-norm defined over groups $\mathcal{G}$, where $\|w_g\|_p^q := \left(\sum_{j \in g} |w_j|^p\right)^{q/p}$. The $\ell_{2,1}$-norm regularizer is used in group Lasso [Yuan and Lin, 2006] which encourages the sparsity on group level. Jacob *et al.* [Jacob, Obozinski, and Vert, 2009] 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 [Liu, Palatucci, and Zhang, 2009; Zhang, 2006].

3.1.2.3 Exclusive Lasso

When $p = 1$, $q = 2$, the $\ell_{1,2}$-norm has recently been studied in the exclusive Lasso (eLasso) regression [Zhou, Jin, and Hoi, 2010] for the multi-task learning. Given a set of observed data $\mathcal{D} = \{X, y\}$ in which $X \in \mathbb{R}^{m \times n}$ is the design matrix of predictors, and $y \in \mathbb{R}^m$ is a response vector. The eLasso is defined (in our notations) as solving the following $\ell_{1,2}$-regularized least squares problem

$$\min_w \frac{1}{2}\|y - Xw\|_2^2 + \frac{\lambda}{2}\sum_{g \in \mathcal{G}} \|w_g\|_1^2, \tag{3.1}$$

where λ is a user-specified term trade-off parameter. Unlike the group Lasso[Yuan and Lin, 2006] regularizer that assumes covariant variables in groups, the eLasso regularizer models the scenario where variables in the same group are exclusively selected in the output. It is assumed in [Zhou, Jin, and Hoi, 2010] that the groups in $\mathcal{G}$ are *disjoint*. In our work, motivated by the practice of multi-label visual classification, we will investigate the optimization of an important variant of eLasso with group overlap and affine transformation of parameter vector.

The remainder of this chapter is organized as follows: We present the label exclusive linear representation and classification framework in Section 3.2. The optimization procedure is described in Section 3.3. Section 3.4 states a kernel-view extension of our method in the setting where features are given in form of kernel matrices. The experimental results on several benchmark visual classification tasks are given in Section 3.5. We conclude this work in Section 3.6.

3.2 Label Exclusive Linear Representation and Classification

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

3.2.1 Label Exclusive Linear Representation

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

$$\min_w \frac{1}{2}\|y - Xw\|_2^2 + \frac{\lambda}{2} \sum_{g \in \mathcal{G}} \|C_g w\|_1^2, \tag{3.2}$$

where w is the linear reconstruction coefficient vector, $\mathcal{G}$ is a group of label subsets, each of which contains several exclusive classes (assumed to be known here, and we will address soon in Section 3.2.2 how to automatically learn $\mathcal{G}$ from reference set), and C_g is the rows of C indexed in g. The first term measures the linear reconstruction error of feature y by Xw, while the second term utilizes the $\ell_{1,2}$-norm to encourage the label exclusion behavior in the predicted label confidence vector Cw. Since both terms are convex, the objective in (3.2) is convex. Apparently, LELR model (3.2) is a variant of the standard eLasso problem (3.1), with the following notable differences:

- The groups in $\mathcal{G}$ may be overlapping to each other (see Section 3.2.2).
- The groups are defined over the affine transformed output Cw, rather than on the original parameter vector w.

We will design later on in Section 3.3 an efficient first-order method to optimize the objective in (3.2). Given the optimal reconstruction coefficient $\hat{w}$, the optimal $\hat{u} = C\hat{w}$ can be regarded as a label confidence vector of the query image. Such a vector can be used for performance evaluation by calculating metrics such as the average precision (AP).

3.2.2 Learn the Exclusive Label Sets

So far, we assume that the set $\mathcal{G}$ of exclusive label groups used in problem (3.2) is known as a prior. Actually, it can be automatically learned from the training data. Here we use the graph shift method [Liu and Yan, 2010] to learn a few groups of exclusive labels as dense subgraphs on a weighted graph $G = \langle V, E \rangle$ defined as follows: the node set $V := \{1, 2, ..., p\}$ contains all the class labels, and the edge set $E \subseteq V \times V$ describes the pairwise exclusiveness between nodes. The weight matrix W associated with G is given by $W_{ij} = 1$ if label i and label j do not simultaneously appear in any training image, and $W_{ij} = 0$ otherwise. The dense subgraphs of G are then determined by the graph-shift method [Liu and Yan, 2010]. The nodes in each dense subgraph naturally form, with high confidence, an exclusive label subset. Note that the exclusive groups learned in this way are typically overlapping to each other. Taking NUS-WIDE-LITE dataset [Chua et al., 2009] as an example, it can be seen in the right part of Figure 3.2 that the labels “tiger”, “airport”, “map”, “whales”, etc., all belong to more than one groups.

3.3 Optimization

In this section, we investigate the optimization problem associated with the LELR model (3.2). Since LELR is a variant of eLasso, one may wish to utilize the existing eLasso solvers for optimization. However, it comes to our notice that the eLasso solvers in literature either suffer from slow convergence rate (e.g., subgradient methods in [Zhou, Jin, and Hoi, 2010]) or are particularly designed for standard eLasso (3.1) with disjoined groups (e.g., proximal gradient method in [Kowalski and Torreesani, 2009]), and thus are not directly applicable to LELR. This motivates us to seek for more suitable tools to optimizie the objective in (3.2). One natural thought is to approximate the non-smooth objective in (3.2) by a smooth function and then solve the latter by utilizing the off-the-shelf smooth optimization algorithms. Next, we derive a Nesterov’s smoothing optimization method to achieve this task.

3.3.1 Smoothing Approximation

Let us re-express LELR (3.2) as follows

$$\min_{w} \left\{ F(w) := f(w) + \lambda h(w) \right\}, \tag{3.3}$$

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

$$\|C_g w\|_1 = \max_{\|u_g\|_\infty \le 1} \langle C_g w, u_g \rangle. \tag{3.4}$$

By utilizing the Nesterov's smoothing approximation method [Nesterov, 2005], the $\|C_g w\|_1$ in (3.4) can be approximated by the following smooth function

$$q_{g,\mu}(w) := \max_{\|u_g\|_\infty \le 1} \langle C_g w, u_g \rangle - \frac{\mu}{2}\|u_g\|_2^2, \tag{3.5}$$

where μ is a parameter to control the approximation accuracy. For a fixed w, denote $u_g(w) \in \mathbb{R}^{|g|}$ the unique minimizer of (3.5). It is standard that

$$u_g(w) = \min\left\{1, \max\left\{-1, \frac{C_g w}{\mu}\right\}\right\}. \tag{3.6}$$

Based on these preliminaries, we now propose to solve the following smooth optimization problem as an approximation to the non-smooth problem (3.3):

$$\min_w \left\{F_\mu(w) := f(w) + \lambda h_\mu(w)\right\}, \tag{3.7}$$

where h_μ is given by

$$h_\mu(w) := \frac{1}{2}\sum_{g\in\mathcal{G}} q_{g,\mu}^2(w). \tag{3.8}$$

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

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

$$h_\mu(w) \le h(w) \le h_\mu(w) + (m\|C\|_1 R|\mathcal{G}|)\mu. \tag{3.9}$$

Proposition 2 shows that for $\mu > 0$, the function h_μ can be seen as a uniform smooth approximation of function h.

Proof of Proposition 2

Proof. Since $0 \in \{u_g : \|u_g\|_\infty \leq 1\}$, by (3.5) we get that

$$0 \leq q_{g,\mu}(w) \leq \max_{\|u_g\|_\infty \leq 1} \langle C_g w, u_g \rangle = \|C_g w\|_1. \tag{3.10}$$

Therefore it holds that

$$h_\mu(w) = \frac{1}{2} \sum_{g \in \mathcal{G}} q_{g,\mu}^2(w) \leq \frac{1}{2} \sum_{g \in \mathcal{G}} \|C_g w\|_1^2 = h(w). \tag{3.11}$$

It is trivial to check that $\|u_g\|_\infty \leq 1$ implies $\|u_g\|_2^2 \leq |g| \leq m$. Therefore,

$$q_{g,\mu}(w) \geq \max_{\|u_g\|_\infty \leq 1} \langle C_g w, u_g \rangle - \frac{\mu m}{2} = \|C_g w\|_1 - \frac{\mu m}{2}. \tag{3.12}$$

Combining (4.9) and (4.10) we get

$$|q_{g,\mu}(w) - \|C_g w\|_1| \leq \frac{\mu m}{2}. \tag{3.13}$$

Thus

$$\begin{aligned} & |q_{g,\mu}^2(w) - \|C_g w\|_1^2| \\ = \;& |q_{g,\mu}(w) - \|C_g w\|_1| \cdot |q_{g,\mu}(w) + \|C_g w\|_1| \\ \leq \;& \frac{\mu m}{2} 2\|C_g w\|_1 \leq \mu m \|C\|_1 R, \end{aligned} \tag{3.14}$$

which implies that

$$q_{g,\mu}^2(w) \geq \|C_g w\|_1^2 - \mu m \|C\|_1 R. \tag{3.15}$$

By summarizing both sides of the above inequality over $g \in \mathcal{G}$, we immediately get

$$h_\mu(w) \geq h(w) - \mu m \|C\|_1 R |\mathcal{G}|. \tag{3.16}$$

Combining (4.9) and (3.16) leads to (3.9). □

Motivated from [Nesterov, 2005, Theorem 1], we derive the following result stating that h_μ is differentiable with Lipschitz continuous gradient:

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

$$\nabla h_\mu(w) = \sum_{g \in \mathcal{G}} q_{g,\mu}(w)(C_g^T u_g(w)) \tag{3.17}$$

is Lipchitz continuous with the constant

$$L_\mu = \left(m + \frac{\|C\|_1 R}{\mu} \right) \|C\|_2^2 |\mathcal{G}|. \tag{3.18}$$

Proof of Theorem 1

Proof. From the standard results (see, e.g. [Nesterov, 2005, Theorem 1]) we have that $q_{g,\mu}(w)$ is well defined and continuously differentiable, and its gradient $\nabla q_{g,\mu}(w) = C_g^T u_g(w)$ is Lipschitz continuous with constant

$$L_{g,\mu} = \frac{\|C_g\|_2^2}{\mu} \leq \frac{\|C\|_2^2}{\mu}. \tag{3.19}$$

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

$$\nabla h_\mu(w) = \sum_{g \in \mathcal{G}} q_{g,\mu}(w) \nabla q_{g,\mu}(w). \tag{3.20}$$

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

$$\begin{aligned}
& \|q_{g,\mu}(w_1)\nabla q_{g,\mu}(w_1) - q_{g,\mu}(w_2)\nabla q_{g,\mu}(w_2)\|_2 \\
= \; & \|q_{g,\mu}(w_1)\nabla q_{g,\mu}(w_1) - q_{g,\mu}(w_1)\nabla q_{g,\mu}(w_2) \\
& + q_{g,\mu}(w_1)\nabla q_{g,\mu}(w_2) - q_{g,\mu}(w_2)\nabla q_{g,\mu}(w_2)\|_2 \\
\leq \; & |q_{g,\mu}(w_1)| \cdot \|\nabla q_{g,\mu}(w_1) - \nabla q_{g,\mu}(w_2)\|_2 \\
& + \|\nabla q_{g,\mu}(w_2)\|_2 \cdot |q_{g,\mu}(w_1) - q_{g,\mu}(w_2)| \\
\leq \; & \left(\frac{\|C\|_2^2 \|C\|_1 R}{\mu} + \|C\|_2^2 m \right) \|w_1 - w_2)\|_2,
\end{aligned} \tag{3.21}$$

where the last inequality follows the basic facts: (i) (4.14), (ii) $q_{g,\mu}(w_1) \leq \|C_g w_1\|_1 \leq \|C\|_1 R$, (iii) $\|\nabla q_{g,\mu}(w_2)\|_2 = \|C_g^T u_g(w_2)\|_2 \leq \|C\|_2\sqrt{m}$, and (iv) $|q_{g,\mu}(w_1) - q_{g,\mu}(w_2)| \leq \|C\|_2\sqrt{m}\|w_1 - w_2\|_2$ (due to the boundness of $\nabla q_{g,\mu}$ in (iii)). By combining (3.20) and (4.15) we establish the validity of (4.13). □

3.3.2 Smooth Minimization via APG

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

$$\nabla F_\mu(w) = X^T(Xw - y) + \lambda \nabla h_\mu(w), \tag{3.22}$$

which is Lipschitz continuous with constant

$$\tilde{L}_\mu = \|X^T X\|_2 + \lambda L_\mu. \tag{3.23}$$

Therefore, we employ the Accelerated Proximal Gradient method [Tseng, 2008] to optimize the smoothed LELR problem (4.16). 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 iteration counter. 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 $1/\tilde{L}_\mu$ which is dominated by the factor μ when it is small. To further accelerate Algorithm 1 for extremely small μ, one may apply the continuation technique as suggested in [Becker, Bobin, and Candes, 2011].

```
Inputs : $X \in \mathbb{R}^{m\times n}$, $y \in \mathbb{R}^m$ , $C$, $\mathcal{G}$, $\lambda$, $\mu$.
Output: $w \in \mathbb{R}^n$
Initialization: Calculate $\tilde{L}_\mu$ by (3.23). Initialize $w_0, v_0$ and let
$\alpha_0 \leftarrow 1$, $t \leftarrow 0$.
repeat
    $u_t = (1-\alpha_t)w_t + \alpha_t v_t$,
    Calculate $\nabla h_\mu(u_t)$ according to (3.17),
    $v_{t+1} = v_t - \frac{1}{\alpha_t \tilde{L}_\mu}\left(X^T(Xu_t - y) + \lambda\nabla h_\mu(u_t)\right)$,
    $w_{t+1} = (1-\alpha_t)w_t + \alpha_t v_{t+1}$,
    $\alpha_{t+1} = \frac{2}{t+1}$, $t \leftarrow t+1$.
until Converges ;
```

Algorithm 1: Smooth minimization for LELR

3.4 A Kernel-view Extension

So far, the smooth minimization Algorithm 1 only applies to LELR (3.2) with raw image features (y, X). However, in the practice of visual classification, the descriptors are often encoded as similarities or kernel matrix, without the raw features available. For the purpose of utilizing feature kernels for LELR, we present in this subsection an extension of LELR to Reproducing Kernel Hilbert Space (RKHS). The intuition of such a kernel trick is to use a non-linear function ϕ to map the reference and query samples from the original space to a higher dimensional RKHS in which we have $\phi(x_i)^T\phi(x_j) = k(x_i, x_j)$ for certain kernel function $k(\cdot,\cdot)$. In this new space, we can write the problem (3.2) as:

$$\min_w \frac{1}{2}\|\phi(y) - \phi(X)w\|_2^2 + \frac{\lambda}{2}\sum_{g\in\mathcal{G}} \|C_g w\|_1^2, \tag{3.24}$$

where $\phi(X) = [\phi(x_1), ..., \phi(x_n)]$. Note that the calculation in APG iteration of Algorithm 1 is characterized by inner product of features, and thus can be straightforwardly extended to solve problem (3.24). Let $K = \phi(X)^T\phi(X)$ be the reference feature kernel matrix, and $z = \phi(X)^T\phi(y)$ be the query kernel vector. The kernel-view of Algorithm 1 for LELR is given in Algorithm 2.

```
Inputs : $K \in \mathbb{R}^{n \times n}$, $z \in \mathbb{R}^n$ , $C$, $\mathcal{G}$, $\lambda$, $\mu$.
Output: $w \in \mathbb{R}^n$
Initialization: Calculate $\tilde{L}_\mu$ by (3.23). Initialize $w_1, v_1$ and let
$\alpha_0 \leftarrow 1$, $t \leftarrow 0$.
repeat
    $u_t = (1 - \alpha_t) w_t + \alpha_t v_t$,
    Calculate $\nabla h_\mu(u_t))$ according to (3.17),
    $v_{t+1} = v_t - \frac{1}{\alpha_t \tilde{L}_\mu} (K u_t - z + \lambda \nabla h_\mu(u_t))$,
    $w_{t+1} = (1 - \alpha_t) w_t + \alpha_t v_{t+1}$,
    $\alpha_{t+1} = \frac{2}{t+1}$, $t := t + 1$.
until Converges ;
```

Algorithm 2: Smooth minimization for LELR in kernel-view

3.5 Experiments

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

3.5.1 Datasets and Features

The PASCAL VOC 2007&2010 are two challenging databases from the PASCAL Visual Object Classes Challenge (VOC) [Everingham et al., 2010]. A total of 20 object classes are collected from four main categories, i.e. *Person*, *Animal*, *Vehicle* and *Indoor*. VOC 2007 and VOC 2010 datasets contain 9,963 and 21,738 images respectively. Both datasets are split into 50% for training/validation and 50% for testing. The distributions of images and objects by class are approximately equal across the training/validation and test sets. We utilize the training set as reference image set. We extract several low-level features including SIFT and its variants [Sande, Gevers, and Snoek, 2010], LBP and HOG by dense sampling strategy in three scales. Each image is represented by Bag-of-Word model with spatial pyramid matching [Lazebnik, Schmid, and Ponce, 2006]. These features are first transformed to kernel space using χ^2 distance and further combined with a detection kernel as in [Chen et al.,].

