Towards Effective Image Annotation by Exploiting Multimodal Data

徐, 行

https://doi.org/10.15017/1543997

Towards Effective Image Annotation by Exploiting Multimodal Data

Xing Xu

Department of Advanced Information Technology
Graduate School of Information Science and Electrical Engineering
Kyushu University

This dissertation is submitted for the degree of

Doctor of Philosophy

July 2015

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Xing Xu
July 2015

Table of contents

List of figures

List of tables

Chapter 1

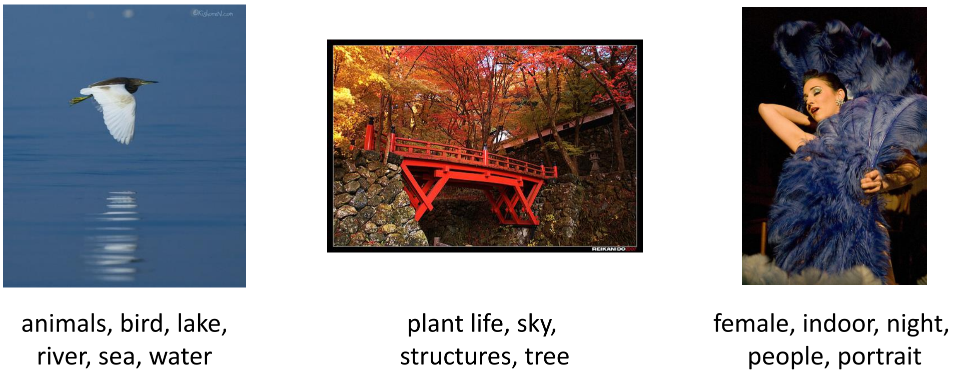
Introduction

1.1 Background

In recent years we have witnessed an explosion in the number of multimedia data in personal archives, social media and content-sharing web sites, such as Flickr, Facebook and Instagram. For example, Flickr has more than two billion images with millions of newly uploaded photos per day. Indeed, given the speed at which new data is released, the traditional manner of manually indexing and searching in such large archives would be prohibitive with extremely high cost. Therefore, how to efficiently manage and access to large archives in a user-oriented and semantically meaningful way is a challenging problem.

In order to automatically index the archive images with the goal of providing easy and efficient access to users, it is necessary to automatically generate text descriptions for the images to describe the corresponding visual contents [1]. This supposes to build systems that can bridge the semantic gap between low-level visual features and high-level semantics. To illustrate this fact, let us consider an important computer vision problem, namely automatic image annotation. Given an unseen image, the goal of automatic image annotation is to predict multiple textual labels describing that image. The candidates of labels are extracted from the texts surrounding the images in the web pages, and the meanings of the labels can range from specific object categories or general patterns, as illustrated in Fig. 1.1, and the images are usually obtained from photo sharing web sites, e.g. Flickr, Picasa and etc.

Note that the image annotation problem is different from the image classification problem [2–4], which aims to assign a single category label exclusively to one image. In contrast, image annotation is the task of assigning multiple labels to a single image. Thus, image annotation is a more generic but difficult problem than classification problem, since we need to consider relationships between labels during annotation.



animals, bird, lake, river, sea, water

plant life, sky, structures, tree

female, indoor, night, people, portrait

Fig. 1.1 Examples of images with manually annotated labels.

During the last decade, there are large efforts to develop various models, such as generative models [5], topic models [6–8], discriminative models [9–11] and nearest neighbor (NN) based models [12–15] (more technical details of these models are described in Chapter 3), have been proposed to tackle the image annotation problem. Among these models, the topic model can be considered as the pioneering work for modeling co-occurrences of visual pixels and textual words. Since the topic model is initially derived from the natural language processing community to cluster and to classify textual documents, it is a natural way to model the words associated to the images.

However, previous topic models based on Latent Dirichlet Allocation (LDA) [6] are all under Dirichlet assumption, where the topic proportions of an image are generated independently, and the presence of one topic is not correlated with the presence of others. Indeed, the annotation words of an image usually have correlation, and the correlation ought to be incorporated in the topic correlations. Thus, how to efficiently capture the topic correlations in the LDA based topic models is an important issue to be investigated.

Furthermore, in the problem setting of image annotation, the ground truth labels (tags) assigned to the images are assumed to be compact and perfect. However, the ground truth labels of the images are assigned by human annotators manually without rigorous supervision. Due to unconsciousness and misconception of the annotators, the tag assignments of the training images tends to be ambiguous, incomplete and even imprecise. Therefore, the dataset can be considered as weakly labeled. Given the weakly labeled dataset for training, it would misguide the learning procedure and result in an unstable annotation model. Therefore, how to overcome the defection in the training images and how to learn a stable annotation model is a primary issue to be tackled in practical scenario.

In addition, if we consider image annotation as searching for a set of tags corresponding to a given image, a promising application of image annotation could be searching for a set of images using a query string (e.g. as is done with Google Image search), which can be considered as a reverse process of image annotation. When we consider image data and text data (tags or documents) come from two different modalities, the so-called cross-modal retrieval is a more general problem that covers both of the image annotation problem and its reverse process. Unlike the image annotation problem that usually focuses on the one-way from unimodal images to unimodal texts, for the cross-modal retrieval problem, the mutual associations between the different modalities need to be stressed. Due to the distinct statistical properties and semantic gap between the different modalities of images and texts, the most difficult problem of cross-modal retrieval is how to eliminate the heterogeneity between modalities to facilitate effective cross-modal matching.

1.2 Objective of this research

In this thesis, firstly the author will focus on the traditional image annotation problem, and particular interests are focused on developing effective topic models to tackle this problem. Secondly, the author analyzes the statistical properties of user-provided ground truth labels (also called "tags") of training images in the image annotation problem, and investigate how the quality of the ground truth labels affects the annotation results. Thirdly, the author extends the image annotation problem to a more general issue of cross-modal retrieval considering mutual associations of images and texts. The major problems that are explored in the thesis can be summarized as follows.

- How to efficiently capture the topic correlations in the LDA based topic models to improve the annotation performance.
- How to learn a robust annotation model given training images in weakly labeled dataset.
- How to eliminate the heterogeneity between different modalities to facilitate effective cross-modal retrieval.

With respect to the first problem, to directly model correlation between topics when generating the topic proportions for a given image, the author uses logistic normal distribution proposed in the correlated topic model (CTM) [16] instead of the Dirichlet assumption to capture the topic correlations. Then, the author extends the idea of modeling topic correlations from popular LDA based models to CTM based models. In the proposed CTM based models, topics are now correlated with each other to mix the topic proportions for one image. The author presents a comprehensive comparison between CTM based and LDA based models on three benchmark datasets, illustrating the superior annotation performance of proposed CTM based models on image annotation task, by means of propagating topic correlation among image features and annotation words.