The NUS-WIDE-LITE [Chua et al., 2009] is a lite version of NUS-WIDE database which contains 269,648 images and the associated 5,018 tags. This lite data set consists of 55,615 images randomly selected from the NUS-WIDE data set. For each image, an 81-D label vector is maintained to indicate its relationship to 81 distinct concepts (tightly related to tags yet relatively high-level). For evaluation, we construct a reference image set of size 27,807 whilst the rest are used for testing. We extract multiple types of global visual features which include 225-D block-wise color moments, 128-D wavelet texture and 75-D edge direction histogram. These features are transformed to kernel space using χ^2 distance and linearly combined into a mean feature kernel.

3.5.2 Evaluation Criteria

Following [Zhou et al., 2010], the criteria to evaluate the performance include *Average Precision* (AP) for each label (or concept) and *Mean Average Precision* (MAP) over all labels. The former is a well-known gauge widely used in the field of image retrieval, whilst the latter is developed to handle the multi-class and multi-label problems. All experiments are conducted on a common PC equipped with 2 Intel quad-core 3.0 GHz CPU and 32GB physical memory.

3.5.3 Results on PASCAL VOC 2007&2010

On VOC 2007, a total number of 11 exclusive label groups are learned, and each group contains 6 labels in average. We compare LELR with two state-of-the-art methods: Locality-constrained Linear Coding (LLC) [Wang et al., 2010] and Super Vector Coding (SVC) [Zhou et al., 2010], and two reported top ranked solutions [Everingham et al., a]: the INRIA_Flat and INRIA_Genetic. Moreover, we are interested in the performance comparison between label exclusive context and label co-occurrence context in linear representation and classification. To do this, we simply replacing the eLasso-type regularizer $\sum_{g\in\mathcal{G}}\|C_g w\|_1^2$ in LELR with a graph Laplacian regularizer that enforces label co-occurrence

$$\min_w \frac{1}{2}\|y - Xw\|^2 + \frac{\lambda}{2}w^T C^T LCw, \tag{3.25}$$

where $L = D - W$, W is a label co-occurrence matrix with the entry W_{ij} counting the number of times an object with label i appears in a training image with an object with label j, and D is a diagonal matrix with $D_{ii} = \sum_j W_{ij}$. We call such a model as label co-occurrence linear representation (LCLR). The objective in (3.25) is quadratic and thus can be optimized with closed form solution.

Table 3.1 lists the quantitative results. As can be seen that our LELR solution outperforms the competing methods in MAP and APs on 18 out of the 20 object classes. On comparison between LELR and LCLR, since both utilize the same features, the improvement of the former over the latter is supposed to stem from the fact that label exclusive context is more helpful than label con-occurrence context in linear representation and classification. The per query time of LELR is ~ 0.13 second.

On VOC 2010, a total number of 11 exclusive label groups are learned on the average of 8 labels per group. The comparing results on VOC 2010 are listed in Table 3.2. In this table, we compare our approach with the Winner'10 system from NUS-PSL team [Everingham et al., b]: the rank-one algorithm NUSPSL_KERNELREGFUSING and the rank-two algorithms NUSPSL_MFDETSVM.

Table 3.1: The APs and MAPs of different image classification algorithms on the PASCAL VOC 2007 dataset. The **INRIA_F** and **INRIA_G** stand for INRIA_Flat and INRIA_Genetic, respectively.

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

We also fuse the results of LELR and a standard SVM classifier trained on the same kernel to further improve the final performance as used in [Chen et al.,]. As can be seen from Table 3.2, LELR outperforms NUSPSL_MFDETSVM in MAP and APs on 18 out of 20 classes, and LELR+SVM outperforms NUSPSL_KERNELREGFUSING in MAP and APs on 14 out of 20 classes. Here we do not report the results by LCLR since it is inferior to the state-of-the-art and also for ease of presentation of the table. The per query time of LELR is ~ 0.2 second.

3.5.4 Results on NUS-WIDE-LITE

On NUS-WIDE-LITE dataset, a total number of 47 exclusive label groups are learned with averagely 9 labels per group (see the right part of Figure 3.2 for some

Table 3.2: Performance comparison of different image classification algorithms on the PASCAL VOC 2010 dataset.

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

exemplar groups). We compare LELR with the following five algorithms: KNN, SVM, LCLR, Entropic Graph Semi-Supervised Classification (EGSSC) [Subramanya and Bilmes, 2009] and Large-scale Multi-label Propagation (LSMP) [Chen et al., 2010]. The last two are semi-supervised methods which make use of the feature information of test samples. LSMP is our proposed multi-label learning approach in large-scale dataset in Chapter 5, which is the state-of-the-art algorithm in large-scale multi-label image annotation. All the algorithms utilize the same features as described in Section 3.5.1.

The MAP results obtained under varying reference set sizes (in percentages of the training set) are shown in Figure 3.3. Figure 3.4 illustrates the detailed APs for each of the 81 concepts, with the whole training set as reference set. Our observations from Figure 3.3 and Figure 3.4 are: (i) under different reference set

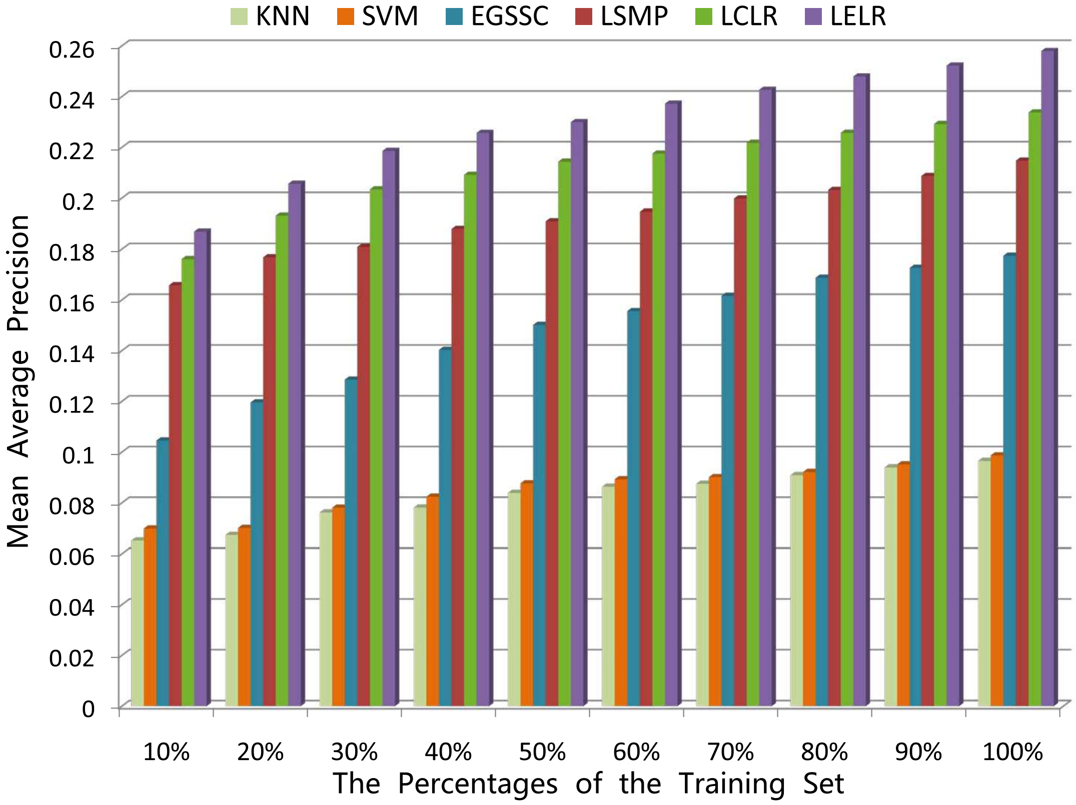


Figure 3.3: The MAP results of our LELR algorithm and the four baselines with varying reference image set sizes (in percentage) on NUS-WIDE-Lite dataset.

sizes, LELR consistently outperforms all the baseline algorithms in MAP; and (ii) in Figure 3.4, LELR and LCLR significantly outperform the other comparing algorithms on some rare concepts (e.g., map, horses, swimmers, waterfall, etc.). This is because LELR and LCLR are a linear representation model which is free of explicit model training and thus is relatively insensitive to the imbalance issue. The LELR per query processing time is ~ 0.75 second.

3.6 Conclusion

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

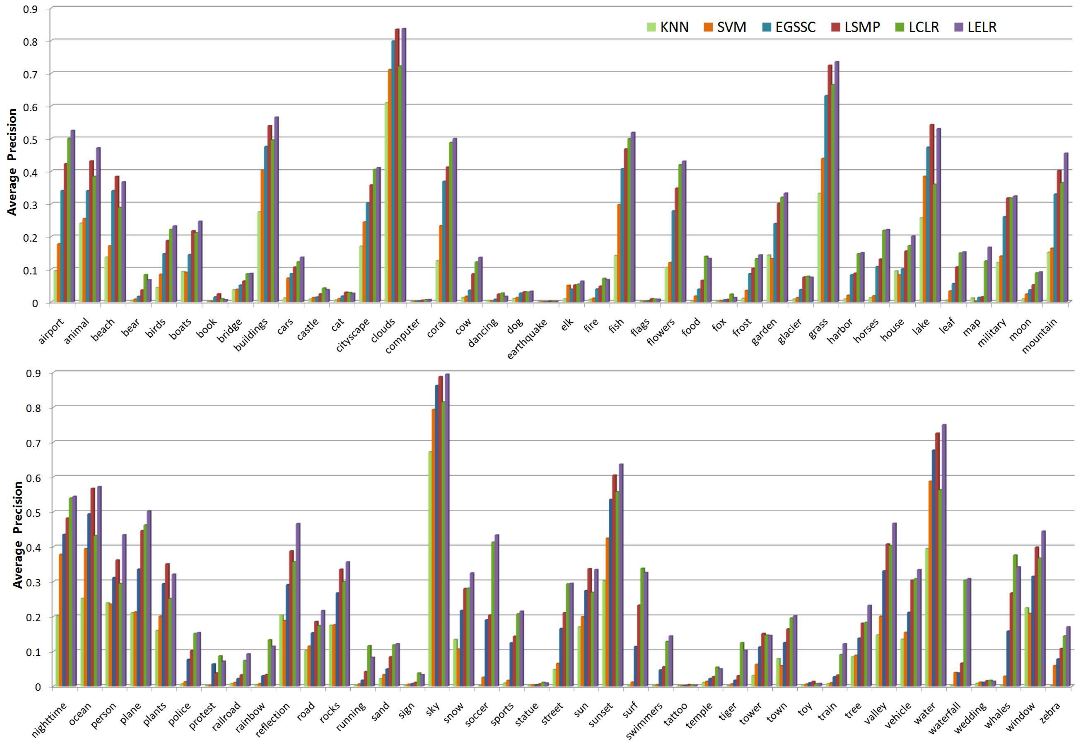


Figure 3.4: The comparison of APs for the 81 concepts using five methods with the whole training set as reference set on NUS-WIDE-LITE.

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

Chapter 4

Multi-Label Learning on Multi-Semantic Space

4.1 Introduction

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 meanings 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 the images, such as Amusement, Awe, Contentment, Excitement, Anger, Disgust, Fear and Sad [Machajdik and Hanbury, 2010; Mikels et al., 2005]. For popular cognitive or affective queries, the returned images can fill many result pages in popular search engines, but many will not satisfy the deeper requirement of complex and multi-semantic retrieval. For

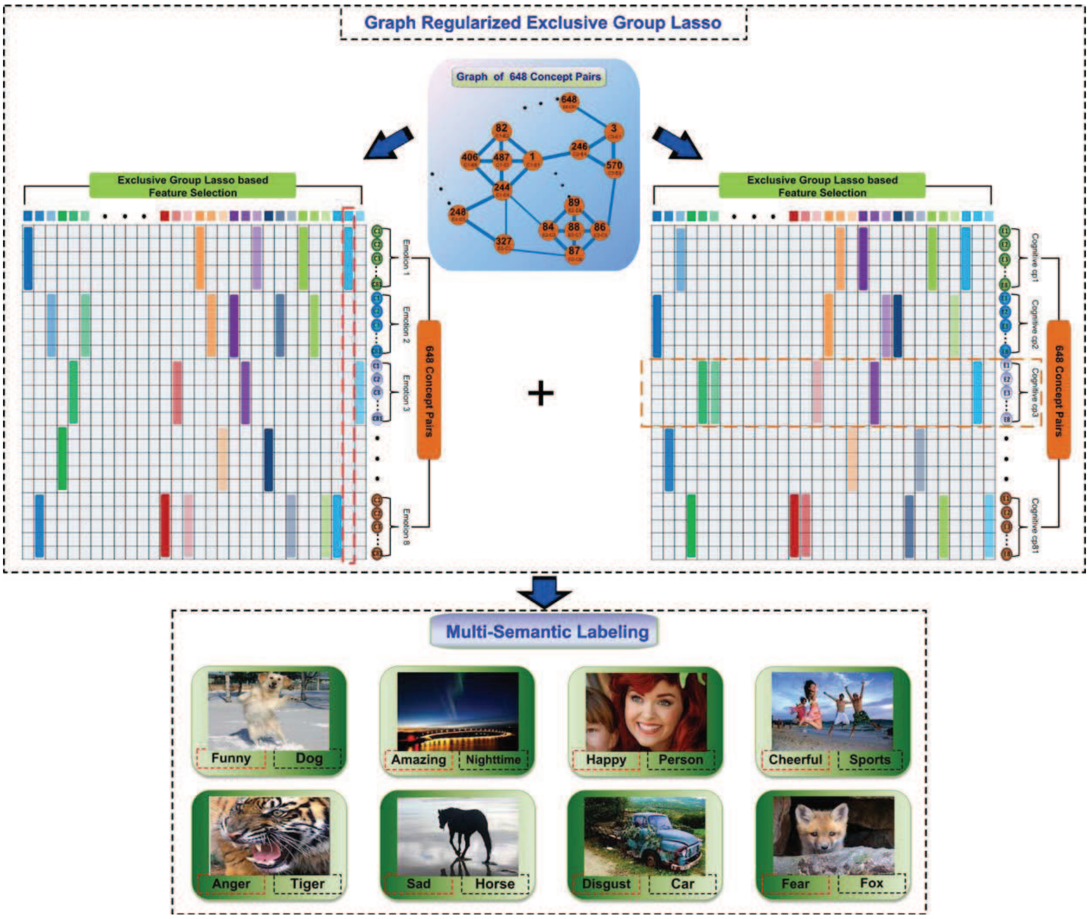


Figure 4.1: System overview of our proposed Multi-Task Learning scheme for Image Annotation with Multi-Semantic Labeling (IA-MSL).

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 precisely 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 and efficient comprehensive semantic image annotation technique.

Learning to annotate the comprehensive semantics to images in multi-semantic spaces is a challenging problem in the real world applications. In this chapter, 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 spaces. The key challenge with IA-MSL is the large number of classes involved in training due to the combination of multiple semantic spaces. Thus, some classes may suffer from the problem of insufficient training samples. One naive solution to avoid this issue is to train the classifiers within each semantic space and then combine the outputs from these semantic spaces for the ultimate combinational semantic prediction in an *enrichment* manner, which imposes the conditional independency assumption. More formally, 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

$$y = y^1 \otimes y^2,$$

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, given an image observation x, we made the semantic space independent assumption, i.e., $P(l^1, l^2 \mid x) = P(l^1 \mid x)P(l^2 \mid x)$, $\forall l^1 \in \mathcal{L}^1, l^2 \in \mathcal{L}^2$. Although simple for implementation, the apparent limitation of enrichment scheme is that it ignores the correlations among the semantic spaces. To deal with such an issue and harness the correlations across the semantic spaces, 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 $\{(l_1, l_2) \mid l_1 \in \mathcal{L}_1, l_2 \in \mathcal{L}_2\}$. 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 space, 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, which performs on the unified generic features, greatly reducing the cost of extracting different types of feature for different semantic spaces. Taking the NUS-WIDE-Emotive dataset as an example, in both emotive space $\mathcal{L}_1$ with 8 concepts and cognitive space $\mathcal{L}_2$ 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. System overview of our proposed Multi-Task Learning scheme for Image Annotation with Multi-Semantic Labeling (IA-MSL) is shown in Figure 4.1. In this figure, the training data are simultaneously labeled in both cognitive (81 concepts) and emotive semantic (8 concepts) spaces. The system is trained with a multi-task linear regression model regularized by a graph term, and an exclusive group lasso term. The graph term (middle of the top part) encourages correlation of the $648 = 8 \times 81$ emotive-cognitive concept pairs among tasks. While the exclusive group lasso term encourages sparse feature sharing across different cognitive concepts under the same emotive category (left of the top part) and different emotive concepts under the same cognitive category (right of the top part), which also captures the negative correlations among different emotive/cognitive categories.

4.1.1 Major Contributions

The major contributions of this chapter are three-fold:

- We propose a novel framework for Image Annotation with Multi-Semantic Labeling (IA-MSL), which exploits the high-level semantic of images from two or more semi-orthogonal label spaces;
- 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.

4.1.2 Related Work

4.1.2.1 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 [Caruana, 1997]. In general, MTL can be addressed through a regularization framework [Evgeniou and Pontil, 2004]. For example, the joint sparsity regularization favors to learn a common subset of features for all tasks [Argyriou, Evgeniou, and Pontil, 2008; Obozinski, Taskar, and Jordan, 2009], while the exclusive sparsity regularization is used for exclusive feature selection across tasks in [Zhou, Jin, and Hoi, 2010]. Our method follows the regularized MTL framework. In contrast to the existing regularization that only considers the model parameters dependent, our proposed regularization is characterized by data as well as model parameters, and thus is much more informative.

4.1.2.2 Group Sparse Inducing Regularization

In this section, we briefly recall some related work and the same representation of Group Sparse Inducing Regularization which is detailed in Section 3.1.2.2. 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 [Yuan and Lin, 2006] which encourages the sparsity on group level. Jacob *et al.* [Jacob, Obozinski, and Vert, 2009] 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 [Liu, Palatucci, and Zhang, 2009; Zhang, 2006]. When $p = 1$, $q = 2$, the $\ell_{1,2}$-norm has recently been studied in the exclusive-Lasso model [Zhou, Jin, and Hoi, 2010] for the multi-task learning and elitist-Lasso model [Kowalski and Torreesani, 2009] 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.

4.2 Image Annotation with Multi-Semantic Labeling

4.2.1 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 more different spaces 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 space 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 spaces, e.g., object category and emotion. Indeed, for each space k, the label l_i^k can be a multi-label vector in this space. In the following descriptions, for simplicity and clarity purpose, we consider without loss of generality that the labels are obtained from $K = 2$ semantic spaces. 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 chapter. 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

$$\min_{W} \left\{ J(W) := \frac{1}{2}\|Y - WX\|^2 \right\}, \tag{4.1}$$

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 (4.1) can be rewritten as:

$$J(W) = \sum_{j=1}^{|\mathcal{L}|} \frac{1}{2}\|Y_j - W_j X\|^2, \tag{4.2}$$

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 (4.2). 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 space, our objective is to differentiate the related categories. Motivated by the exclusive feature sharing previously considered in [Zhou, Jin, and Hoi, 2010], 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 space, 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 [Zhou, Jin, and Hoi, 2010].

- **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 (4.2). 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 [Cilibrasi and Vitanyi, 2007] 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 (4.2). Figure 4.1 gives an illustration of the pipeline of the proposed framework.

4.2.2 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 spaces, 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:

$$\Omega(W) := \frac{1}{2} \sum_{i=1}^{d} \left(\|W^i_{\mathcal{G}^1}\|^2_{2,1} + \|W^i_{\mathcal{G}^2}\|^2_{2,1} \right), \tag{4.3}$$

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 [Zhou, Jin, and Hoi, 2010]. Similar to the analysis in [Zhou, Jin, and Hoi, 2010], 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.

4.2.3 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

$$\Psi(W) := \frac{1}{2} \sum_{j,k=1}^{|\mathcal{L}|} P_{jk} \parallel W_j - W_k \parallel^2 . \tag{4.4}$$

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 (4.4) as the following compact form

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

Generally speaking, the similarity matrix P can be defined based on any reasonable co-currency measurement such as Google distance [Cilibrasi and Vitanyi, 2007] and Flickr distance [Wu et al., 2008]. In our implementation, P is obtained by applying the Normalized Google similarity Distance (NGD) proposed by Cilibrasi and Vitanyi [Cilibrasi and Vitanyi, 2007]. 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 chapter 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 [Wu et al., 2008].

4.2.4 Graph Regularized Exclusive Group Lasso

Based on the discussion in the previous two subsections, we propose to extend the naive MTL model (4.1) to the following graph regularized exclusive Lasso MTL:

$$\min_{W} \left\{ F(W) := \underbrace{J(W) + \lambda\Omega(W)}_{\text{exclusive group Lasso}} + \underbrace{\gamma\Psi(W)}_{\text{graph regularizer}} \right\}. \tag{4.5}$$

As aforementioned, Equation (4.5) 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 (4.5) 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 (4.5), the parameters W can be learned in a joint manner. It is straightforward to verify that the objective $F(W)$ in (4.5) 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 (4.5). 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.

4.3 Optimization

The non-smooth structure of $\Omega(W)$ makes the optimization of problem (4.5) a non-trivial task. The general purpose subgradient method as used in [Zhou, Jin, and Hoi, 2010] 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 [Nesterov, 2005] to achieve this purpose.

4.3.1 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 [Nesterov, 2005], function $\Omega(W)$ can be approximated by the following smooth function

$$\Omega_\mu(W) = \frac{1}{2} \sum_{i=1}^{d} \left(q^2_{\mathcal{G}^1,\mu}(W^i) + q^2_{\mathcal{G}^2,\mu}(W^i) \right), \tag{4.6}$$

where

$$q_{\mathcal{G}^k,\mu}(W^i) := \max_{\|V^{i,k}_{\mathcal{G}^k}\|_{2,\infty} \leq 1} \langle W^i, V^{i,k} \rangle - \frac{\mu}{2} \|V^{i,k}\|_2^2. \tag{4.7}$$

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

$$\Omega_\mu(W) \leq \Omega(W) \leq \Omega_\mu(W) + C\mu, \tag{4.8}$$

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

Proposition 2 shows that for fixed $\mu > 0$, the function Ω_μ can be seen as a uniform smooth approximation of function Ω.

Proof of Proposition 1

Proof. Since $0 \in \{V^i : \|V^i\|_{2,\infty} \leq 1\}$, by (4.7) we get that for $k = 1, 2$:

$$0 \leq q_{\mathcal{G}^k,\mu}(W^i) \leq \max_{\|V^{i,k}_{\mathcal{G}^k}\|_{2,\infty}\leq 1} \langle W^i, V^{i,k}\rangle = \|W^i_{\mathcal{G}^k}\|_2. \tag{4.9}$$

Therefore by definition of Ω in (4.3) we get the validity of the first inequality in (4.8). Since $\|V^i_{\mathcal{G}^k}\|_{2,\infty} \leq 1$,

$$q_{\mathcal{G}^k,\mu}(W^i) \geq \max_{\|V^{i,k}_{\mathcal{G}^k}\|_{2,\infty}\leq 1} \langle W^i, V^{i,k}\rangle - \frac{\mu}{2} = \|W^i_{\mathcal{G}^k}\|_{2,1} - \frac{\mu|\mathcal{L}^k|^2}{2}. \tag{4.10}$$

Combining (4.9) and (4.10) we get

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

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 (4.8). □

For a fixed W^i, denote $V^{i,k}(W^i)$ the unique minimizer of (4.7) 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 2. *Function $\Omega_\mu(W)$ is well defined, convex and continuously differentiable with gradient*

$$\nabla\Omega_\mu(W) = \left[\nabla\Omega_\mu(W^1), ..., \nabla\Omega_\mu(W^d)\right], \tag{4.11}$$

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

$$\nabla\Omega_\mu(W^i) = q_{\mathcal{G}^1,\mu}(W^i)V^{i,1}(W^i) + q_{\mathcal{G}^2,\mu}(W^i)V^{i,2}(W^i). \tag{4.12}$$

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

$$L_\mu = \left(\frac{2\sqrt{2}R}{\mu} + |\mathcal{L}^1|^2 + |\mathcal{L}^2|^2\right) d. \tag{4.13}$$

Proof of Theorem 1

Proof. Fixe an $i \in \{1, ..., d\}$. Analog to the standard analysis and results (see, e.g. [Nesterov, 2005, 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

$$L_{k,\mu}^i = \frac{1}{\mu}. \tag{4.14}$$

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 (4.12), and consequently (4.11).

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

$$\begin{aligned}
& \|q_{\mathcal{G}^k,\mu}(W^i)\nabla q_{\mathcal{G}^k,\mu}(W^i) - q_{\mathcal{G}^k,\mu}(U^i)\nabla q_{\mathcal{G}^k,\mu}(U^i)\|_2 \\
= \; & \|q_{\mathcal{G}^k,\mu}(W^i)\nabla q_{\mathcal{G}^k,\mu}(W^i) - q_{\mathcal{G}^k,\mu}(W^i)\nabla q_{\mathcal{G}^k,\mu}(U^i) \\
& + q_{\mathcal{G}^k,\mu}(W^i)\nabla q_{\mathcal{G}^k,\mu}(U^i) - q_{\mathcal{G}^k,\mu}(U^i)\nabla q_{\mathcal{G}^k,\mu}(U^i)\|_2 \\
\leq \; & |q_{\mathcal{G}^k,\mu}(W^i)| \cdot \|\nabla q_{\mathcal{G}^k,\mu}(W^i) - \nabla q_{\mathcal{G}^k,\mu}(U^i)\|_2 \\
& + \|\nabla q_{\mathcal{G}^k,\mu}(U^i)\|_2 \cdot |q_{\mathcal{G}^k,\mu}(W^i) - q_{\mathcal{G}^k,\mu}(U^i)| \\
\leq \; & \left(\frac{\sqrt{2}R}{\mu} + |\mathcal{L}^k|^2\right)\|W^i - U^i\|_2 \qquad (4.15)
\end{aligned}$$

where the last equality follows the basic facts: (i) constant in (4.14), (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 (4.6) and (4.15) we establish the validity of (4.13). □

4.3.2 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 (4.5):

$$\min_W \left\{F_\mu(W) := J(W) + \lambda\Omega_\mu(W) + \gamma\Psi(W)\right\}. \qquad (4.16)$$

```
Input: $X \in \mathbb{R}^{d \times n}$, $Y \in \mathbb{R}^{|\mathcal{L}| \times d}$ , $\mathcal{G}^1$, $\mathcal{G}^2$, $\lambda$, $\gamma$, $\mu$.
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 (4.11), (4.12), and $L_\mu$
  according to (4.13).
  $V_{t+1} = V_t - \frac{1}{\alpha_t L_\mu}\big(-(Y - WX)X^T + \lambda\nabla\Omega_\mu(U_t) + \gamma LW\big)$,
  $W_{t+1} = (1 - \alpha_t)W_t + \alpha_t V_{t+1}$,
  $\alpha_{t+1} = \frac{2}{t+1}$, $t \leftarrow t + 1$.
until Converges
```

Algorithm 3: Smooth minimization for Problem (4.16)

Given a fixed $\mu > 0$, by Theorem 2 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 [Nesterov, 2004] and BFGS [Nocedal and Wright, 2006], to optimize the smooth objective (4.16). In our implementation, for simplicity and efficiency, we employ the Accelerated Proximal Gradient method [Tseng, 2008] to optimize the smoothed problem (4.16). The algorithm is formally described in Algorithm 3. 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 3 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 [Becker, Bobin, and Candes, 2011].

4.4 Experiments

To validate the effectiveness of IA-MSL, we conduct extensive experiments on two large scale image datasets: NUS-WIDE-Emotive; and NUS-WIDE-Object&Scene [Chua et al., 2009]. 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 [Chua et al., 2009] obtained after noisy tag removal. It is also annotated in two sematic spaces: 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.

4.4.1 Datasets

NUS-WIDE-Emotive dataset is an emotion version of the publicly available NUS-WIDE-LITE [Chua et al., 2009] 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 the 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 currently available from NUS-WIDE. For the emotive semantic concepts, we adopt the similar categories as studied in [Machajdik and Hanbury, 2010; Mikels et al., 2005]: Amusement, Awe, Contentment, Excitement, Anger, Disgust, Fear, Sad to represent the 8 different types of positive and negative emotions. 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, 75-D edge direction histogram, 64-D color histogram, 144-D color correlogram and 500-D bag of visual words [Chua et al., 2009].

NUS-WIDE-Object&Scene [Chua et al., 2009] are two subsets from NUS-WIDE. In this chapter, 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.

4.4.2 Baselines and Evaluation Criteria

We systematically compare our proposed IA-MSL with six baseline algorithms as listed in Table 4.1. Amongst them,

Table 4.1: The baseline algorithms.

Name	Methods
`SVM`	Support Vector Machine
`SVM-E`	The enrichment of SVM from individual spaces.
`NMTL`	Naive MTL as in (4.2)
`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

- 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 (4.2).

- The SVM-E and NMTL-E are two enrichment (recall the definition of enrichment method in Section 4.1) methods of SVM and NMTL, respectively.

- 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 (4.5), 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 4.2, on each semantic space of NUS-WIDE-Emotive.

Many measurements are used to evaluate multi-label 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) [Hanley and

Table 4.2: The baseline algorithms for comparison in individual semantic spaces of NUS-WIDE-Emotive.

Name	Methods
SVM	Support Vector Machine
LNP	Linear Neighborhood Propagation [Wang and Zhang, 2006]
EGSSC	Entropic Graph Classification [Subramanya and Bilmes, 2009]
LSMP	Large-scale Multi-label Propagation [Chen et al., 2010]

McNeil, 1982], 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 desktop PC equipped with Intel dual-core CPU (frequency: 3.0 GHz) and 32G bytes physical memory.

4.4.3 Experiment-I: NUS-WIDE-Emotive

On NUS-WIDE-Emotive, we category all labels into 648 (8 emotions × 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 4.1. Table 4.3 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 simultaneously outperforms the competing methods in MAUC and AUCs on all of the 8 emotive classes.

Table 4.3: 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**

- 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 bypass this problem by training and 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 4.3, we illustrate in Figure 4.2 the objective value ($F_\mu(W)$) in (4.16) 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

Table 4.4: 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**

Table 4.5: 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**

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 about 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 4.1. Table 4.4 lists the results for 8 emotive classes. Table 4.5 lists the corresponding results for 81 cognitive object categories. To make the table compacter, we sort the 81 concepts according to the descending order of training sample number and evenly divide them into 8 groups. The AUCs in Table 4.5 are obtained by averaging over each of these 8 concept groups. From the results in both tables we can see that IA-MSL also outperforms the base-

Table 4.6: The unitary semantic annotation results on NUS-WIDE-LITE.

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

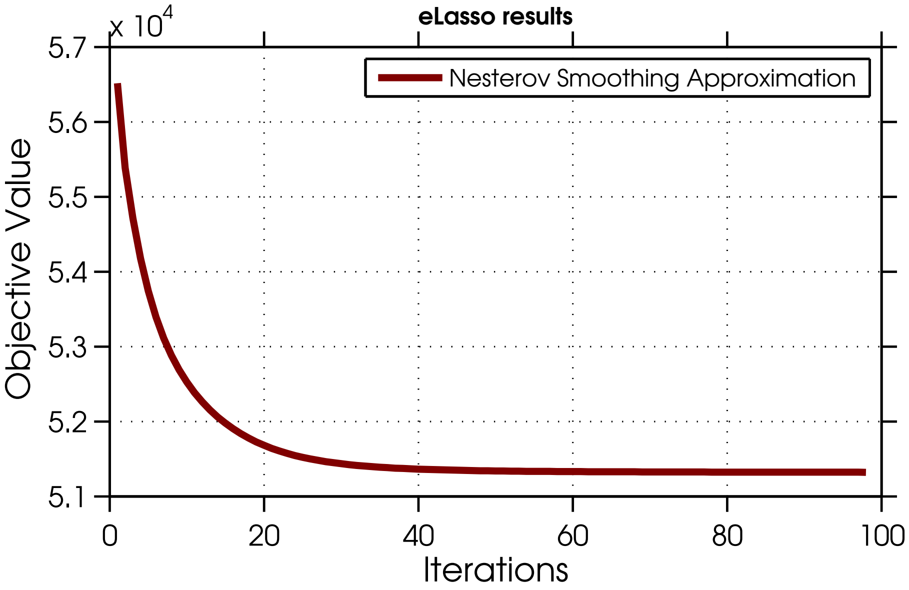


Figure 4.2: Convergence curve of IA-MSL on NUS-WIDE-EMOTIVE dataset.

lines 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 4.6. LSMP is the algorithm focusing on large-scale multi-label image annotation, which will be introduced in Chapter 5. 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 4.3.