For the second problem, the author invesigates a practical issue of image annotation in weakly labeled datasets that are collected from Internet. To make this problem simple, the author first addresses a special case of image annotation with *incomplete labeling*, where the multiple objects in each training image are not fully labeled. The author leverages the structured output learning method to boost the performance of conventional one-versus-all support vector machine (OVA-SVM) [9] classifiers, and proposes an image specific structured loss function which is more appropriate to explore the dependency of predicted multiple labels. Experimental evaluation demonstrates that the proposed method is efficient to handle the issue of incomplete labeling, and it performs superior than several existing methods on image annotation task.

After discussing the special case above, the author considers a more general case of tag completion with *defective tag assignments*, where ground truth tags cover both incomplete and noisy situations. The author proposes an image-tag re-weighting scheme and formulates a re-weighted empirical loss term, which is more efficient for optimization in tag completion problem under defective tag assignments. In particular, the weighted penalty of pairwise image-tag is derived from image-image similarities and tag-tag associations, which ensures that missing and noisy tags have discriminant importance rather than identical importance in previous works. Experimental evaluations verify that using proposed re-weighting scheme achieves superior tag completion results on various degree of incomplete and noisy settings.

For the third problem, specifically, the author casts the cross-modal retrieval tasks such as text-to-image search and image-to-text search to modeling images and associated text through latent subspace learning. To handle the diversities of different modalities of data, the author develops a novel framework that consists of two procedures: (1) a coupled dictionary learning method is first developed to generate homogeneous sparse representations for different modalities by associating and jointly updating their dictionaries; (2) a coupled feature mapping scheme is then used to project the derived sparse representations from different modalities into a common subspace in which cross-modal retrieval can be performed. Experiments on a variety of cross-modal retrieval tasks demonstrate that the proposed framework outperforms the state-of-the-art approaches.

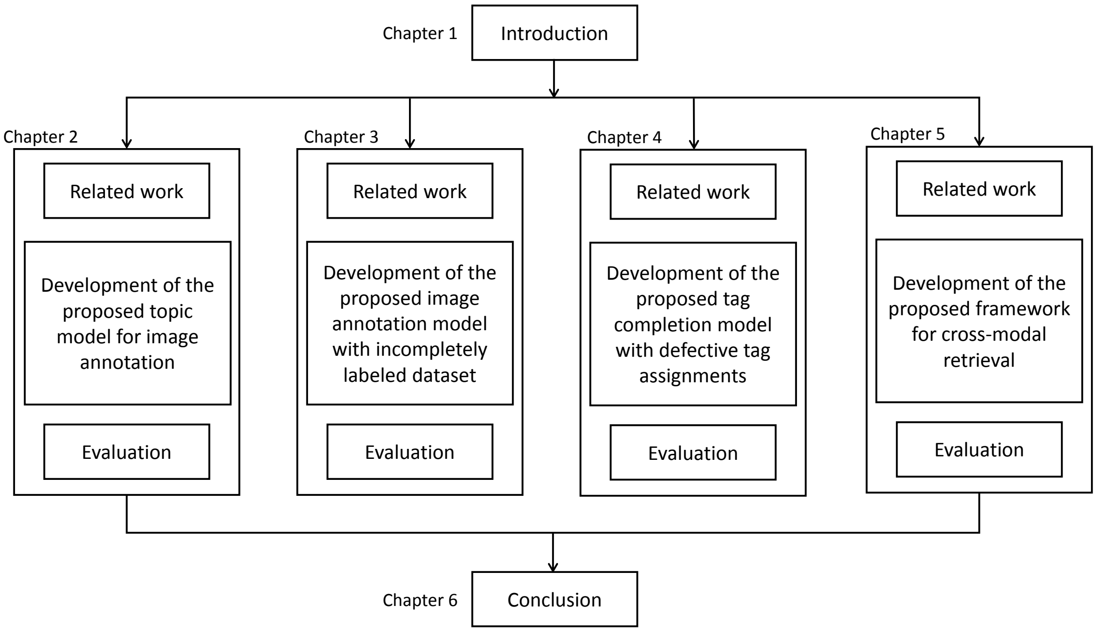

Chapter 1
Introduction
Chapter 2
Chapter 3
Chapter 4
Chapter 5
Related work
Development of the proposed topic model for image annotation
Evaluation
Related work
Development of the proposed image annotation model with incompletely labeled dataset
Evaluation
Related work
Development of the proposed tag completion model with defective tag assignments
Evaluation
Related work
Development of the proposed framework for cross-modal retrieval
Evaluation
Chapter 6
Conclusion

Fig. 1.2 Structure of the thesis.

1.3 Structure of this thesis

The structure of this thesis is as follows (see Fig. 1.2). The background and objective of the thesis are given in this chapter. In Chapter 2 "Modeling the topic correlation of images and texts via latent topic model," the author presents the details of the proposed CTM based topic models for image annotation task. This work has been published in [17, 18]. In Chapter 3 "Exploring image-label associations in incompletely labeled datasets," the author proposes a new scheme to model the image-label associations and embeds the scheme in several existing approaches that are designed for the issue of weakly labeling. This work has been published in [19]. In Chapter 4 "Tag completion by image-tag re-weighting with defective tag assignments," the author extends several existing tag completion methods by re-weighting image-tag associations for the issue of defective tag assignments. This work has been published in [20]. In Chapter 5 "Cross-modal learning for images and texts," the author turns to a more general problem of cross-modal retrieval and develops a novel framework to learning effective subspace from multimodal data of images and texts. This work has been published in [21]. Finally, in Chapter 6, the author concludes the thesis and presents the future works.

Chapter 2

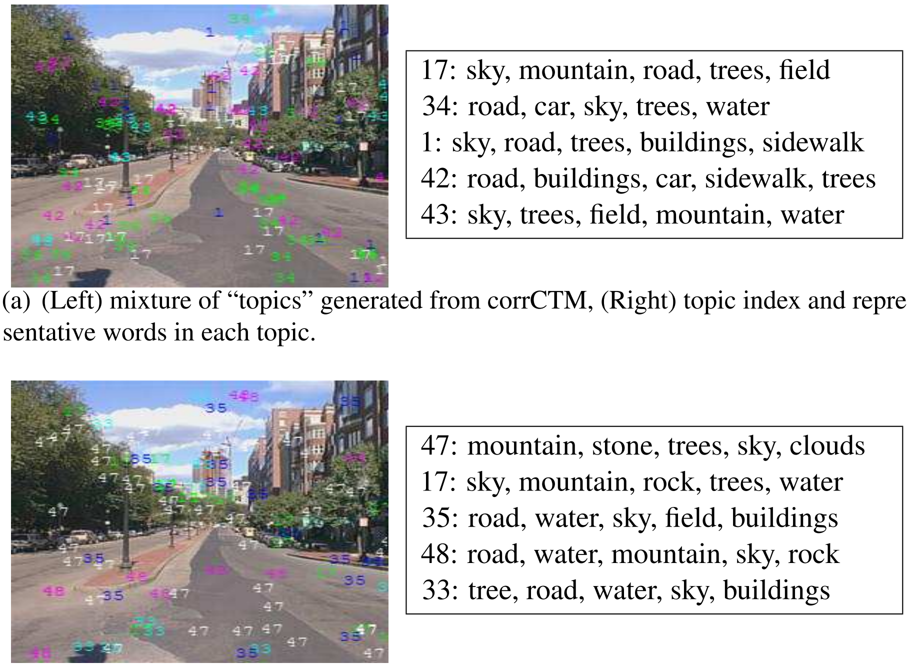
Modeling topic correlation of images and texts via latent topic model

2.1 Introduction

Automatic image annotation is an active and challenging problem in computer vision research. Given an unseen image with no caption, the goal of image annotation is to predict relevant textual words from a word vocabulary to describe the visual contents of that image.

The primary work in this chapter is built on the special case of latent topic models. The concept of *latent topic* is introduced to reduce original higher dimensional image feature space and annotation word space to lower dimensional topic space, under the assumption that each image feature and annotation word can exhibit multiple components of “topics.” Previous LDA based topic models for image annotation are all under Dirichlet assumption: topic proportions of an image are randomly drawn from a Dirichlet distribution [6]. Under Dirichlet assumption, each topic proportion is assigned independently, which leads to an unrealistic limitation that the presence of one topic is not correlated with the presence of others. Actually, annotation words of an image have correlation upon image content, which could be modeled by topic correlations. To the author’s knowledge, there have been no attempt in the past which directly considers topic correlation in annotation problem. Inspired by the pioneering work of CTM [16], which produces correlated topic proportions for words in one document using logistic normal mapping, the author adopts the methodology of CTM and extends it to the image annotation task.

The main contributions of the work in this chapter are threefold. First, the author makes the first trial to extend CTM to image annotation problem, and redesigns the popular LDA based topic models such as correspondence LDA (corrLDA) [22], supervised LDA with



17: sky, mountain, road, trees, field
34: road, car, sky, trees, water
1: sky, road, trees, buildings, sidewalk
42: road, buildings, car, sidewalk, trees
43: sky, trees, field, mountain, water

(a) (Left) mixture of "topics" generated from corrCTM, (Right) topic index and representative words in each topic.

47: mountain, stone, trees, sky, clouds
17: sky, mountain, rock, trees, water
35: road, water, sky, field, buildings
48: road, water, mountain, sky, rock
33: tree, road, water, sky, buildings

(b) (Left) mixture of "topics" generated from corrLDA, (Right) topic index and representative words in each topic.

Fig. 2.1 Modeling topic correlation to formulate semantically compact topic.

binary response (sLDA-bin) [8] as CTM based models: corrCTM and sCTM-bin, respectively. In the proposed two CTM based models, topic correlation is explicitly incorporated. Second, the author derives a general mean-field variational algorithm for parameter estimation in these CTM based models. Third, the author makes an informative comparison of annotation performance between the proposed CTM based and previous LDA based models through standard measure criteria, which is rarely provided in previous works [8, 22–24].

Through the experiments, it can be observed that modeling topic correlation has two promising advantages that improve annotation performance. Firstly, image would be assigned with correlated topics, each topic tends to be semantically compact and represented by more co-occurred words. For example, in Fig. 2.1 the topic representation in the proposed corrCTM and previous corrLDA models are different, and the final annotation words predicted by corrCTM are {"*sky*," "*trees*," "*buildings*," "*road*," "*sidewalk*"}, which are more precise than {"*sky*," "*mountain*," "*water*," "*rock*," "*trees*"} by corrLDA. Secondly, images share similar empirical topic proportions would have similar visual content and related annotation words in proposed sCTM-bin model, whereas this internal correlation is overlooked in previous sLDA-bin, visualized example can be seen in Fig. 2.6.

In the next section, the author discusses the related work of topic models in the field of image annotation, and then takes two typical CTM based topic models: corrCTM and sCTM-bin, for instance, to depict a general framework of modeling topic correlation in the CTM based models.

2.2 Related work

Topic models have been widely applied and developed in the natural language processing community, such as text classification, clustering, sentimental analysis and etc. The representative methods include Latent Semantic Analysis (LSA) [7], probabilistic Latent Semantic Analysis (pLSA) [25], and LDA [6]. They are effective at discovering the underlying topics in text documents, and at modeling more than single words. Topic models have been extended to the image domain by replacing text words with image descriptors. The resulting models have been applied to image recognition. For example, in [26], LDA was used to classify 13 scene categories of images.

When applying the topic model to image annotation problem, unlike the original topic models that contend with the problem of unimodal text, here we need to consider two modalities (image and words). The essential problem in applying the topic models to image annotation task is how to define latent variables (topics) to associate images and words, and then develop efficient algorithm to estimate these variables [27]. A seminal work of Blei *et al.* proposed two association models: Gaussian Multinomial LDA (GM-LDA) and corrLDA, which extend the basic LDA model to learn the joint distribution of texts and image features. In order to capture the correlations between the two modalities, the two models use a set of shared latent variables to represent the underlying causes of cross-correlations in the two modalities. Specifically, GM-LDA samples the latent variables of each region feature and word using a multinomial distribution specific to each sample. Parameters of multinomial distributions are sampled with a Dirichlet distribution, which is tuned with a hyper parameter. With this model, it is possible to represent multiple region features and words within a sample as a mixture of multiple topics. This property makes the model highly expressive. To overcome this drawback, in corrLDA each word direclty shares a hidden topic variable with a randomly selected image region, it achieved better performance than GM-LDA.

Motivated by the success of corrLDA model, later several approaches have been proposed to extend it. Wang *et al.* [24] proposed a supervised LDA (sLDA) model that can perform image annotation and image classification simultaneously. In sLDA, a linear regression module is added to corrLDA, in order to allow a real-valued response variable (class category) to be predicted from the empirical topic proportions of an image. Putthividhya *et al.* [8] then extended the sLDA model to handle a multi-variate binary response variables of the annotation data, their model is called sLDA-bin and compares favorably with corrLDA on image annotation task. Moreover, unlike the corrLDA that shares a set of latent topics between images and words, a new scheme of correlating separate latent topics of images and words by a regression module is introduced in [23], and this model is termed topic regression multi-modal LDA (trmmLDA).

In general, for the image annotation problem, previous LDA based topic models can be categorized into two types: 1) C-type, which builds correspondence between each image feature and each annotation word. Each annotation word is restricted to be associated with one specific image feature, and this type includes corrLDA, sLDA and trmmLDA; 2) R-type, which treats empirical topic proportions of whole image as a reduced dimensional "feature vector," annotation words as discrete response values. Then the regression models (linear Gaussian regression or logistic regression) are used to predict annotations. This type contains sLDA-bin. In the next section, the author selects typical LDA based models: corrLDA in C-type and sLDA-bin in R-type, and extend them to the proposed CTM based models, which are termed corrCTM and sCTM-bin, respectively.