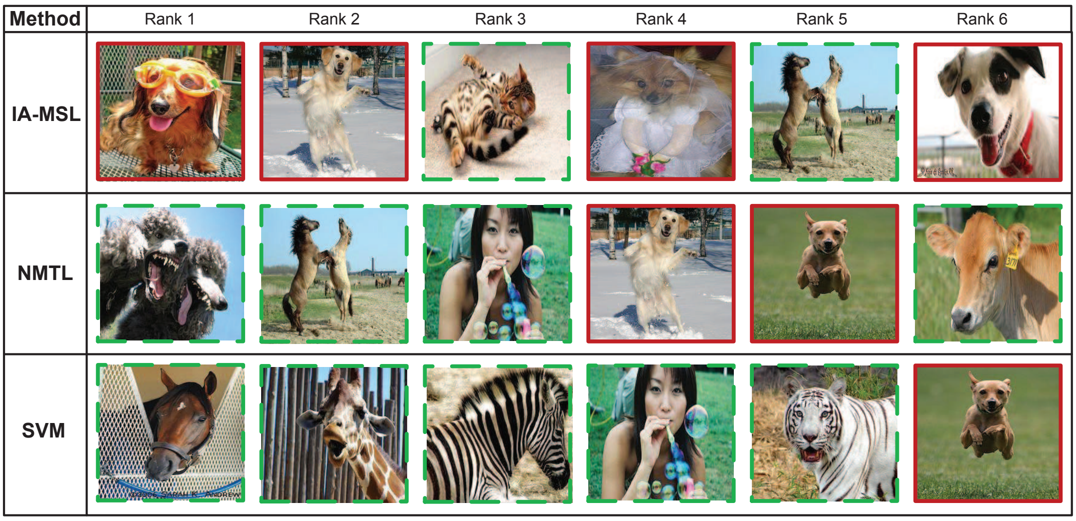


Figure 4.3: Some exemplar results of query search and ranking by IA-MSL (top row), NMTL (middle row) and SVM (bottom row) on NUS-WIDE-Emotive with the query: "Amusement Dog". The red border indicates correct result while the green one incorrect.

4.4.4 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 4.1. Table 4.7 lists the quantitative results. To make the results more compactly, we sort the 1033 concepts in the descent order of training sample number and evenly divide them into 5 groups. The AUCs in Table 4.7 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 noted that on Group 5, all the involved

Table 4.7: 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**

Table 4.8: 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**

comparing algorithms return AUC value 0. This is unsurprising since Group 5 is composed of those concepts with very few or even zero training samples, and thus all the algorithms fail including. A typical running time for training on this dataset is about 470 seconds. The per query time of IA-MSL is about 0.08 second.

Specially, in the setting of unitary semantic image annotation, we have also compared IA-MSL with the algorithms listed in Table 4.2. Table 4.8 and Table 4.9 list the corresponding results for 31 objects and 33 scenes, respectively. In order to present the concise and compact results, we sort both the 31 objects and 33 scenes based on 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, we observe again that IA-MSL also outperforms the baselines for unitary semantic annotation.

Table 4.9: 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**

4.5 Conclusion

In this chapter, we proposed the IA-MSL method to explore multi-semantic meaning of images based on two or more semi-orthogonal label spaces from multi-semantic. We formulated this challenging problem as a multi-task discriminative analysis model, where individual tasks are defined by learning the linear discriminative model for individual complex semantic concepts. We considered all the tasks in a joint manner by imposing two types of regularization on parameters, the graph Laplacian regularization and exclusive group lasso regularization. A Nesterov-type smoothing approximation method is developed for model optimization. The proposed algorithm was tested on two image benchmarks built for multi-semantic annotation. We demonstrated 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 aforementioned 8 Emotive Categories to expand our search range towards real world search scenario.

Chapter 5

Multi-Label Learning in Large-Scale Dataset

5.1 Introduction

Generally, there are three crucial subtasks in graph-based multi-label learning algorithms: 1) graph construction; 2) the choice of loss function; and 3) the choice of regularization term. As argued in [Zhu, 2005], graph construction is supposed to be more dominating than the other two factors in terms of performance. Unfortunately, it is also the area that is most inadequately studied. In Section 5.3.2, we propose a novel hashing-based scheme for efficient large-scale graph construction. The solutions to the last two subtasks may affect the final accuracy as well as the proper optimization strategy (thus the convergence speed). As reported in [Delalleau, Bengio, and Le Roux, 2005], early work on semi-supervised learning can only handle $10^2 \sim 10^4$ unlabeled samples. Consequently, a large number of recent endeavors has been devoted to the scalability of semi-supervised learning

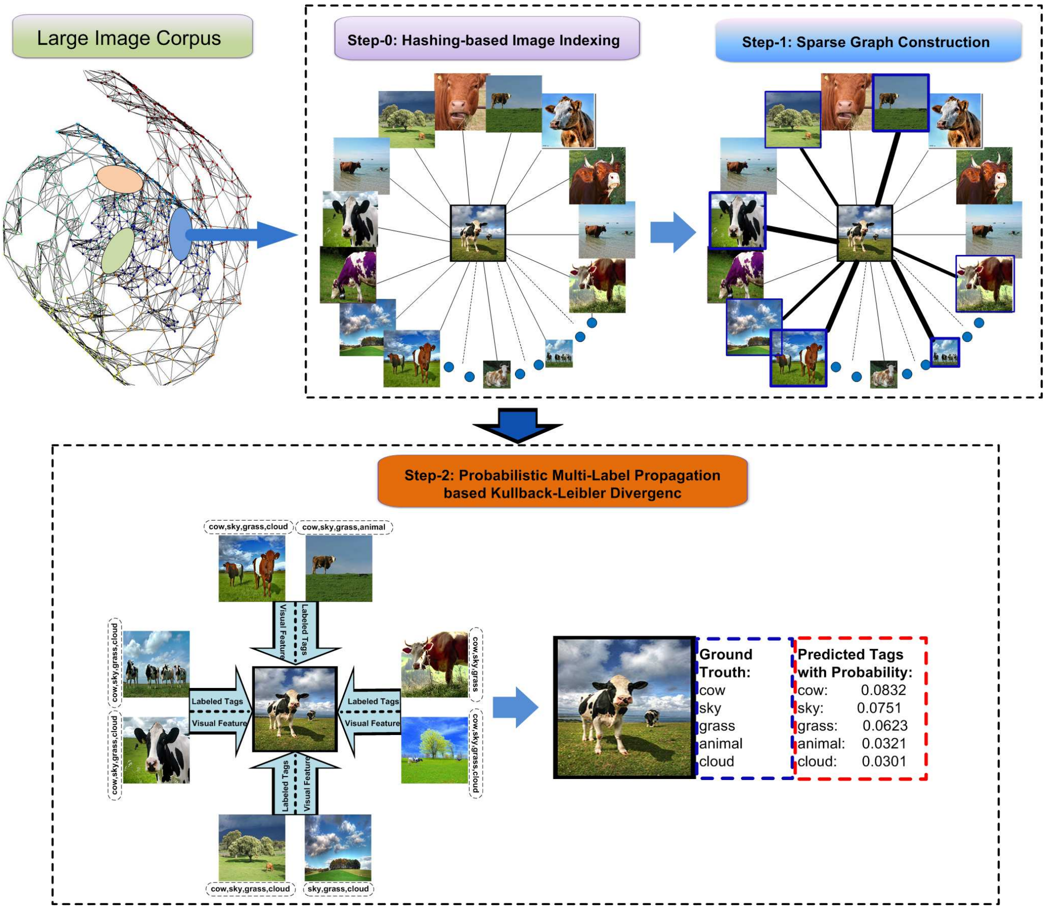


Figure 5.1: Flowchart of our proposed scheme for multi-label propagation. Step-0 and step-1 are the proposed hashing-based l_1-graph construction scheme, which perform neighborhood selection and weight computation respectively; Step-2 is the probabilistic multi-label propagation based Kullback-Leibler divergence.

methods to large-scale datasets.

The seminal work in [Subramanya and Bilmes, 2009] is most similar to our work in this chapter. Unlike previous approaches, this method models the multi-class label confidence vector as a probabilistic distribution, and utilizes the Kullback-Leibler (KL) divergence to gauge the pairwise discrepancy. The underlying philosophy is that such soft regularization term will be less vulnerable to noisy annotation or outliers. Here we adopt the same representation and distance measure, but apply it to a different scenario of multi-label image annotation, which demands a new solution.

Several algorithms were recently proposed to exploit the inter-relations among different labels [Liu et al., 2009]. For example, Qi et al. [Qi et al., 2007] proposed a unified Correlative Multi-Label (CML) framework to simultaneously classify labels and model correlations between them. Chen et al. [Chen et al., 2008] formulated this problem as a sylvester equation. They first constructed two graphs at the sample level and category level associated with a quadratic energy function respectively, and then obtain the labels of the unlabeled images by minimizing the combination of the two energy functions. Liu et. al. [Liu, Jin, and Yang, 2006] utilized constrained nonnegative matrix factorization (CNMF) to optimize the consistency between image similarity and label similarity. Unfortunately, most of the aforementioned algorithms are of high complexity and unsuitable to scale up to the large-scale datasets.

5.2 Motivation

Most existing works in the line of graph-based label propagation suffer (or partially suffer) from these disadvantages: 1) they consider each tag independently when handling multi-label propagation problem, 2) the derived labels for one image are not rankable, and 3) the graph construction process is time-consuming. And as reviewed in Section 2.3 most recent large-scale algorithms focus on the single label case, but the scalability to large number of labels is unclear. To address the above issues, we propose a new large-scale graph-based multi-label propagation approach by minimizing the Kullback-Leibler divergence of the image-wise label confidence vector and its propagated version via the so-called hashing-based ℓ_1-graph, which is efficiently derived with Locality Sensitive Hashing approach followed by sparse ℓ_1-graph construction within the individual hashing buckets. Finally, an efficient and convergence provable iterative procedure is presented for problem optimization. The main contributions of our proposed scheme can be summarized as follows:

- We propose a probabilistic collaborative multi-label propagation formulation for large-scale image annotation, which is founded on Kullback-Leibler divergence based label similarity measurement and scalable ℓ_1-graph construction.

- We also propose a novel hashing-based scheme for efficient large-scale graph construction. *Locality sensitive hashing* [Indyk and Motwani, 1998; And, ; Mu, Shen, and Yan, 2010] is utilized to speed up the candidate selection of similar neighbors for one image, which makes the ℓ_1-graph construction process scalable.

The remainder of this chapter is organized as follows. In Section 5.3, we elaborate on the proposed probabilistic collaborative multi-label propagation (LSMP) algorithm. Section 5.4 presents analysis on algorithmic complexity and convergence properties. Experimental results on both middle-scale and large-scale image datasets are reported in Section 5.5. Section 5.6 concludes this work along with future work discussion.

5.3 Large-Scale Multi-Label Propagation

5.3.1 Scheme Overview

Our proposed large-scale multi-label propagation framework includes three concatenating parts: 1) An efficient k-nearest-neighbor (k-NN) search based on *locality sensitive hashing* (LSH) approach; 2) sparse ℓ_1-graph construction within hashing buckets; and 3) multi-label propagation based on Kullback-Leibler divergence. Figure 5.1 gives an illustration of the algorithmic pipeline.

5.3.2 Hashing-based ℓ_1-Graph Construction

The first step of the proposed framework is the construction of an directed weighted graph $\mathcal{G} =< V,\ E >$, where the cardinality of the node set V is $m = l + u$ (denote the labeled and unlabeled data respectively), and the edge set $E \subseteq V \times V$ describes the graph topology. Let V_l and V_u be the sets of labeled and unlabeled vertices respectively. $\mathcal{G}$ can be equivalently represented by a weight matrix $\mathbf{W} = \{w_{ij}\} \in \mathbb{R}^{m \times m}$. To efficiently handle the large-scale data, we enforce the constructed graph to be sparse. The weight between two nodes w_{ij} is nonzero only when $j \in \mathcal{N}_i$, where $\mathcal{N}_i$ denotes the local neighborhood of the i-th image. The graph construction can thus be decomposed into two sub-problems: 1) how to determine the neighborhood of a datum; and 2) how to compute the edge weight w_{ij}.

5.3.2.1 Neighborhood Selection

For the first problem, the conventional strategies in previous work can be roughly divided into two categories:

- k-nearest-neighbor based neighborhood: w_{ij} is nonzero only if x_j is among the k-nearest neighbors to the i-th datum. Obviously, graphs constructed in this way may ensure a constant vertex degree, avoiding over-dense subgraphs and isolated vertices.

- ϵ-ball neighborhood: given a pre-specified distance measure between two nodes $d_{\mathcal{G}}(x_i,\ x_j)$ and a threshold ϵ. Any vertex x_j that satisfies $d_{\mathcal{G}}(x_i,\ x_j) \leq \epsilon$ will be incorporated in the neighborhood of the vertex x_i, resulting in nonzero w_{ij}. It is easy to observe that the weight matrix of the constructed graph is symmetric. However, for some vertices beyond a distance from the others, there is probably no edge connecting to other vertices.

Although dominating the graph-based learning literature, the above two schemes are both computation-intensive on large-scale dataset, since a linear scan is required to process a single sample and the overall complexity is $\mathcal{O}(n^2)$ (n is the number of all samples). For a typical image data set to annotate, there are $10^4 \sim 10^5$ images, from each of which high-dimensional features are extracted. A naive implementation based on either of these two schemes usually takes several days to accomplish graph construction, which is definitely unaffordable in terms of efficacy. Instead, in our implementation we use the *locality-sensitive hashing* (LSH) to enhance the efficacy on large-scale data sets.

The basic idea of LSH is to store proximal samples into the same bucket, which greatly saves the retrieval time at the expense of additional storage of hash bits. LSH is a recently proposed hashing algorithm family. The most attractive property of LSH is the theoretic guarantee that the collision probability of two samples (i.e., projected into the same bucket) is proportional to their similarity in feature space. The most popular LSH approach relies on random projection followed by a threshold-based binarization. Formally, given a random projection direction v, the whole dataset is splitted into two half-spaces, according to the rule $h(x_i) = \text{Boolean}(v^T x_i > 0)$. The hash table typically consists of k independent bits, namely the final hash bits are obtained via sequential concatenation $H(x_i) = \langle h_1(x_i), \ldots, h_k(x_i) \rangle$. In the retrieval phase, the k-NN candidate set can be safely confined to be the buckets whose Hamming distances to the query sample are below a pre-specified small threshold. Prior investigation at the theoretic aspect reveals that a sublinear retrieval complexity is feasible by the LSH method, which is a crucial acceleration for the scenario of large-scale image search. Note that in our implementation, LSH is run for multiple times in all the experiments, and the neighborhoods are the combined to avoid the case of isolated subgraphs.

5.3.2.2 Weight Computation

A proper inter-sample similarity definition is the core for graph-based label propagation. The message transmitted from the neighboring vertices with higher weights will be much stronger than the others. Generally, the more similar a sample is to another sample, the stronger the interaction (thus larger weight) exists between them. Below are some popular ways to calculate the pairwise weights:

- *Unweighted k-NN similarity*: The similarity w_{ij} between x_i and x_j is 1 if x_j is among the k-NN of x_i; otherwise 0. For undirected graph, the weight matrix is symmetric and therefore $w_{ij} = w_{ji}$ is enforced.

- *Exponentially weighted similarity*: For all chosen k-NN neighbors, their weights are determined as below:

$$w_{ij} = \exp\left(-\frac{d_{\mathcal{G}}(x_i, x_j)}{\sigma^2}\right), \tag{5.1}$$

where $d_{\mathcal{G}}(x_i, x_j)$ is the ground truth distance and σ is a free parameter to control the decay rate.

- *Weighted linear neighborhood similarity* [Roweis and Saul, 2000; Wang and Zhang, 2006]: In this scheme sample x_i is assumed to be linearly reconstructed from its k-NN. The weights are obtained via solving the following optimization problem:

  $$\min_{w_{ij}} \parallel x_i - \sum_{j \in \mathcal{N}_i} w_{ij} x_j \parallel^2 . \tag{5.2}$$

  Typically additional constraints are given to w_{ij}. For example, in [Wang and Zhang, 2006], the constraints $w_{ij} \geq 0$ and $\sum_j w_{ij} = 1$ are imposed.

In our implementation, we adopt a scheme similar to the idea in [Roweis and Saul, 2000; Wang and Zhang, 2006], based on the linear reconstruction assumption. Moreover, prior work [Tang et al., 2009] reveals that minimizing the ℓ_1 norm over the weights is able to suppress the noise contained in data. The constructed graph is non-parametric and is comparably more robust than the other graph construction strategies. Meanwhile, the graph constructed by datum-wise one-vs-all sparse reconstruction of samples can remove considerable label-unrelated links between those semantically unrelated samples to reduce the incorrect information for label propagation.

Suppose we have an over-determined system of linear equations:

$$\begin{bmatrix} x_{i_1} & x_{i_2} & \cdots & x_{i_k} \end{bmatrix} \times \mathbf{w}_i = x_i, \tag{5.3}$$

where x_i is the feature vector of the i-th image to be reconstructed, $\mathbf{w}_i$ is the vector of the unknown reconstruction coefficients. Let $X \in \mathbb{R}^{d\times k}$ be a data matrix, each column of which corresponds to the feature vector of one of its k-NN. In practice, there are probably noises in the features, and a natural way to recover these elements and provide a robust estimation of $\mathbf{w}_i$ is to formulate $x_i = X\mathbf{w}_i + \xi$, where $\xi \in \mathbb{R}^d$ is the sparse noise term. We can then solve the following l_1-norm minimization problem with respect to both reconstruction coefficients and feature noise:

$$\begin{aligned} \arg_{w,\ \xi} \min \quad & \parallel \xi \parallel_1 \\ s.t. \quad & x_i = X\mathbf{w}_i + \xi, \\ & \mathbf{w}_i \geq \mathbf{0}, \quad \parallel \mathbf{w}_i \parallel_1 = 1. \end{aligned} \tag{5.4}$$

This optimization problem is convex and can be transformed into a general linear programming problem. There exists a globally optimal solution, and the optimization can be solved efficiently using many available l_1-norm optimization toolboxes like ℓ_1-MAGIC [Candès, Romberg, and Tao, 2006].