2.3 Proposed models

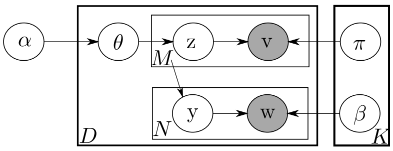
2.3.1 Data representation and notation

The data representation and notation follow the instruction from statistical text document analysis, and each image is represented as a bag of "codewords." Given a training set of images with annotation words, the following notations are used. Each image is a collection of M visual feature codewords, denoted as $\mathbf{v} = \{v_{1:M}\}$, where v_m is a unit-basis vector of size V_s with exactly one non-zero entry representing the index of current visual feature in the visual feature dictionary of size V_s. Similarly, for one image annotated with N words $\mathbf{w} = \{w_{1:N}\}$, each word w_n is denoted as a unit-basis vector of size V_t, where only one element takes value 1 and 0 otherwise, and V_t is the word dictionary size. Thus, a collection of D training image-word pairs can be denoted as $\{\mathbf{v}_{1:D}, \mathbf{w}_{1:D}\}$.

2.3.2 Modeling topic correlation in LDA based topic models

The LDA model is a generative model for a collection of exchangeable discrete data. LDA has been mainly used to model text corpora, where the notion of exchangeability corresponds to the "bag of words" assumption that is commonly employed in such model.

Later, the model of correspondence latent Dirichlet allocation (corrLDA) was extended from LDA model. The corrLDA finds conditional relationship between latent topic representations of sets of image regions and sets of words. In the corrLDA model, each image and its



(a) Graphical representation.

1. Draw topic proportions $\theta|\{\mu, \Sigma\} \sim$ Dir(α)
2. For each visual feature v_m, $m \in \{1, ..., M\}$:
   (a) Draw topic assignment $z_m|\theta \sim$ Mult(θ)
   (b) Draw feature point $v_m|z_m \sim$ Mult(π_{z_m})
3. For each textual word w_n, $n \in \{1, ...N\}$:
   (a) Draw feature index $y_n \sim$ Unif$(1, ..., M)$
   (b) Draw textual word $w_n \sim$ Mult$(\beta_{z_{y_n}})$

(b) Generative procedure.

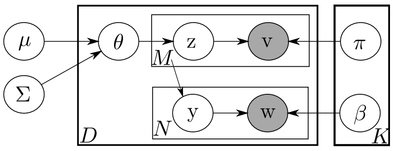
Fig. 2.2 Illustration of the corrLDA model. Note that in (a), gray circles represent observed variables, white circles represent latent variables.

corresponding words (label) are represented as a pair $(\mathbf{v}, \mathbf{w})$, where **v** and **w** represents the feature vectors of image regions and words. The region feature vectors are assumed to be distributed as a multivariate Gaussian distribution with diagonal covariance, and the words are assumed to be distributed as a multinomial distribution over the vocabulary. Moreover, different words or different regions in an image can come from different topics, and the words or the image as a whole can be viewed as combination of multiple "topics."

Let θ be a K-dimensional topic proportions from Dirichlet distribution, and let topics $\pi_{1:K}$ be K multinomials over a fixed visual feature vocabulary with size V_s, $\beta_{1:K}$ be K multinomials over a fixed text word vocabulary with size V_t. The graphical representation and the generative procedure of the corrLDA are shown in Fig. 2.2. The topic-feature matrix π_{z_m} is a V_s dimensional multinomial distribution, where z_m ranges from 1 to K. The topic-word matrix $\beta_{z_{y_n}}$ is a V_t dimensional multinomial distribution, where z_{y_n} also ranges from 1 to K. "Dir" means Dirichlet distribution, "Mult" means multinomial distribution, and "Unif" indicates uniform distribution.

A problem with drawing the topic mixture proportions from a Dirichlet distribution in LDA is that Dirichlet distributions are too simple and exhibit a near independence structure, where the correlations between the components of a Dirichlet random variable are restricted to sum to one. Therefore, the corrLDA also has the problem that the correlation of different topics cannot be efficiently captured by the Dirichlet distributions, and the relationship between image regions and words cannot be properly represented by the topic representations.

To tackle the problem in LDA model, the CTM is proposed to capture the topic correlation and is applied to use topic representation for document classification problem. Compared with LDA, the CTM uses logistic normal instead of Dirichlet to capture the topic correlation. As the logistic normal distribution can form a richer class of distributions and better captures inter-component correlations, the CTM gives a better fit than LDA in representing textual



(a) Graphical representation.

1. Draw topic proportions $\theta|\{\mu,\Sigma\} \sim$ Norm(μ,Σ)
2. For each visual feature v_m, $m \in \{1,...,M\}$:
   - **(a)** Draw topic assignment $z_m|\theta \sim$ Mult$(f(\theta))$
   - **(b)** Draw feature point $v_m|z_m \sim$ Mult(π_{z_m})
3. For each textual word w_n, $n \in \{1,...N\}$:
   - **(a)** Draw feature index $y_n \sim$ Unif$(1,...,M)$
   - **(b)** Draw textual word $w_n \sim$ Mult$(\beta_{z_{y_n}})$

(b) Generative procedure.

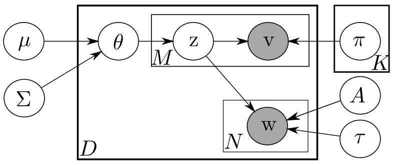
Fig. 2.3 Illustration of the corrCTM model.

documents. Furthermore, Malisiewicz *et al.* [28] has applied CTM to object recognition in images and found that CTM also has a better fit for visual representation than LDA by modeling correlated topic space rather than independent topic space.

Motivated by such success of CTM in recognition task, here the author adopts the association model of CTM for the task of annotation. In order to capture the correlation of topics, the hyper parameter of topic prior distribution is modeled as multivariate Normal distribution instead of Dirichlet, and the dependencies of different topics by covariance matrix is captured. Then the logistic normal function $f(\theta_i) = \left(\exp\ \theta_i / \sum_{j=1}^{K} \exp\ \theta_j\right)$ is used to project the multivariate normal to topic proportions for each image, where K is the topic number. The graphical representation and the generative procedure of the proposed corrCTM are shown in Fig. 2.3.

In Fig. 2.3 (b), “Norm” means normal distribution, “Mult” means multinomial distribution, and “Unif” indicates uniform distribution. In particular, firstly M image features v_m are generated from correlated topic proportions θ, conditional on the topic-feature multinomial π Secondly, for each of the N text words, one of the M features is selected and corresponding assigned to a text word w_n, conditional on the topic-word multinomial β. The generative process of corrCTM is identical to the process of corrLDA in Fig. 2.2 (a), except that the topic proportions for the image are drawn from a logistic normal distribution rather than a Dirichlet distribution. From the generative process of corrCTM, it could be learned that topic correlations are modeled and generated through the covariance matrix Σ of prior multivariate normal distribution. After the logistic normal procedure, the topic proportions are generated depending on the correlations. And topic assignment for image features and text words in the latter part would again rely on the correlated mixture of topics in the current image.

Then the author embeds the scheme to another model sLDA-bin, and develops sCTM-bin that can model the topic correlation. Similarly, the generative process of sCTM-bin is identical to the general process of sLDA-bin, except that topic proportions θ are drawn



(a) Graphical representation.

1. Draw topic proportions $\theta|\{\mu, \Sigma\} \sim \mathrm{N}(\mu, \Sigma)$
2. For each visual feature v_m, $m \in \{1, ..., M\}$:
   - **(a)** Draw topic assignment $z_m|\theta \sim \mathrm{Mult}(f(\theta))$
   - **(b)** Draw feature point $v_m|z_m \sim \mathrm{Mult}(\pi_{z_m})$
3. Given the empirical topic proportions $\bar{\mathbf{z}} = \frac{1}{M}\sum_{m=1}^{M} z_m$, for each textual word w_i in word vocabulary, $i \in \{1, ...V_t\}$:
   - **(a)** Draw textual word $w_i \sim \mathrm{Bern}(f(\bar{\mathbf{z}}, A_i, \tau_i))$

(b) Generative procedure.

Fig. 2.4 Illustration of the sCTM-bin model.

form logistic norm rather than Dirichlet. Specifically, the generative procedure of the sCTM-bin models are depicted in Fig. 2.4. Here $f(\bar{\mathbf{z}}, A_i, \tau_i)$ is a logistic regression function $f(\bar{\mathbf{z}}, A_i, \tau_i) = \sigma(A_i^T \bar{\mathbf{z}} + \tau_i)$, where A_i is a V_t dimensional vector of regression coefficients, τ_i is a regularization constant for A_i, $\sigma(x) = 1/(1 + \exp(-x))$. “Bern” means Bernoulli distribution.

It can be observed from the graph representations of the proposed corrCTM and sCTM-bin models that, different from the previous LDA based models, in the proposed models topic correlation is first incorporated in prior multivariate covariance Σ, then topic proportions θ_i, θ_j of an image are generated relying on Σ through logistic normal process $f(\theta)$. Moreover, topic assignment for image features and annotation words would subsequently depend on the correlated mixture of topic proportions, by means of building topic correspondence (C-type) or regression from empirical topic proportions (R-type).

2.3.3 Variational inference in proposed models

To learn the parameters of corrCTM that maximize the likelihood of training data, the variational Expectation Maximization (VEM) framework developed in [6] is employed to iteratively estimate the model parameters of latent variables. Generally, in the E-step, the posterior probability function is derived to approximate the lower-boundary of the joint likelihood, and then maximize the lower-boundary to calculate model parameters in the M-step. In other words, the two alternative procedures in VEM consists of variational inference and parameter estimation. The former E-step is to calculate a set of variational parameters to obtain the approximate lower-bound on likelihood of each sample. The latter M-step is to estimate the model parameters that maximize the log likelihood of the whole training samples.

To simplify the formulation, in the graphical representations of the models in Fig. 2.3 and Fig. 2.4, a symbol Ψ is used to represent the model parameters in each model, e.g. $\{\mu,\Sigma,\pi,\beta\}$ in corrCTM and $\{\mu,\Sigma,\pi,A,\tau\}$ in sCTM-bin, and a symbol Δ is used to denote the latent variables in each model, e.g. $\{\theta,z,y\}$ in corrCTM and $\{\theta,z\}$ in sCTM-bin. In addition, a symbol Ω is used to denote the parameters of varational distributions introduced during the VEM procedure, details can be referred to Appendix A. Then during the learning procedure, the goal is to maximize the lower boundary $\mathscr{L}(\Omega;\Psi)$ of log joint likelihood of observed image-word pairs {**v, w**},

$$\log p(\mathbf{v},\mathbf{w}|\Psi) \geq \mathscr{L}(\Omega;\Psi), \tag{2.1}$$

where Ω contains variational parameters approximated to latent variables in Δ. The variational EM framework is utilized to maximize $\mathscr{L}(\Omega;\Psi)$. In the E step, the variational parameters Ω is approximated to lower bound the marginal likelihood , then maximize the lower boundary in M step to obtain model parameters Ψ. The detailed deriving process is presented in Appendix A, and here the update rules of the variational parameters γ, ν for θ, ϕ for z, and model parameters μ, Σ are summarized as follows.

E-step: For each image, optimize values for variational parameters in Ω. The update rules of γ, ν, ϕ which differ from LDA based models are

$$\frac{\partial\mathscr{L}}{\partial\gamma} = -\Sigma^{-1}(\gamma-\mu) + \sum_{m=1}^{M}\phi_{mi} - \frac{M}{\zeta}\exp\,(\gamma+\nu^2/2), \tag{2.2}$$

$$\frac{\partial\mathscr{L}}{\partial\nu_i^2} = -\Sigma_{ii}^{-1}/2 - (M/2\zeta)\exp\,(\gamma_i+\nu_i^2/2) + 1/(2\nu_i^2), \tag{2.3}$$

$$\hat{\phi_{mi}} \propto \pi_{i,v_m}\exp\,\left(\gamma_i + \sum_{n=1}^{N}\lambda_{nm}\log\,\beta_{i,w_n}\right), \tag{2.4}$$

$$\hat{\phi_{mi}} \propto \pi_{i,v_m}\exp\left[\gamma_i + \sum_{j=1}^{V_t}\left(\frac{\lambda(\xi_j)}{2M}A_j - \frac{\lambda(\xi_j)}{M^2}\mathrm{diag}(A_jA_j^T) \;+\; 2A_jA_j^T\sum_{n\neq m}\phi_{ni}\right)\right], \tag{2.5}$$

where i is topic index. Intermediate variables such as ζ, ξ, λ are the same as [8, 22], and the other variational parameters specific in corrCTM and sCTM-bin follow the similar iteration rules in [8, 22]. The update rule of ϕ_{mi} in Equation 2.4 is for corrCTM and the one in Equation 2.5 for sCTM-bin, respectively.

M-step: For all images in a training set, to estimate model parameter Ψ, maximize lower boundary of summational log likelihood $\sum_{d=1}^{D}\mathscr{L}_d(\Omega_d;\Psi_d)$ for all D images, given Ω_d obtained from the E-step. The update rules of μ, Σ, which are different from LDA based models, are

$$\hat{\mu} = (1/D)\sum_{d=1}^{D} \gamma_d, \tag{2.6}$$

$$\hat{\Sigma} = (1/D)\sum_{d=1}^{D} \left(I\, \nu_d^2 + (\gamma_d - \hat{\mu})(\gamma_d - \hat{\mu})^T\right), \tag{2.7}$$

where d is image index in the training set. The other model parameters such as π, β in corrCTM, and π, A, τ in sCTM-bin follow the similar iteration rules as in [8, 22].