5.3.3 Problem Formulation

Let $M_l = \{x_i, r_i\}_{i=1}^{l}$ be the set of labeled images, where x_i is the feature vector of the i-th image and r_i is a multi-label vector (its entry is set to be 1 if it is assigned with the corresponding label, otherwise 0). Let $M_u = \{x_i\}_{i=l+1}^{l+u}$ be the set of unlabeled images, and $M = \{M_l, M_u\}$ is the entire data set. The graph-based multi-label propagation is intrinsically a transductive learning process, which propagates the labels of M_l to M_u.

For each x_i, we define the probability measure p_i over the measurable space

$(Y, \mathcal{Y})$. Here $\mathcal{Y}$ is the σ-field of measurable subsets of Y and $Y \subset \mathbb{N}$ (the set of natural numbers) is the space of classifier outputs. $|Y| = 2$ yields binary classification while $|Y| > 2$ implies multi-label. In this paper, we focus on the multi-label case. Hereafter, we use p_i and r_i for the i-th image, both of which are subject to the multinomial distributions, and $p_i(y)$ is the probability that x_i belongs to class y. As mentioned above, $\{r_j, j \in V_l\}$ encodes the supervision information of the labeled data. If it is assigned a unique label by the annotator, r_j becomes the so-called "one-hot" vector (only the corresponding entry is 1, the rest is 0). In case being associated with multiple labels, r_j is represented to be a probabilistic distribution with multiple non-zero entries.

We propose the following criterion to guide the propagation of the supervision information, which is based on the concept of KL divergence defined on two distributions:

$$D_1(p) = \sum_{l=1}^{l} D_{KL}\big(r_i \parallel p_i\big) + \mu \sum_{i=1}^{m} D_{KL}\big(p_i \parallel \sum_{j \in N(i)} w_{ij} p_j\big), \tag{5.5}$$

and the optimal solution $p^* = \arg_p \min D_1(p)$.

Here $D_{KL}(r_i \parallel p_i)$ denotes the KL divergence between r_i and p_i, whose formal definition for the discrete case is expressed as $D_{KL}(r_i \parallel p_i) = \sum_y r_i(y) \log \frac{r_i(y)}{p_i(y)}$. The first term in $D_1(p)$ trigger a heavy penalty if the estimated value p_i deviates from the pre-specified r_i. Note that unlike most traditional approaches, there is no constraint for the rigid equivalence between p_i and r_i. Such a relaxation is able to mitigate the bad effect of noisy annotations. The second term of D_1 stems from the assumption that p_i can be linearly reconstructed from the estimations of its neighbors, thus penalizing the inconsistency between the p_i and its neighborhood estimation. Unlike previous works [Wang and Zhang, 2006] using squared-error (optimal under a Gaussian loss assumption), the adopted KL-based loss penalizes *relative error* rather than *absolute error* in the squared-error case. In other words, they can be regarded as the regularization terms from prior supervision and local coherence respectively. μ is a free parameter to balance these two terms.

If $\mu, w_{ij} \geq 0$, then $D_1(p)$ is convex. Since no closed-form solution is feasible, standard numerical optimization approaches such as interior point methods (IPM) or method of multipliers (MOM) can be used to solve the problem. However, most of these approaches guarantee global optima yet are tricky to implement (e.g., an implementation of MOM to solve this problem would have seven extraneous parameters) [Subramanya and Bilmes, 2009]. Instead, we utilize a simple alternating minimization method in this work.

Alternating minimization is an effective strategy to optimize functions of the form $f(x,\ y)$ where x, y are two sets of variables. In many cases, simultaneous optimizing over x and y is computationally intractable or unstable, while optimizing over one set of variables with the other fixed is relatively easier. Formally, a typical alternating minimization loops over two sub-problems, i.e., $x^{(t)} = \arg_x \min f(x,\ y^{(t-1)})$ and $y^{(t)} = \arg_y \min f(x^{(t)}, y)$. An example for alternating optimization is the well-known Expectation-Maximization (EM) algorithm. Note that D_1 in Equation (5.5) is not amenable to alternating optimization. We further propose a modified version by introducing a new group of variables $\{q_i\}$, which is shown as below:

$$\begin{aligned} D_2(p,q) &= \sum_{l=1}^{l} D_{KL}(r_i \parallel q_i) + \mu \sum_{i=1}^{m} D_{KL}(p_i \parallel \sum_{j\in\mathcal{N}(i)} w_{ij} q_j) \\ &\quad + \eta \sum_{i=1}^{m} D_{KL}(p_i \parallel q_i). \end{aligned} \tag{5.6}$$

In the above, a third measure q_i is introduced to decouple the original term

$\mu \sum_{i=1}^{m} D_{KL}\big(p_i \parallel \sum_{j \in N(i)} w_{ij} p_j\big)$. q_i can actually be regarded as a relaxed version of p_i. To enforce consistency between them, the third term $\sum_{i=1}^{m} D_{KL}(p_i \parallel q_i)$ is incorporated. The proof of convexity of $D_1(p)$ and $D_2(p, q)$ is given below.

Proof of Convexity of $D_1(p)$ and $D_2(p, q)$

Proof. The convexity of $D_1(p)$ is obvious if $D_{KL}(r_i \parallel p_i)$ and $D_{KL}(p_i \parallel \sum_{j \in N(i)} w_{ij} p_j)$ prove convex. Consequently, to justify the convexity of $D_1(p)$, first we elaborate on the convexity of KL divergence defined on two probability mass functions, which has already been studied in the fields of both information theory [Cover and Thomas, 1991] and convex optimization [Boyd and Vandenberghe, 2004].

Specifically, for $D_{KL}(p \parallel q)$ defined on two pairs of probability mass functions (p_1, q_1) and (p_2, q_2), the convexity of D_{KL} equivalently implies the following fact:

$$\begin{aligned} & D_{KL}(\lambda p_1 + (1-\lambda) p_2 \parallel \lambda q_1 + (1-\lambda) q_2) \leq \lambda D_{KL}(p_1 \parallel q_1) \\ & + (1-\lambda) D_{KL}(p_2 \parallel q_2), \end{aligned} \tag{5.7}$$

where $\lambda \in [0, 1]$. The correctness of the above inequality is clear by applying the log-sum inequality [Cover and Thomas, 1991], i.e.,

$$\left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \leq \sum_{i=1}^{n} a_i \log \frac{a_i}{b_i},$$

on both the left and right sides of the following inequality:

$$\begin{aligned} & D_{KL}(\lambda p_1 + (1-\lambda) p_2 \parallel \lambda q_1 + (1-\lambda) q_2) = \\ & \sum_{y} (\lambda p_1(y) + (1-\lambda) p_2(y)) \log \frac{\lambda p_1(y) + (1-\lambda) p_2(y)}{\lambda q_1(y) + (1-\lambda) q_2(y)}. \end{aligned}$$

It is easily verified that

$$\begin{aligned} & D_{KL}(\lambda p_1 + (1-\lambda)p_2 \parallel \lambda q_1 + (1-\lambda)q_2) \le \\ & \sum_y \lambda p_1(y) \log \frac{\lambda p_1(y)}{\lambda q_1(y)} + \sum_y (1-\lambda)p_2(y) \log \frac{(1-\lambda)p_2(y)}{(1-\lambda)q_2(y)} \\ & = \lambda D_{KL}(p_1 \parallel q_1) + (1-\lambda) D_{KL}(p_2 \parallel q_2). \end{aligned} \tag{5.8}$$

Thus $D_{KL}(r_i \parallel p_i)$ is convex. And likewise the convexity of $D_{KL}(p_i \parallel \sum_{j\in N(i)} w_{ij}p_j)$ can be justified, observing that $\sum_{j\in N^k(i)} w_{ij}p_j$ is a convex, linear combination of several variables. Hence $D_1(p)$ is convex.

Using similar tricks, $D_2(p,q)$ is also demonstrated to be convex. □

5.3.4 Part I: Optimize p_i with q_i Fixed

With $\{q_i,\ i = 1\ldots m\}$ fixed, the optimization problem is reduced to the following form:

$$\begin{aligned} p^* &= \arg_p \min D_2(p,q) && (5.9) \\ s.t. & \quad \sum_y p_i(y) = 1,\ p_i \ge \mathbf{0},\ \ \forall\ i. \end{aligned}$$

The above constrained optimization problem can be easily transformed into an unconstrained one using the Lagrange multiplier:

$$p^* = \arg_p \min D_2(p,q) + \sum_{i=1}^m \lambda_i (1 - \sum_y p_i(y)). \tag{5.10}$$

For brevity, let $\mathcal{L}_p \triangleq D_2(p,q) + \sum_{i=1}^m \lambda_i(1 - \sum_y p_i(y))$. Recall that any locally optimal solutions should be subject to the zero first-order derivative, i.e.,

$$\begin{aligned} \frac{\partial \mathcal{L}_p}{\partial p_i(y)} &= \mu\big(\log p_i(y) + 1 - \log \sum_{j\in\mathcal{N}(i)} w_{ij}q_j(y)\big) \\ &\quad + \eta\big(\log p_i(y) + 1 - \log q_i(y)\big) - \lambda_i \\ &= 0. \end{aligned} \tag{5.11}$$

From Equation (5.11), it is easily verified that (let $\gamma = \mu + \eta$):

$$p_i(y) = \exp\left(\frac{\mu \log \sum_{j \in \mathcal{N}(i)} w_{ij} q_j(y) + \eta \log q_i(y) - \gamma + \lambda_i}{\gamma}\right).$$

Recall that λ_i is the Lagrange coefficient for the i-th sample and unknown. Based on the fact $\sum_y p_i(y) = 1$, λ_i can be eliminated and finally we obtain the updating rule:

$$p_i(y) = \frac{\exp\left(\frac{\mu}{\gamma} \log \sum\limits_{j \in \mathscr{N}(i)} w_{ij} q_j(y)) + \frac{\eta}{\gamma} \log q_i(y)\right)}{\sum_y \exp\left(\frac{\mu}{\gamma} \log \sum\limits_{j \in \mathscr{N}(i)} w_{ij} q_j(y) + \frac{\eta}{\gamma} \log q_i(y)\right)}. \tag{5.12}$$

5.3.5 Part II: Optimize q_i with p_i Fixed

The other step of the proposed alternating optimization is to update q_i with p_i fixed. Unfortunately, it proves that the same trick used in subsection 5.3.4 cannot be applied to the optimization of q_i, due to the highly non-linear term $\log\left(\sum_{j \in \mathcal{N}_i} w_{ij} q_j(y)\right)$. To ensure that q_i is still a valid probability vector after updating, we set the updating rule as:

$$q_i^{new} = q_i^{old} + U\boldsymbol{h}, \tag{5.13}$$

where the column vector of matrix $U \in \mathbb{R}^{d \times (d-1)}$ is constrained to be summed 0. Denote $\boldsymbol{e}$ to be a column vector with its all entries equal to 1, then we have $\boldsymbol{e}^T U = \boldsymbol{0}$. An alternative view of this relationship is that U is the complementary subspace of the one spanned by $\frac{1}{\sqrt{n}}\boldsymbol{e}$, thus $UU^T = I - \frac{1}{n}\boldsymbol{e}\boldsymbol{e}^T$ also holds.

Vector $\boldsymbol{h}$ in each iteration should be carefully chosen so that the updated value of q_i^{new} results in a non-trivial decrease of the overall objective function. Denote $\mathcal{L}_q \triangleq D_2(p, q)$ and the value of q_i at the t-th iteration as $q_i^{(t)}$, we have

$$\nabla \mathcal{L}_h(q_i^{(t)}) \triangleq \frac{\partial \mathcal{L}_q(q_i^{(t)} + U^T\boldsymbol{h})}{\partial \boldsymbol{h}} = U^T \frac{\partial \mathcal{L}_q}{\partial q_i}\Big|_{q_i = q_i^{(t)}}. \tag{5.14}$$

Note that in each iteration $\boldsymbol{h}$ is typically initialized as 0, thus $h = -\alpha \nabla \mathcal{L}_h(q_i^{(t)})$ is a candidate descent direction (α is a parameter to control the step size). By substituting it into Equation (5.13), we obtain the following updating rule:

$$\begin{aligned} q_i^{(t+1)} &= q_i^{(t)} - \alpha U U^T \frac{\partial \mathcal{L}_q}{\partial q_i}\Big|_{q_i = q_i^{(t)}} \\ &= q_i^{(t)} - \alpha (I - \frac{1}{n} \boldsymbol{e}\boldsymbol{e}^T) \frac{\partial \mathcal{L}_q}{\partial q_i}\Big|_{q_i = q_i^{(t)}}. \end{aligned} \tag{5.15}$$

```
Input: An directed weighted sparse graph $\mathcal{G} =< V,\ E >$ of the
whole image dataset $M = \{M_l, M_u\}$, where $M_l = \{x_i, r_i\}_{i=1}^{l}$ is
the labeled image set and $M_u = \{x_i\}_{i=l+1}^{l+u}$ is the set of unlabeled
images. $x_i$ is the feature vector of the $i$-th image and $r_i$ is a multi-
label confidence vector for $x_i$.
Output: The convergent probability measures $p_i$ and $q_i$.
Initialization: Randomly initialize $\{p_i \geq 0,\ \sum_y p_i(y) = 1\}$ and
$\{q_i \geq 0,\ \sum_y q_i(y) = 1\}$.
for $p_i$ and $q_i$ are not convergent do
    Optimize $p_i$ with $q_i$ Fixed:
    $p_i(y) = \frac{\exp\left(\frac{\mu}{\gamma} \log \sum_{j \in \mathcal{N}(i)} w_{ij} q_j(y)) + \frac{\eta}{\gamma} \log q_i(y)\right)}{\sum_y \exp\left(\frac{\mu}{\gamma} \log \sum_{j \in \mathcal{N}(i)} w_{ij} q_j(y) + \frac{\eta}{\gamma} \log q_i(y)\right)}.$
    Optimize $q_i$ with $p_i$ Fixed:
    $q_i^{(t+1)} = q_i^{(t)} - \alpha(I - \frac{1}{n}\boldsymbol{e}\boldsymbol{e}^T)\frac{\partial \mathcal{L}_q}{\partial q_i}$, where $\alpha$ lies in the range defined in
    Equation (5.18).
end for
```

Algorithm 4: Probabilistic Collaborative Multi-Label Propagation

In this way, the pursuit of the descent direction with respect to q_i is transformed into an equivalent problem taking $\boldsymbol{h}$ as variable, which is further solved by calculating $\frac{\partial \mathcal{L}_q}{\partial q_i}$. For completeness, we list the concrete value of an entry of $\frac{\partial \mathcal{L}_q}{\partial q_i}$:

$$\frac{\partial \mathcal{L}_q}{\partial q_i(y)} = -\frac{r_i(y)}{q_i(y)} - \mu \sum_{\forall k:\ i \in \mathcal{N}_k} \frac{w_{ki} p_k(y)}{\sum_{j \in \mathcal{N}_k} w_{kj} q_j(y)} - \eta \frac{p_i(y)}{q_i(y)}. \tag{5.16}$$

One practical issue is the feasible region of parameter α. An arbitrary α probably cannot ensure that the updated $p_i^{(t+1)}$ in Equation (5.15) stays within

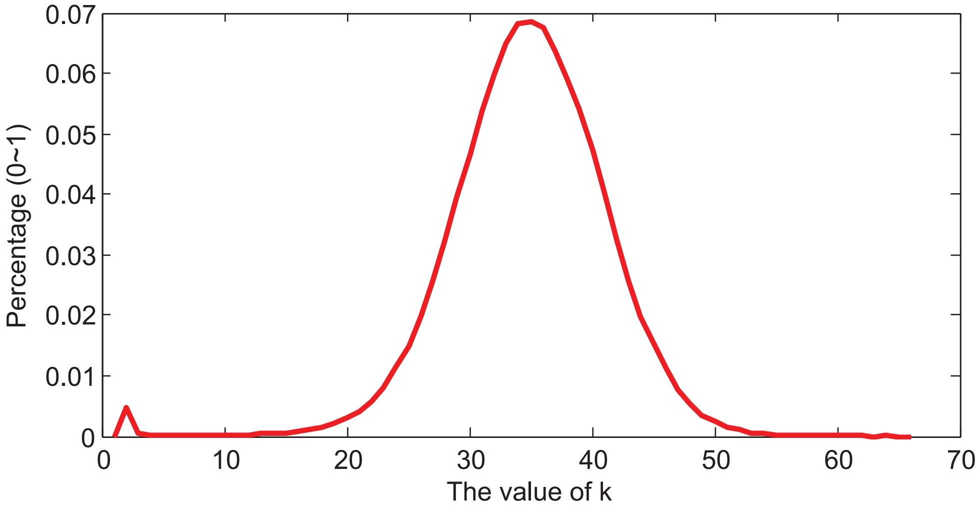


Figure 5.2: The distribution of the number of nearest neighbors (denote as k) in our proposed LSMP.

the range $[0, 1]$. A proper value of α should ensure:

$$\mathbf{0} \leq q_i - \alpha U U^T \frac{\partial \mathcal{L}_q}{\partial q_i}\Big|_{q_i = q_i^{(t)}} \leq \mathbf{1}. \tag{5.17}$$

Denote $\boldsymbol{v} = UU^T \frac{\partial \mathcal{L}_q}{\partial q_i}\big|_{q_i = q_i^{(t)}}$. It is easy to verify that

$$0 \leq \alpha \leq \min\left\{\max\left\{\frac{q_i(y)}{\boldsymbol{v}(y)}, \frac{q_i(y) - 1}{\boldsymbol{v}(y)}, \epsilon\right\}\right\}. \tag{5.18}$$

In practice, α can be adaptively determined from $q_i^{(t)}$. The whole process of optimization is illustrated in Algorithm 4. The resultant p_i is adopted to infer the image tags, as it connects both r_i and q_i.

5.4 Algorithmic Analysis

5.4.1 Computational Complexity

Overall speaking, the computational complexity of the proposed algorithm consists of two components: the cost of hashing-based ℓ_1-graph construction, and the cost of KL-based label propagation. The efficacy of traditional graph construction as in [Yuan, Li, and Zhang, 2007; Tang et al., 2009] hinges on the complexity of k-NN retrieval, which is typically $\mathcal{O}(n^2)$ (n is the number of images) for a naive linear-scan implementation. Our proposed LSH-based scheme guarantees a sublinear complexity by aggregating visually similar images into the same buckets, greatly reducing the cardinality of the set of candidate neighbors. Formally, recent work points out the lower bound of LSH is only slightly high than $\mathcal{O}(n \log(n))$, which drastically reduces the computational overhead of graph construction compared with traditional $\mathcal{O}(n^2)$ complexity.

On the other hand, for our proposed KL-guided label propagation procedure, it has $\mathcal{O}(n\ k\ l)$ computation in each iteration, where k denotes the averaged number of nearest neighbors for a graph vertex and l is the total number of labels. Actually, most label propagation methods based on local confidence exchange have the same complexity. The consumed time in real calculation mainly hinges on the value of k. In Figure 5.2 we plot the distribution of k obtained via the proposed ℓ_1-regularized weight computation, which reaches its peek value around $k = 35$. This small k value indicates that ℓ_1 penalty term is able to select much compacter reconstruction basis for a vertex. In contrast, to obtain nearly optimal performance, previous works usually take $k > 100$ (see Figure 5.3). In implementation, we find that the subtle reduce of k results in a drastic reduce of the running time (see more details in the experimental section).

5.4.2 Algorithmic Convergence

The above two updating procedures are iterated until converged. For the experiments on NUS-WIDE dataset, generally about 50 iterations are required for

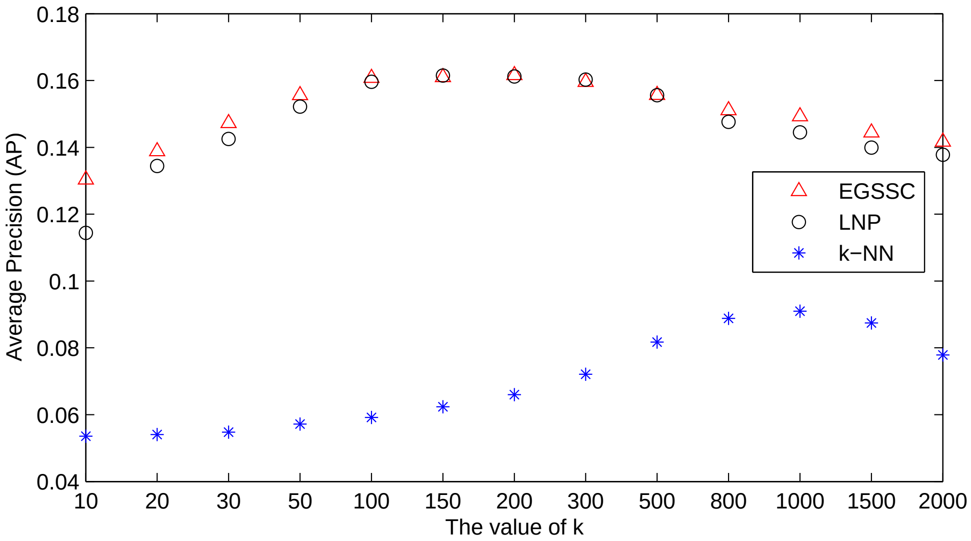


Figure 5.3: The performance of three baseline algorithms with respect to the number of nearest neighbors (denote as k).

the convergency of the solution. An exemplar convergency curve is shown in Figure 5.4.

5.5 Experiments

To validate the effectiveness of our proposed approach on large-scale multi-label datasets, we conduct extensive experiments on the real-world image dataset NUS-WIDE [Chua et al., 2009], which contains 269,648 images accompanied with totally 5,018 unique tags. Images in this dataset are crawled from the photo sharing website Flickr by using its public API. The underlying image diversity and complexity make it a good testbed for large-scale image annotation experiments. Moreover, a subset of NUS-WIDE (known as NUS-WIDE-Lite) obtained after noisy tag removal is also publicly available. We provide quantitative study on both the lite dataset and the full NUS-WIDE dataset, with an emphasis on the

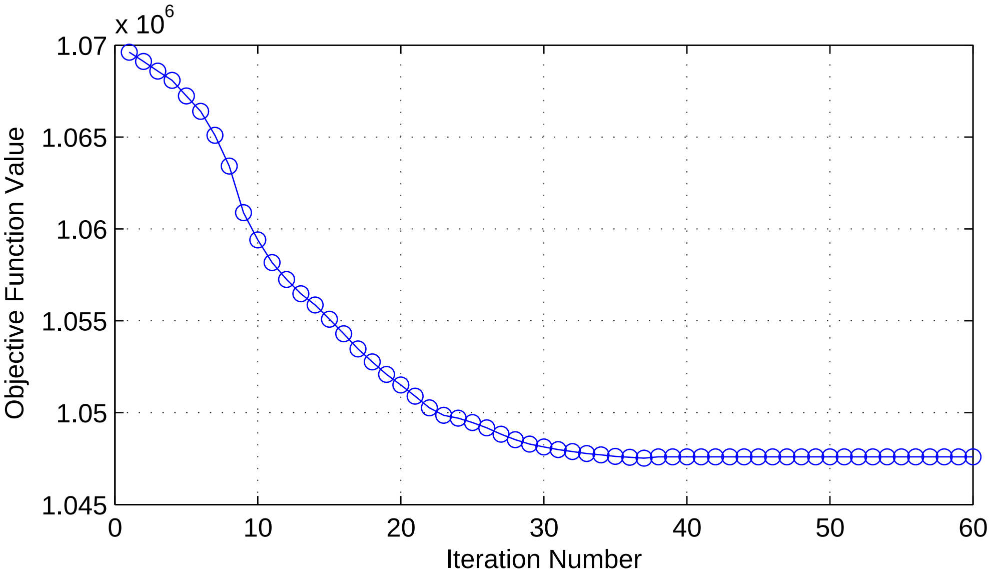


Figure 5.4: Convergence curve of our proposed Algorithm on NUS-WIDE dataset.

comparison with five state-of-the-art related algorithms in terms of accuracy and computational cost.

5.5.1 Datasets

NUS-WIDE [Chua et al., 2009]: The dataset contains 269,648 images and the associated 5,018 tags. For evaluation, we construct two image pools from the whole dataset: the pool of labeled images is comprised of 161,789 images whilst the rest are used for the pool of unlabeled images. For each image, an 81-D label vector is maintained to indicate its relationship to 81 distinct concepts (tightly related to tags yet relatively high-level). Moreover, to testify the performance stability of various algorithms, we vary the percentage of labeled images selected from the labeled image pool (in implementation it is varying from 10% to 100% increased by a step of 10%. We introduce the variable $\tau \in [0,\ 1]$ for it). The sampled labeled images are then amalgamated with the whole set of unlabeled images (107,859 in all). We extract multiple types of local visual features from

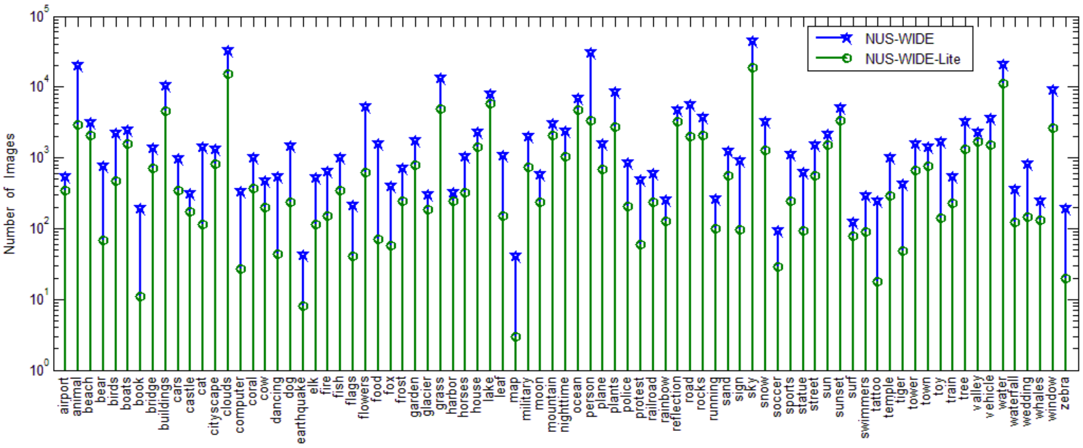


Figure 5.5: The distribution of 81 concepts in the training data of NUS-WIDE and NUS-WIDE-Lite when $\tau = 100\%$.

the images (225-D block-wise color moments, 128-D wavelet texture and 75-D edge direction histogram).

NUS-WIDE-Lite: As stated above, this dataset is a lite version of the whole NUS-WIDE database. It consists of 55,615 images randomly selected from the NUS-WIDE dataset. And the labels of each image are also like those of NUS-WIDE, an 81-D label vector is set to indicate its relationship to 81 distinct concepts. As done on NUS-WIDE, three types of local visual features are also extracted for this dataset. We randomly select about half of the images as labeled and the rest to be unlabeled. Again, we use the same sampling strategy on the labeled set to perform the stability test. Figure 5.5 illustrates the distribution of 81 concepts in the training data of NUS-WIDE and NUS-WIDE-Lite when $\tau = 100\%$.

5.5.2 Baselines and Evaluation Criteria

In the experiments, five baseline algorithms as shown in Table 5.1 are evaluated for comparative study. Amongst them, the *support vector machines* (SVM) is originally developed to solve binary-class or multi-class classification problem. Here we use its multi-class version by adopting the one-vs-one method. The selected baselines includes several state-of-the-art algorithms for semi-supervised learning. The *linear neighborhood propagation* (LNP) [Wang and Zhang, 2006] bases on a linear-construction criterion to calculate the edge weights of the graph, and disseminates the supervision information by a local propagation and updating process. The EGSSC [Subramanya and Bilmes, 2009] is an entopic graph-regularized semi-supervised classification method, which is based on minimizing a Kullback-Leibler divergence on the graph built from k-NN Gaussian similarity as introduced in sub-section 5.3.2.1 and 5.3.2.2. The SGSSL [Tang et al., 2009] is a sparse graph-based method for semi-supervised learning by harnessing the labeled and unlabeled data simultaneously, which considers each label independently.

The criteria to compare the performance include *Average Precision* (AP) for each label (or concept) and *Mean Average Precision* (MAP) for all labels. The former is a well-known gauge widely used in the field of image retrieval, whilst the latter is developed to handle the multi-class or multi-label cases. For example, in our application MAP is obtained by averaging the APs on 81 concepts. All experiments are conducted on a common desktop PC equipped with Intel dual-core CPU (frequency: 3.0 GHz) and 32G bytes physical memory.

For the experiments on NUS-WIDE-Lite, the proposed method is compared with all the five baseline algorithms. While on the NUS-WIDE, the results from SGSSL is not reported due to its incapability to handle dataset in such large scale.

Table 5.1: The Baseline Algorithms.

Name	Methods
KNN	k-Nearest Neighbors [Duda, Stork, and Hart, 2000]
SVM	Support Vector Machine [Collobert et al., 2006]
LNP	Linear Neighborhood Propagation [Wang and Zhang, 2006]
EGSSC	Entropic Graph Classification [Subramanya and Bilmes, 2009]
SGSSL	Sparse Graph-based Semi-supervised Learning [Tang et al., 2009]

5.5.3 Experiment-I: NUS-WIDE-LITE (56k)

In this experiment, we compare the proposed algorithm with five baseline algorithms. The results with varying numbers of labeled images (controlled by the parameter τ) are presented in Figure 5.6. Below are the parameters and the adopted values for each method: for KNN, there is only one parameter k for tuning, which stands for the number of nearest neighbors and is trivially set as 500. For SVM algorithm, we adopt the RBF kernel. For its two parameters γ and C, we set $\gamma = 0.6$ and $C = 1$ in experiments after fine tuning. For LNP algorithm, one parameter α is adjusted, which is the fraction of label information that each image receives from its neighbors. The optimal value is $\alpha = 0.95$ in our experiments. There are three parameters μ, ν and β in EGSSC, where μ and ν are used for weighting the Kullback-Leibler divergence term and Shannon entropy term respectively and β ensures the convergence of the two similar probability measures. The optimal values are set as $\mu = 0.1$, $\nu = 1$ and $\beta = 2$ here. For our proposed algorithm, we set $\mu = 10$ and $\eta = 5$. MAP of these six methods is illustrated in Figure 5.7.

Our observations from Figure 5.6 are described as follows:

- Our proposed algorithm LSMP outperforms the other baseline algorithms significantly when selecting different proportions of labeled set. For example,

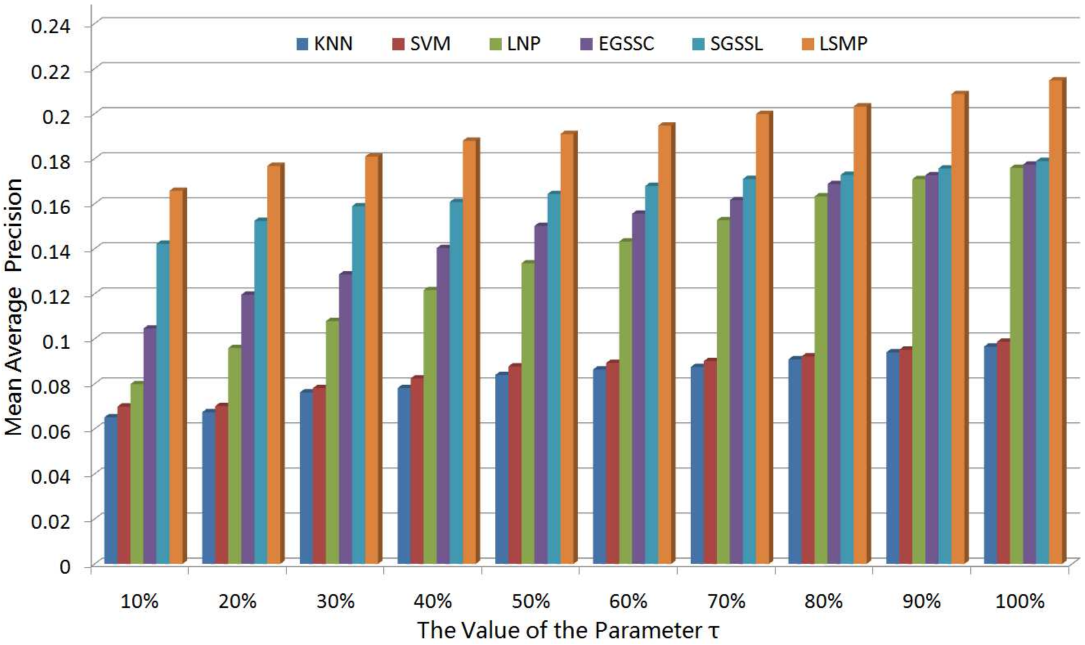


Figure 5.6: The results of the comparison of LSMP and the five baselines with varying parameter τ on NUS-WIDE-Lite dataset.

with 10 percent of labeled images selected, LSMP has an improvement 16.6% over SGSSL, 58.5% over EGSSC, 107.6% over LNP, 137.2% over SVM, and 154.5% over KNN. The improvement is supposed to stem from the fact that our proposed algorithm encodes the label information of each image as a unit confidence vector, which imposes extra inter-label constraints. In contrast, other methods either consider the visual similarity graph only, or considers each label independently.