Generally speaking, the computational complexity of two proposed CTM based models corrCTM and sCTM-bin are higher than their LDA based versions corrLDA and sLDA-bin. The reason is that the proposed CTM based models introduce the new parameters μ and Σ of logistic normal distribution and the correspondingly variational parameters γ and ν. Computing μ (Equation 2.6) and γ (Equation 2.2) both require complexity of $\mathcal{O}(NK)$, and computing Σ (Equation 2.7) and ν (Equation 2.3) both require complexity of $\mathcal{O}(NK^2)$. Note that, for the update rules of the parameters in the CTM based models, the rules for γ in Equation 2.2, and ν in Equation 2.3 are not closed-form. Thus the optimal values of these parameters can be calculated iteratively. In practice, the (Limited-memory BFGS) LBFGS[1] algorithm is used to update γ and a constrained Newton's method to update ν.

In addition, the scheme of capturing topic correlation can also be observed in the update rules of the parameters. For example, in Equation 2.4 and 2.5, the per-feature variational distribution over topics ϕ_{mi} depends on variational hyper parameter γ_i and ν_i of different topic proportions. As the variance ν_i is derived from prior covariance matrix Σ, the covariance matrix Σ controls topic proportions of each image, and different topic components assigned to feature point (ϕ_{mi}) would be correlated.

2.4 Experimental result

2.4.1 Configuration

The proposed two models are evaluated on three benchmark datasets: LabelMe [29], PASCAL VOC07 [30] and Corel 5K [31]. The detailed information of the three datasets are as follows:

- **LabelMe**: The annotation words that occur less than 3 times are removed from the original dataset, resulting in a word dictionary with 136 entries and 2,687 images (2,149 for training, 538 for testing).

[1] Matlab implementation of the LBFGS algorithm can be found in http://www.caam.rice.edu/~heinken/

Table 2.1 Statistics of three datasets: LabelMe, PASCAL VOC07 and Corel 5K.

	LabelMe	PASCAL VOC07	Corel 5K
Number of labels	136	186	260
Number of training images	2,149	4,340	4,500
Number of testing images	538	1,860	499

- **PASCAL VOC07**: For this 20-category dataset, after removing the words that appears less than 5 times, a subset containing 6,200 images (4,340 for training, 1,860 for testing) with 186 words is preserved.
- **Corel 5K**: The original Corel 5K dataset is used in the experiment without pruning. This dataset includes 5,000 images (4,500 for training, 500 for testing) and 260 words.

Table 2.1 summarizes the statistics of these datasets.

Following the configuration in [8, 22], the author uses the 128-dimensional SIFT [32] descriptor as a feature descriptor to represent each image. The SIFT feature has shown to be effective in image classification and recognition tasks. The public tool VLFeat[2] is used to extract 128-dimensional dense SIFT descriptor on 20×20 gray-scale patches in each image. And then a codeword dictionary is built for LabelMe and PASCAL VOC07 with the size 500, and for Corel 5K, with the size 1000. To train the model, the stopping criteria for log likelihood change is less than 0.001%, and the the max number of iterations is 20 for variation inference, 100 for parameter estimation. All the experiments are conducted on a computer with dual 4-core Intel i7 2.8GHz CPU, 8GB RAM.

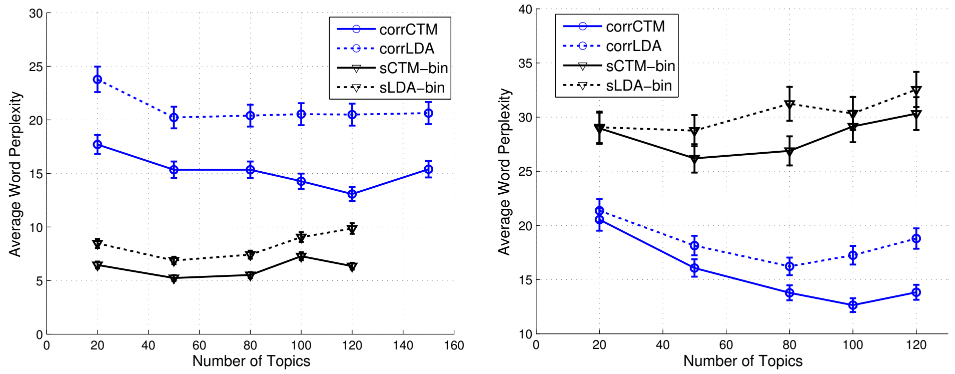
2.4.2 Predictive perplexity

To measure the quality of codewords and annotations predicted by the models, two standard measures: "feature perplexity" and "word perplexity," which are proposed in [6] are adopted. Introduced from conventional language modeling, the perplexity is equivalent to the inverse of the geometric mean likelihood (in Equation 2.8 and 2.9, denoted as $\mathscr{P}(*)$), a lower perplexity score indicates better predictive performance),

$$\mathscr{P}(feature) = \exp\left(-\frac{\sum_{d=1}^{D}\sum_{m=1}^{M_d}\log p(v_m|z_m)}{\sum_{d=1}^{D}M_d}\right), \tag{2.8}$$

$$\mathscr{P}(word) = \exp\left(-\frac{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\log p(w_n|\mathbf{v}_d)}{\sum_{d=1}^{D}N_d}\right). \tag{2.9}$$

[2] http://www.vlfeat.org/



(a) LabelMe

(b) PASCAL VOC07

Fig. 2.5 Word perplexity of CTM based and LDA based models on datasets: (a) LabelMe, (b) PASCAL VOC07.

Fig. 2.5 shows the word perplexity of the proposed CTM based models and previous LDA based models on two standard datasets with various number of topics. Note that on LabelMe dataset, the results of the models of sLDA-bin and sCTM-bin overfit with 150 topics are not shown as they overfit with 150 topics. As seen in Fig. 2.5, both corrCTM and sCTM-bin give lower perplexity score for different number of topics compared with their LDA based versions corrLDA and sLDA-bin. It shows the superior predictive capabilities of capturing topic correlation, propagating this correlation to build correspondence between topics and words in C-type models, or to perform regression from topics to words in R-type models. These capabilities are also hinted in the previous exemplars Fig. 2.1 and Fig. 2.6. Note that, in Fig. 2.6, the second column is a test image and its histogram of empirical topic proportions, the third to fifth columns are "neighbor images" selected according to nearest Hellinger distance [16]. The italic words indicate predicted annotations for the test image, while the normal words are ground truth annotations for "neighbor images."

2.4.3 Visualizing topic correlation

Since topic representation can decompose image-word relations to image-topic and word-topic relation in corrCTM, the author then investigates how the proposed corrCTM better propagates the correlation to word-topic. Taking dataset Corel 5K for example, the author first establishes the normalized co-occurred frequency matrix for pair-wise words in the vocabulary dictionary, and then obtains empirical frequently co-occurred word pairs by setting a threshold to the matrix. Finally, there are 1434 word pairs can be obtained from the total 260 words, given a global threshold 0.25, which means that when one word occurs its partner would also occur with conditional probability larger than 0.25.