- With the increasing number of labeled images, the performances of all algorithms consistently increase. When $\tau \leq 0.6$, the algorithm SGSSL outperforms the other two state-of-art algorithms LNP and EGSSC significantly. However, when $\tau > 0.6$, the improvement of SGSSL over the others is lower. The proposed method keeps higher MAP value than other five methods over

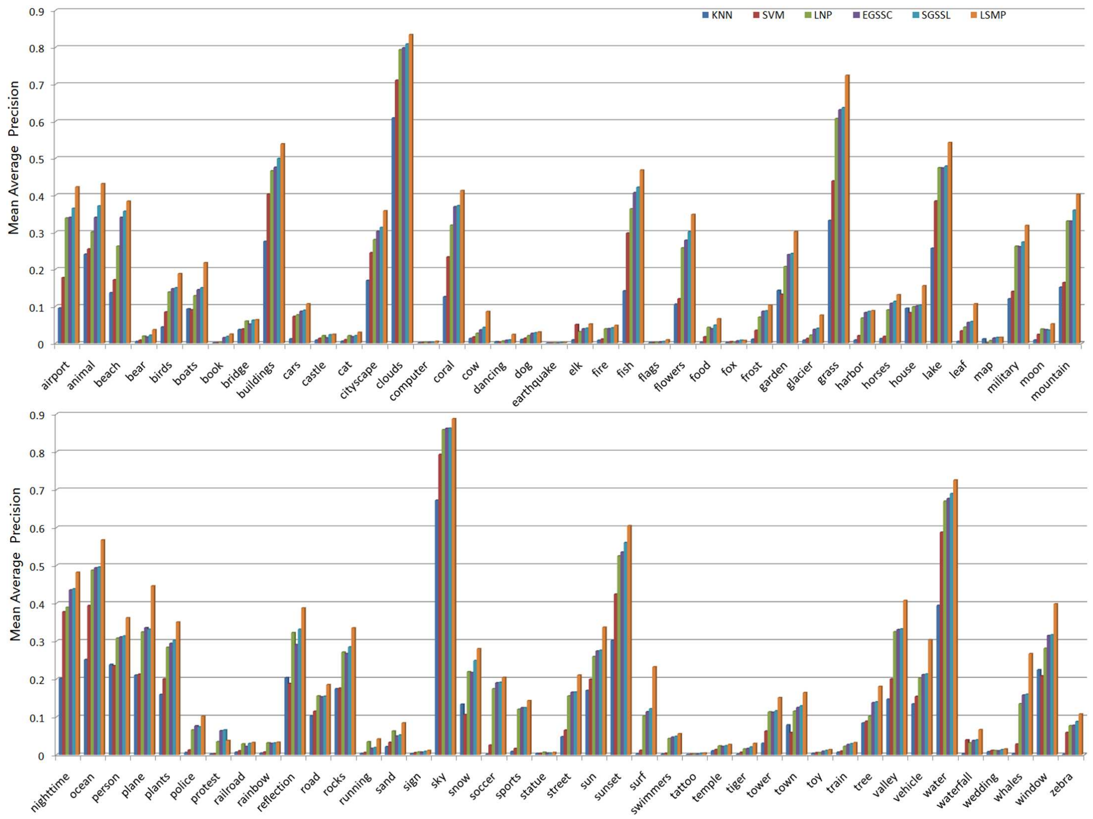


Figure 5.7: The comparison of APs for the 81 concepts using six methods with $\tau = 1$.

all values of τ.

Recall that the proposed algorithm is a probabilistic collaborative multi-label propagation algorithm, wherein $p_i(y)$ expresses the probability for the i-th image to be associated with the y-th label. A direct application for this probabilistic implication is the tag ranking task. Some exemplar results of tag ranking are shown in Figure 5.8.

5.5.4 Experiment-II: NUS-WIDE (270k)

In this experiment, we compare the proposed LSMP algorithm with four state-of-the-art algorithms on the large-scale NUS-WIDE dataset for multi-label image

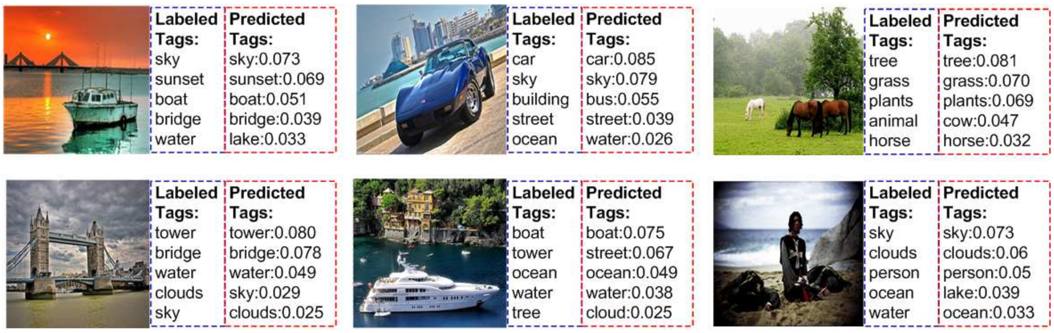

Labeled Tags: sky sunset boat bridge water
Predicted Tags: sky:0.073 sunset:0.069 boat:0.051 bridge:0.039 lake:0.033
Labeled Tags: car sky building street ocean
Predicted Tags: car:0.085 sky:0.079 bus:0.055 street:0.039 water:0.026
Labeled Tags: tree grass plants animal horse
Predicted Tags: tree:0.081 grass:0.070 plants:0.069 cow:0.047 horse:0.032
Labeled Tags: tower bridge water clouds sky
Predicted Tags: tower:0.080 bridge:0.078 water:0.049 sky:0.029 clouds:0.025
Labeled Tags: boat tower ocean water tree
Predicted Tags: boat:0.075 street:0.067 ocean:0.049 water:0.038 cloud:0.025
Labeled Tags: sky clouds person ocean water
Predicted Tags: sky:0.073 clouds:0.06 person:0.05 lake:0.039 ocean:0.033

Figure 5.8: The tags ranking results of LSMP in NUS-WIDE-LITE.

annotation. As in previous experiments, we modulate the parameter τ to vary the percentage of the labeled images used in the experiments and carefully tune the optimal parameters in each method for fair comparison. For KNN, the optimal value is $k = 1000$. For SVM algorithm, we set $\lambda = 0.8$ and $C = 2$. For LNP method, the optimal value is $\alpha = 0.98$. In the experiment of EGSSC, the best values are $\mu = 0.5$, $\nu = 1$ and $\beta = 1$. For our proposed LSMP algorithm, $\mu = 15$ and $\eta = 8$. The results of all algorithms are shown in Figure 5.9 and the results with respect to each individual concept are presented in Figure 5.10. From Figure 5.10, we can observe that

- On the large-scale real-world image dataset, the proposed algorithm outperforms other algorithms significantly at all values of τ. For example, when

Table 5.2: Executing time (unit: hours) comparison of different algorithmson the NUS-WIDE dataset.

Algorithms	Graph Construction Time	Label Estimation Time	Total Time
KNN	143.6	0.7	144.3
SVM	0	132.5	132.5
LNP	143.6	0.2	143.8
EGSSC	143.6	2.4	146
LSMP	31.4	0.3	31.7

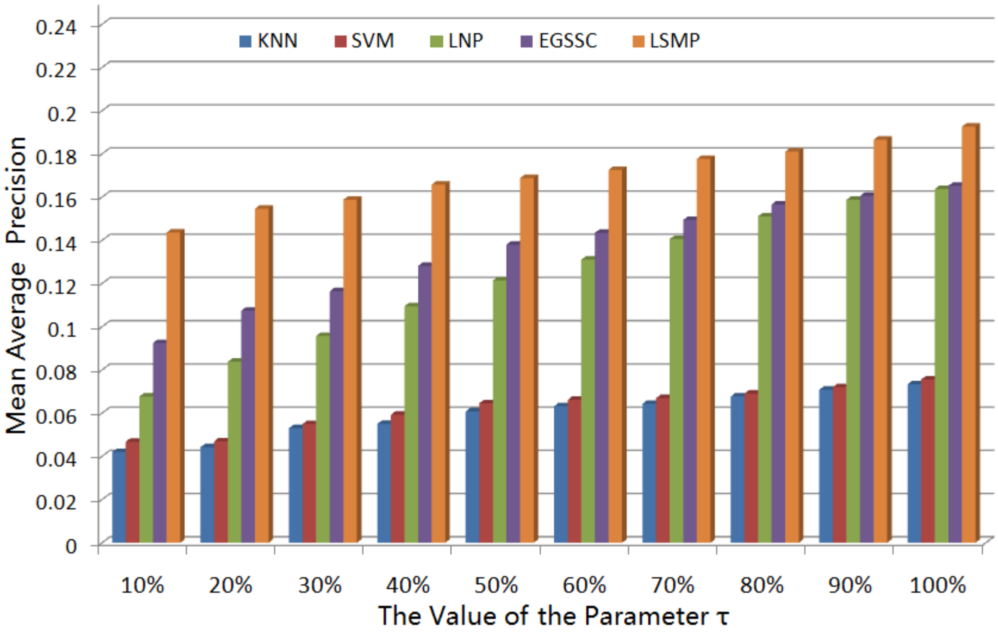


Figure 5.9: The results of the comparison of LSMP and the four baselines with varying parameter τ on NUS-WIDE.

$\tau = 0.1$, LSMP has an improvement 53.5% over EGSSC, 112.6% over LNP, 197.2% over SVM, and 220.5% over KNN. Compared with the performance on NUS-WIDE-Lite, the best performance of LSMP in NUS-WIDE is 0.193, which is smaller than the MAP value in the Lite version. The performance degradation is primarily attributed to the increase of data scale (the size of labeled image pool in NUS-WIDE is 170K, while for the Lite version it is only 27K).

- With the increasing parameter τ, the performances of all algorithms also increase. When $\tau \leq 0.6$, the algorithm EGSSC outperforms LNP significantly, but for $\tau > 0.6$, the improvement of EGSSC than LNP is negligible. The proposed method LSMP also keeps higher MAP value than all baselines over all feasible values of τ similar to the case on NUS-WIDE-LITE, which

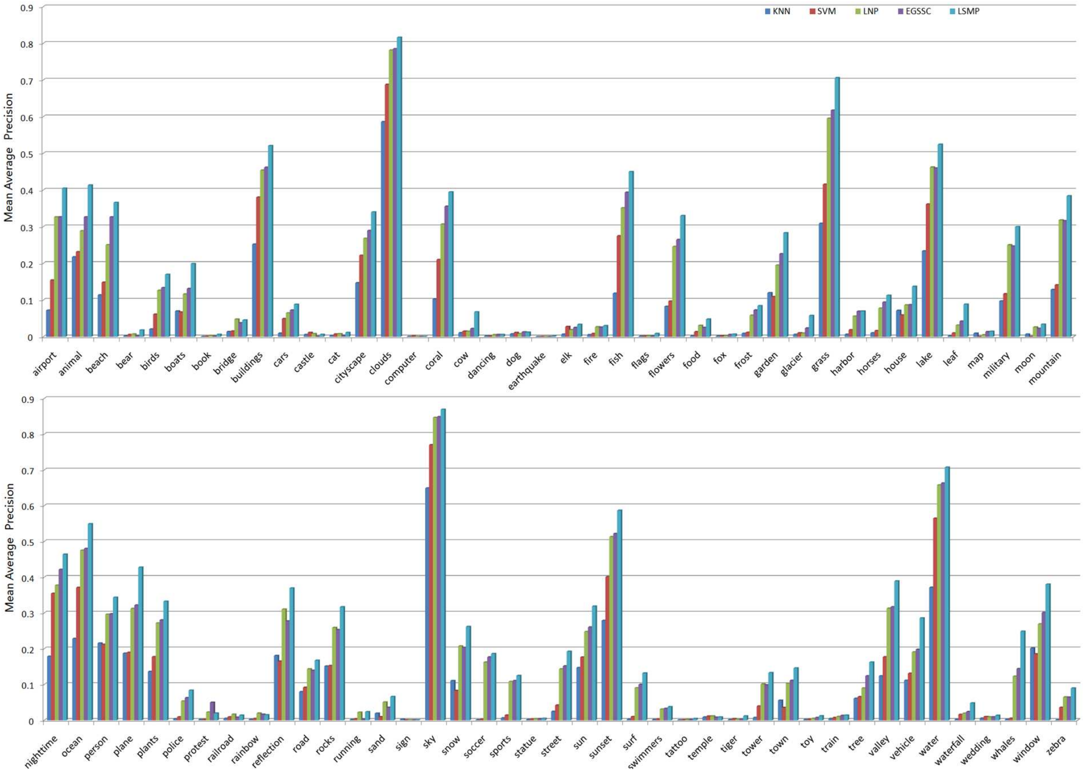


Figure 5.10: The comparison of APs for the 81 concepts with $\tau = 1.0$ on NUS-WIDE.

validates the robustness of our proposed algorithm.

We also provide the recorded running time for different algorithms on NUS-WIDE, as shown in Table 5.2. A salient efficacy improvement can be observed from our proposed method.

5.6 Conclusion

In this chapter we proposed and validated an efficient large-scale image annotation method. Our contributions lie in both the hashing-accelerated ℓ_1-graph construction, and KL-divergence oriented soft loss function and regularization term in graph-based modeling. The optimization framework utilizes the inter-label relationship and finally returns a probabilistic label vector for each image, which is more robust to noises and can be used for tag ranking. The proposed algorithm is tested on several publicly-available image benchmarks built for multi-label annotation, including the publicly available largest NUS-WIDE data set. We showed the superiority of our proposed method in terms of both accuracy and efficacy. Our future work will follow two directions: 1) extend the image annotation datasets to web-scale and further validate the scalability of our proposed method; and 2) develop more elegant algorithms for KL-based label propagation which shows better convergent speed.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this dissertation, we first addressed the multi-label learning problems for semantic image annotation using two paradigms: multi-label learning on traditional single semantic space and multi-label learning on multiple semantic spaces. We then presented a novel and efficient sparse graph based multi-label learning scheme for large-scale image annotation. We summarize our research as follows:

1) We presented a label exclusive context regularized multi-label linear representation framework for semantic image annotation, which is formulated as an eLasso model with group overlaps and affine transformation.

2) We proposed a multi-semantic multi-label learning framework for semantic image annotation, in which the multi-task linear discriminative model is 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) We introduced an efficient KL-divergence based multi-label learning framework for large-scale image annotation, which is based on hashing-accelerated ℓ_1-graph construction.

The validity and the performances of these proposed approaches were demonstrated by extensive experiments on the challenging real-world benchmarks: PASCAL VOC 2007&2010, NUS-WIDE-Emotive dataset, and NUS-WIDE dataset. In this chapter, we summarize this dissertation with a review of our main research contributions, and discuss new directions for future research.

6.1.1 Multi-Label Learning with Label Exclusive Context

In this dissertation, we proposed a Label Exclusive Linear Representation (LELR) model to incorporate label exclusive context into a multi-label linear representation framework for multi-label learning. The proposed label exclusive context described the negative relationship among class labels. Given a set of exclusive label groups, the proposed LELR model enforces repulsive assignment of the labels from each group to a query image. For the solution of LELR, we formulated it as an eLasso model with group overlaps and affine transformation. Such a variant of eLasso was efficiently optimized with Nesterov-type smoothing approximation method. Extensive experiments on the challenging real-world visual classification tasks validate that LELR is a powerful model to boost the performance of linear representation and classification.

6.1.2 Multi-Label Learning on Multi-Semantic Space

To handle and explore the annotation problem of images contained comprehensive semantics, we developed a novel and promising approach called Image Annotation with Multi-Semantic Labeling (IA-MSL), to annotate multi-semantic meaning of images based on two or more semi-orthogonal label spaces from multi-semantic. We formulated this challenging problem as a multi-task discriminative analysis model, where individual tasks are defined by learning the linear discriminative model for individual complex semantic concepts. We considered all the tasks in a joint manner by imposing two types of regularization on parameters: 1) the graph Laplacian regularization to deal with the problem of insufficient training samples; and 2) the exclusive group lasso regularization for competitive feature selection. For model optimization, we introduced a Nesterov-type smoothing approximation method. The proposed algorithm was tested on two image benchmarks built for multi-semantic annotation: NUS-WIDE-Emotive dataset, and NUS-WIDE-Object&Scene. We validated the superiority of IA-MSL in terms of both accuracy and efficacy.