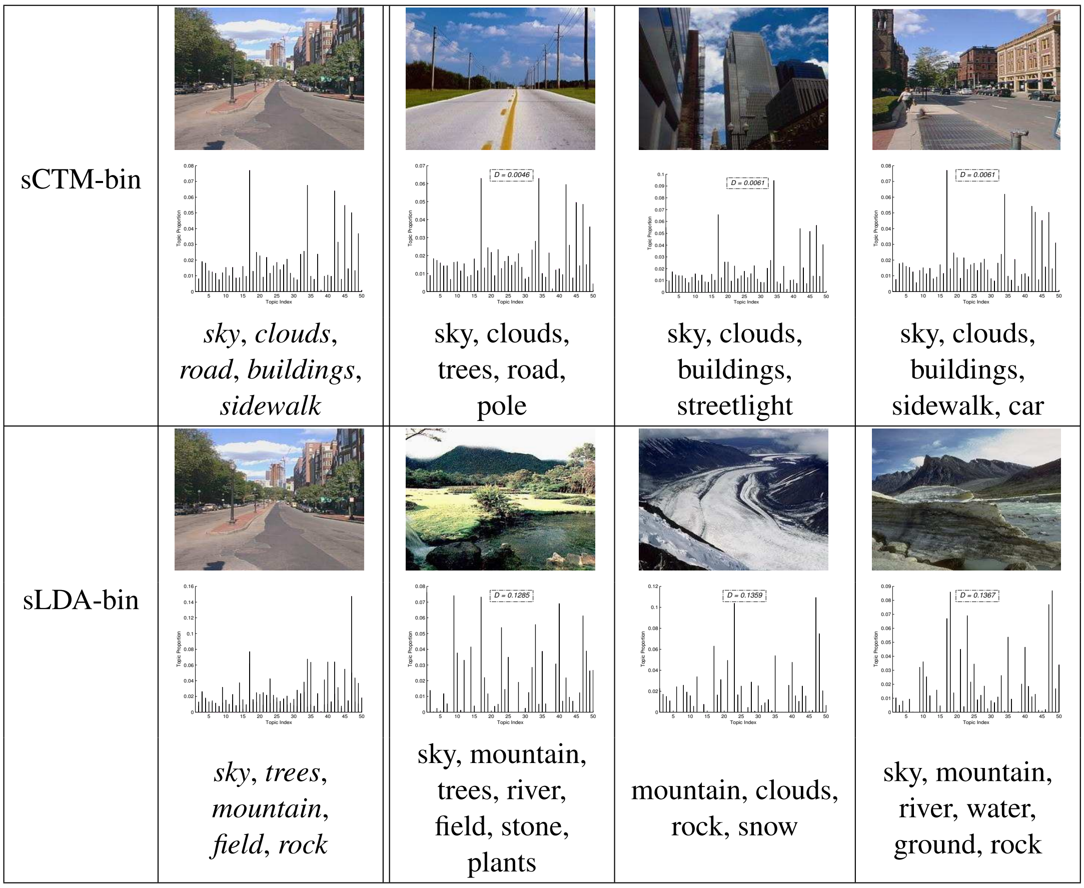


Fig. 2.6 Modeling topic correlation to formulate correlated empirical topic proportions.

The goal is to examine how well the proposed corrCTM fits the empirical co-occurred word pairs. According to the caption-topic multinomial matrix β in Fig. 2.2 (b), which is a $V_t \times K$ matrix and each column implies the multinomial distribution among words for current topic. Top 10 caption words are selected as most representative for each topic, and then make a statistic to recall the count of the appeared co-occurred word pair in the empirical co-occurred word pairs. The recall count for both datasets compared with corrLDA (#recall count / #total empirical count) is reported in Table 2.2. And Table 2.3 shows some typical co-occurred word pairs in top-10 words for topics in Corel 5K, where the total topic number K is 80, and different word pairs from various semantic meanings are marked with different symbols (†, ‡ and §). It is indicated that corrCTM owns the ability to cluster more related word pairs to formulate compact topic representation compared with corrLDA. The reason is that from the generative procedure of corrCTM, once correlated topic proportions are assigned to image features, the co-occurred words to related topics may also be probable. As a result, probability value for co-occurred word pairs would be lifted in the topic-word multinomial β.

Table 2.2 Recall count for co-occurred words: corrCTM *vs.* corrLDA

	PASCAL VOC 2007	Corel 5K
corrCTM	232/621	348/1,434
corrLDA	185/621	276/1,434

Table 2.3 Typical co-occurred word pairs in top-10 representative words of each topic: corrCTM (top panel) and corrLDA (bottom panel)

Topic index	Typical word pairs
Topic 1	cars†, prototype†, tracks†, street, turn, marsh‡, roofs, bengal§, forest‡, tiger§
Topic 4	plane†, jet†, sky†, sun, birds§, fly§, clouds§, snow, sand†, dunes†
Topic 27	snow†, ice†, polar†, frozen†, bear, mountain§, water, rocks§, grass, sky
Topic 48	island†, beach†, sand, sea†, water†, sky, people, kauai†, sunset, buildings
Topic 72	ocean†, coral†, fish†, rocks§, reefs§, water, orchid, boat§, sky, fan
Topic 1	water, sky†, tree, people, clouds†, grass, mountain, buildings, sun, snow
Topic 9	sky†, jet†, plane†, mountain, tree, water, sun, people, clouds, buildings
Topic 30	tree†, grass†, flowers§, people, field, house, mountain, sky, water, garden§
Topic 41	ice†, people, mountain§, sky, frost, snow§, clouds, water, rocks§, landscape
Topic 67	cars, buildings§, street†, people, sidewalk, lights†, window†, post, store‡, shops‡

It can be learned that modeling topic correlation has two promising advantages that improve annotation performance. Firstly, image would be assigned with correlated topics, where each topic tends to be semantically compact and represented by more co-occurred words (see Fig. 2.1 the topic representation in the proposed corrCTM and the previous corrLDA models). Secondly, as shown in Fig. 2.6, images sharing similar empirical topic proportions would have similar visual content and related annotation words in the proposed sCTM-bin, whereas this internal correlation is overlooked in the previous sLDA-bin.

2.4.4 Evaluation of annotation performance

Previous works [8, 22, 24] rarely provide comprehensive statistic of annotation performance by modern measures, resulting in ambiguous judgment. To provide a comprehensive comparison, the standard measures widely used in recent works [13, 14, 33] are used for evaluation. The standard measures include 4 types: (1) percentage mean precision P, (2) percentage mean recall R, (3) $F1$ score $F1 = 2\frac{P*R}{P+R}$, (4) total recalled words count $N+$. The two proposed CTM based models are compared with their LDA based versions. To present the best performance of each model, the optimal topic numbers for each model is selected via cross validation. And for each test image, top-5 words is predicted by each model.