6.1.3 Multi-Label Learning in Large-Scale Dataset

We further developed and validated an efficient sparse graph multi-label learning method for large-scale image annotation, whereby both the efficacy and accuracy of annotation were enhanced. Different from previous large-scale approaches that propagate over individual label independently, we encoded the tag information of each image to the proposed large-scale multi-label propagation (LSMP) scheme, in which the Kullback-Leibler divergence was employed for problem formulation. We then performed the multi-label propagation on the hashing-accelerated ℓ_1-graph, which was efficiently derived with Locality Sensitive Hashing approach followed by sparse ℓ_1-graph construction within the individual hashing buckets. An efficient and convergence provable iterative procedure was also presented for problem optimization. Finally, the whole optimization framework returned a probabilistic label vector for each image, which was more robust to noise and could be used for tag ranking. Extensive experiments on several publicly-available image benchmarks well validated the effectiveness and scalability of the proposed approach.

6.2 Future Work

Despite the significant progress made in this thesis, there remain several open exciting challenges for multi-label learning of semantic image annotation. In the followings, we discuss some interesting topics that we will explore in our future research agenda.

1) Multi-Label Learning with Label Exclusive Context

The implementation and optimization of the proposed Label Exclusive Linear Representation (LELR) model should be improved for multi-label learning with large number of categories (e.g. ImageNET [Deng et al., 2009] which contains 5247 categories.). Since LELR is a variant of eLasso, one may wish to utilize the existing eLasso solvers for optimization. However, we observe that the eLasso solvers in literature either suffer from slow convergence rate (e.g., subgradient methods in [Zhou, Jin, and Hoi, 2010]) or are particularly designed for standard eLasso with disjoined groups (e.g., proximal gradient method in [Kowalski and Torreesani, 2009]), and thus are not directly applicable to LELR. In this thesis, we first approximate the non-smooth objective in by a smooth function and then solve the latter by utilizing the off-the-shelf Nesterov's smoothing optimization method. However, from the experimental results of LELR model, we found that the executing time of LELR increases with the size of concept set in image dataset. For example, the per query time of LELR in PASCAL VOC 2007&2010 containing 20 concepts is about 0.2 second, and the per query time in NUS-WIDE-LITE including 81 concepts is about 0.75 second. This motivates us to seek more efficient approach to optimizie the objective function of LELR in order to handle large number of concepts in real-world problem.

2) Multi-Label Learning on Multi-Semantic Space

The proposed Image Annotation with Multi-Semantic Labeling (IA-MSL) method should be extended towards real world search scenario. Due to the popularity of photo sharing websites, the contents of images are enriched and more diverse than ever before. How to effectively annotate these images on a wide variety of semantics and topics for improved image search performance is a challenging problem. In this thesis, the proposed IA-MSL method has been designed to annotate images simultaneously with labels in two or more semantic spaces. But with the increasing of the number of semantic space in image corpus, a large number of classes will be involved in training due to the combination of multiple semantic spaces. As a result, many classes will suffer from the problem of insufficient training samples. The worst case is that some classes do not have training samples. This motivates us to further explore the IA-MSL algorithm and expand the search range towards real world search scenario.

3) Multi-Label Learning in Large-Scale Dataset

More elegant algorithms for the proposed KL-based large-scale multi-label propagation (LSMP) scheme should be developed in order to get better convergent speed. As proven in this thesis, the objective function of LSMP is convex, and hence LSMP has a global optima for the solution. But there is no closed form solution for the objective function, which may affect the convergent performance. Since no closed-form solution is feasible, standard numerical optimization approaches such as interior point methods (IPM) or method of multipliers (MOM) can be used to solve the problem. However, most of these approaches guarantee global optima yet are tricky to implement (e.g., an implementation of MOM to solve this problem would have seven extraneous parameters) [Subramanya and Bilmes, 2009]. Although we adopt a simple alternating minimization method to tackle the objective function and the implementation of LSMP is efficient, the convergent performance may be improved if a more suitable algorithms is chosen and exploited to solve the objective function of LSMP.

References

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

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

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

Boyd, Stephen and Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1993. *Classification and Regression Trees*. Chapman and Hall.

Cai, L. and T. Hofmann. 2004. Hierarchical document categorization with support vector machines. In *ACM International Conference on Information and Knowledge Management*.

Candès, Emmanuel J., Justin K. Romberg, and Terence Tao. 2006. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, February.

Cao, L., J. Luo, and T. Huang. 2008. Annotating photo collections by label propagation according to multiple similarity cues. In *ACM International Conference on Multimedia*.

Caruana, R. 1997. Multi-task learning. *Machine Learning*, 28(1):41–75.

Chang, E., G. Sychay K. Goh, and G. Wu. CBSA. 2003. Contentbased soft annotation for multimodal image retrieval using bayes point machines. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(1):26–38.

Chapelle, O., P. Haffner, and V. N. Vapnik. 1999. Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10:1055–1064.

Chen, Gang, Yangqiu Song, Fei Wang, and Changshui Zhang. 2008. Semi-supervised multi-label learning by solving a sylvester equation. In *SIAM International Conference on Data Mining*.

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

Chen, X., Y. Mu, S. Yan, and T.-S. Chua. 2010. Efficient large-scale image annotation by probabilistic collaborative multi-label propagation. In *ACM International Conference on Multimedia*.

Choi, Myung Jin, Joseph J. Lim, Antonio Torralba, and Alan S. Willsky. 2010. Exploiting hierarchical context on a large database of object categories. In *IEEE International Conference on Computer Vision and Pattern Recognition*.

Chu, W. and Z. Ghahramani. 2005. Preference learning with gaussian processes. In *International Conference on Machine Learning*.

Chua, T.-S., J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. 2009. NUS-WIDE: A real-world web image database from national university of singapore. In *ACM International Conference on Image and Video Retrieval*.

Cilibrasi, R. and P. M. B. Vitanyi. 2007. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383.

Collobert, Ronan, Fabian H. Sinz, Jason Weston, and Léon Bottou. 2006. Large scale transductive svms. *Journal of Machine Learning Research*, 7:1687–1712, September.

Cortes, C. and V. Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Cover, T. M. and J. A. Thomas. 1991. *Elements of Information Theory*. Wiley Series in Telecommunications.

Delalleau, Olivier, Yoshua Bengio, and Nicolas Le Roux. 2005. Efficient non-parametric function induction in semi-supervised learning. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 96–103.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE International Conference on Computer Vision and Pattern Recognition*.

Desai, C., D. Ramanan, and C. Fowlkes. 2009. Discriminative models for multi-class object layout. In *IEEE International Conference on Computer Vision*.

Duda, R., D. Stork, and P. Hart. 2000. *Pattern Classification*. JOHN WILEY.

Elisseeff, A. and J. Weston. 2002. A kernel method for multi-labelled classification. *Advances in Neural Information Processing Systems, MIT Press*.

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

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

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

Evgeniou, Theodoros and Massimiliano Pontil. 2004. Regularized multi–task learning. In *ACM International Conference on Knowledge Discovery and Data mining*.

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

Frate, F. D., F. Pacifici, G. Schiavon, and C. Solimini. 2007. Use of neural networks for automatic classification from high-resolution images. *IEEE Transactions on Geoscience and Remote Sensing*, 45(4):800–809.

Freund, Y. and R. E. Schapire. 1997. Learning multilabel scene classification. *Pattern Recognition*, 55(1):119–139.

Furnkranz, J., E. Hullermeier, E. Loza Mencia, and K. Brinker. 2008. Multilabel classification via calibrated label ranking. machine learning. *Machine Learning*, 73(2):133–153.

Godbole, S. and S. Sarawagi. 2004. Discriminative methods for multi-labeled classification. In *Eighth Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 22–30.

Griffiths, T. and Z. Ghahramani. 2005. Infinite latent feature models and the indian buffet process. In *Neural Information Processing Systems*.

Hanjalic, A. 2006. Extracting moods from pictures and sounds: Towards truly personalized TV. *Signal Processing Magazine*, 23(2):90–100.

Hanley, J. A. and B. J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36.

Hayashi, T. and M. Hagiwara. 1998. Image query by impression words-the IQI system. *IEEE Transactions on Consumer Electronics*, 44(2):347–352.

Hullermeier, E., J. Furnkranz, W. Cheng, and K. Brinker. 2008. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172(16-17):1897–1916.

Indyk, P. and R. Motwani. 1998. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Symposium on Theory Computing*.

Jacob, Laurent, Guillaume Obozinski, and Jean-Philippe Vert. 2009. Group lasso with overlap and graph lasso. In *International Conference on Machine Learning*.

Ji, S., L. Tang, S. Yu, and J. Ye. 2008. Extracting shared subspace for multi-label classification. In *ACM International Conference on Knowledge Discovery and Data mining*, pages 381–389.

Kang, F., R. Jin, and R. Sukthankar. 2006. Correlated label propagation with application to multi-label learning. In *IEEE International Conference on Computer Vision and Pattern Recognition*.

Karlen, Michael, Jason Weston, Ayse Erkan, and Ronan Collobert. 2008. Large-scale manifold transduction. In *International Conference on Machine Learning*.

Kesorn, Kraisak. 2010. *Multi-Model Multi-Semantic Image Retrieval.* PhD Thesis, Queen Mary, University of London.

Kowalski, M. and B. Torreesani. 2009. Sparsity and persistence: Mixed norms provide simple signals models with dependent coefficient. *Signal, Image and Video Processing*, 3(3):251–264.

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

Lazebnik, S., C. Schmid, and J Ponce. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE nternational Conference on Computer Vision and Pattern Recognition.*

Lew, M., N. Sebe, C. Djeraba, and R. Jain. 2006. Content-based multimedia information retrieval: State-of-the-art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2(1):1–19.

Liu, Dong, Xian-Sheng Hua, Linjun Yang, Meng Wang, and Hong jiang Zhang. 2009. Tag ranking. In *International World Wide Web Conference.*

Liu, Han, Mark Palatucci, and Jian Zhang. 2009. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *International Conference on Machine Learning*, pages 649–656.

Liu, H.R. and S.C. Yan. 2010. Robust graph mode seeking by graph shift. In *International Conference on Machine Learning.*

Liu, Y., D. Zhang, and G. Lu. 2008. Region-based image retrieval with high-level semantics using decision tree learning. *Pattern Recognition*, 41(8):2554–2570.

Liu, Yi, Rong Jin, and Liu Yang. 2006. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In *Proceedings of National Conference on Artificial Intelligence.*

Machajdik, Jana and Allan Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. In *ACM International Conference on Multimedia.*

McCallum, A. 1999. Multi-label text classification with a mixture model trained by em. In *Working Notes of the AAAI'99 Workshop on Text Learning.*

Mencia, E. L. and J. Furnkranz. 2008a. Pairwise learning of multilabel classifications with perceptrons. In *International Joint Conference on Neural Networks*, pages 2899–2906.

Mencia, E. Loza and J. Furnkranz. 2008b. Efficient pairwise multilabel classification for large-scale problems in the legal domain. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 50–65.

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

Mojsilovic, A. and B. Rogowitz. 2001. Capturing image semantics with low-level descriptors. In *IEEE International Conference on Image Processing*, pages 18–21.

Mu, Yadong, Jialie Shen, and Shuicheng Yan. 2010. Weakly-supervised hashing in kernel space. In *IEEE International Conference on Computer Vision and Pattern Recognition.*

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

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

Nocedal, Jorge and Stephen J. Wright. 2006. *Numerical Optimization*. Springer-Verlag.

Obozinski, G., B. Taskar, and M.I. Jordan. 2009. Joint covariate selection and joint subspace selection for multiple classification problems. *Journal of Statistics and Computing*, 20(2):231–252.

Qi, Guo-Jun, Xian-Sheng Hua, Yong Rui, Jinhui Tang, Tao Mei, and Hong-Jiang Zhang. 2007. Correlative multi-label video annotation. In *ACM International Conference on Multimedia*.

Quinlan, J. R. 1986a. *Induction of decision trees*. Springer Machine Leaning.

Quinlan, J. R. 1986b. Induction of decision trees. *Machine Learning*, pages 81–106.

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning*. California, USA.

Raez, A. M., L. A. U. Lopez, and R. Steinberger. 2004. Adaptive selection of base classifiers in one-against-all learning for large multi-labeled collections. In *4th International Conference on Advances in Natural Language Processing*, pages 1–12.

Rousu, J., C. Saunders, S. Szedmak, and J. Shawe-Taylor. 2004. On maximum margin hierarchical multi-label classification. In *NIPS Workshop on Learning With Structured Outputs*.

Roweis, S.T. and L.K. Saul. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326.

Sande, K., T. Gevers, and C. Snoek. 2010. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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

Sethi, I. K. and I. L. Coman. 2001. Mining association rules between low-level image features and high-level concepts. *SPIE Data Mining and Knowledge Discovery*, 3:279–290.

Shi, R., H. Feng, T. S. Chua, and C. H. Lee. 2004a. Anadaptive image content representation and segmentation approach to automatici mage annotation. In *Proceedings of the International Conference on Image and Video Retrieval*, pages 545–554.

Shi, R., H. Feng, T. S. Chua, and C. H. Lee. 2004b. Image classification into object/non-object classes. In *Proceedings of the International Conference on Image and Video Retrieval*, pages 393–400.

Shotton, J., J. Winn, C. Rother, and A. Criminisi. 2006. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *European Conference on Computer Vision*, pages 1–15.

Sindhwani, V. and S. S. Keerthi. 2006. Large scale semi-supervised linear svms. In *ACM SIGIR Conference on Research and Development in Information Retrieval*.

Subramanya, Amarnag and Jeff Bilmes. 2009. Entropic graph regularization in non-parametric semi-supervised classification. In *Neural Information Processing Systems.*

Tang, Jinhui, Shuicheng Yan, Richang Hong, Guo-Jun Qi, and Tat-Seng Chua. 2009. Inferring semantic concepts from community-contributed images and noisy tags. In *ACM International Conference on Multimedia.*

Tsang, Ivor W. and James T. Kwok. 2006. Large-scale sparsified manifold regularization. In *Neural Information Processing Systems.*

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

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

Ueda, N. and K. Saito. 2002. Parametric mixture models for multi-labeled text. In *Neural Information Processing Systems.*

Vapnik, V. 1995. *The Nature of Statistical Learning Theory.* Springer.

Wang, F. and C. Zhang. 2006. Label propagation through linear neighborhoods. In *International Conference on Machine Learning.*

Wang, Jingjun, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. 2010. Locality-constrained linear coding for image classification. In *IEEE International Conference on Computer Vision and Pattern Recognition.*

Wang, Wei-Ning, Ying-Lin Yu, and Sheng-Ming Jiang. 2006. Image retrieval by emotional semantics: A study of emotional space and feature extraction. In *IEEE International Conference on Systems, Man and Cybernetics.*

Weiss, G. M. and F. J. Provost. 2003. Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, 19:315–354.

Wong, R. C. F. and C. H. C. Leung. 2008. Automatic semantic annotation of real-world web images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1933–1944.

Wright, J., A.Y. Yang, A. Ganesh, S.S Sastry, and Yi Ma. 2009. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–226.

Wu, Lei, Xian-Sheng Hua, Nenghai Yu, Wei-Ying Ma, and Shipeng Li. 2008. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In *ACM International Conference on Multimedia*.

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

Yan, R., J. Tesic, and J. R. Smith. 2007. Model-shared subspace boosting for multi-label classification. In *ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 834–843.

Yan, S.C. and H. Wang. 2009. Semi-supervised learning by sparse representation. In *SIAM International Conference on Data Mining*.

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

Yu, K., S. Yu, , and V. Tresp. 2005. Multi-label informed latent semantic indexing. In *ACM SIGIR Conference on Research and Development in Information Retrieval*.

Yuan, J., J. Li, and B. Zhang. 2007. Exploiting spatial context constraints for automatic image region annotation. In *ACM International Conference on Multimedia*.

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

Yuan, X. and S.C. Yan. 2010. Visual classification with multi-task joint sparse representation. In *IEEE International Conference on Computer Vision and Pattern Recognition*.

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

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

Zhou, D., B. Scholkopf, and T. Hofmann. 2005. Semi-supervised learning on directed graphs. In *Neural Information Processing Systems*.

Zhou, Xi, Kai Yu, Tong Zhang, and Thomas Huang. 2010. Image classification using super-vector coding of local image descriptors. In *European Conference on Computer Vision*.

Zhou, Y., R. Jin, and Steven C.H. Hoi. 2010. Exclusive lasso for multi-task feature selection. In *International Conference on Artificial Intelligence and Statistics*.

Zhu, Guangyu, Shuicheng Yan, and Yi Ma. 2010. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *ACM International Conference on Multimedia*.

Zhu, S., X. Ji, W. Xu, and Y. Gong. 2005. Multi-labelled classification using maximum entropy method. In *ACM SIGIR Conference on Research and Development in Information Retrieval*.

Zhu, X., Z. Ghahramani, and J. Lafferty. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *International Conference on Machine Learning*.

Zhu, Xiaojin. 2005. *Semi-supervised learning with graphs*. Carnegie Mellon University.

Zhu, Xiaojin. 2006. *Semi-Supervised Learning Literature Survey*. Carnegie Mellon University.