Table 2.4 summarizes the overall evaluations from the proposed results as well as those reported by previous models for image annotation on three datasets. It can be learned that, on most of the measures, the CTM based models defeat their LDA based versions, on all three datasets. Precisely, for LabelMe, PASCAL VOC07, Corel 5K datasets, corrCTM improves corrLDA 3.03%, 0.85%, 2.03% on F1 score and 11, 5, 12 on $N+$, sCTM-bin outperforms sLDA-bin for 3.54%, 2.74% , 4.22% on F1 score, 4, 6, 2 on $N+$. In general, the proposed models gain 3.54%, 0.85%, 2.03% in terms of F1 score, and 8, 5, 4 with regards to $N+$ compared with the best results from both sides of the CTM based models and the LDA based models on three datasets respectively.

It is notable that the composition of datasets, such as variety of image contents, balance of image counts and word counts, highly affects the performance of the different types of models. It can be observed that R-type (sLDA-bin, sCTM-bin) models are better on uniform dataset LabelMe, in which the regression for topic-word seems to be robust, but more word counts will reduce the regression efficiency. On the other hand, C-type (corrLDA, corrCTM) models, which build correspondence between topic-word support larger number of topics , and do not easily suffer from over fitting on datasets, e.g. PASCAL VOC07, Corel 5K, which contain much more various image content and annotations.

Fig. 2.7 demonstrates some exemplar annotation results from the proposed CTM based models. In most scenarios, annotations predicted by the CTM based models are more related to the contents of image and ground truth annotations, while the LDA based models incline to prefer weird words (e.g. “boat,” “train,” “bird,” “snow”) which are actually irrelevant to image content.

Table 2.4 Comparison of annotation performance between proposed CTM *vs.* previous LDA based models. Promising results in both parts are highlighted in bold.

	LabelMe					PASCAL VOC07					Corel 5K				
Method	K[3]	P	R	F1	$N+$	K	P	R	F1	$N+$	K(L)[4]	P	R	F1	$N+$
corrLDA [22]	50	43.6	34.4	38.46	54	80	17.8	27.9	21.28	49	80	24.9	35	29.09	53
sLDA-bin [8]	50	44.6	35.3	39.4	57	50	12.2	22	15.69	31	80	17.5	25.8	20.85	74
corrCTM	120	46.9	37.2	41.49	**65**	100	18.3	28	**22.13**	**54**	120	26.7	37.3	31.12	65
sCTM-bin	50	48.7	38.4	**42.94**	61	50	14.5	25.3	18.43	37	80	21	31.1	25.07	**76**

Ground truth	sky, trees, brushes, field	sand, beach, mountain, sea, water	bus, buildings, red	airplane, green, sheep, fence, farm
corrLDA	sky, trees, mountain, rock, water	sky, water, river, mountain, clouds	bus, buildings, street, motorcycle, station	airplane, bird, sheep, stone, snow
sLDA-bin	trees, buildings, ground, road, clouds	coast, sun, boat, stone, shrub	truck, train, station, railway, motor	horse, cow, house, mountain, snow
corrCTM	sky, trees, clouds, grass, field	sea, water, sand, beach, coast	bus, buildings, street, window, shop	airplane, trees, sheep, green, farm
sCTM-bin	sky, trees, plants, field, hill	sea, water, sky, beach, sand,	bus, car, buildings, street, red	clouds, fence, sheep, farm, home

Fig. 2.7 Annotations for exemplar images: left two images are from LabelMe, right two images are from PASCAL VOC07.

2.5 Summary

In this chapter, the author has developed a series of CTM based topic models to capture topic correlation, and provides a comprehensive comparison between the proposed CTM based and previous LDA based models on annotation performance. It indicates that modeling topic correlation could benefit annotation performance, and that the CTM based models have either superior or competitive annotation results compared with LDA based models. In addition, more co-occurred words can be recalled by the CTM based models through modeling topic correlation. However, the computation of estimating model parameters μ, Σ in the CTM based models are more complex, and the approximation for variational parameters are sometimes numerical unstable in practice. Therefore, it remains a challenge to apply the CTM based model on large-scale datasets, e.g., ESP Game, IAPRTC-12. This would be a feasible direction of the future work.

Chapter 3

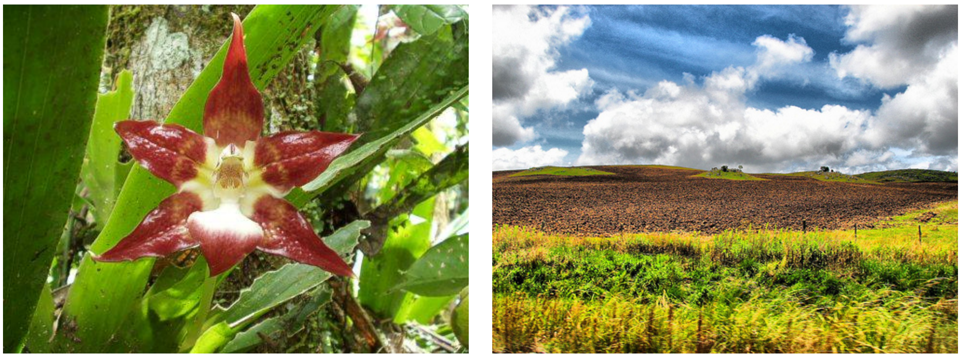
Exploring image-label associations in incompletely labeled datasets

3.1 Introduction

In Chapter 2, the author discusses the traditional image annotation problem and develops a scheme to model topic correlation in the previous LDA based topic models. In the problem setting of traditional image annotation, the ground truth labels assigned to the images are assumed to be compact and perfect. In other words, the ground truth labels completely describe the content of the images. However, this assumption is not reasonable as the ground truth labels are assigned by human annotators manually. Due to unconsciousness and misconception of the annotators, the ground truth labels indeed tend to be ambiguous, incomplete and even imprecise.

In this chapter, the author considers a practical issue of *incomplete labeling* that the images in the training set are not completely annotated with all relevant labels. As shown in Fig. 3.1, the two images from benchmark datasets IAPRTC-12 and NUS-WIDE (both datasets are collected from social media such as Flickr) have few annotated labels and suffer from the problem of incomplete labeling. Potentially correct labels such as {*"flower," "plant," "tree," "trunk"*}, {*"sky," "grass"*} are missed from the ground truth labels of two images, respectively.

Thus, when applying the traditional annotation methods on incompletely labeled datasets, the annotation performance can hardly achieve optimal since the insufficient annotations of the datasets could potentially lead to biased estimation of the prediction model. For example, a recent study of [19] reports that the annotation performance of a baseline method OVA-SVM decays drastically when the degree of incompleteness of training set increases.



(a) *Ground truth:* bloom, leave (b) *Ground truth:* clouds

Fig. 3.1 Examples of incomplete labeling: (a) an image from IAPRTC-12 dataset and (b) an image from NUS-WIDE dataset.

To tackle the issue of incomplete labeling, one group of recent ongoing researches [10, 34–36] directly modify conventional image annotation prototypes such as multi-label ranking [34, 36], binary SVM [10], and ridge regression [35], by incorporating additional consistency between visual and semantic cues in images and labels. In addition, the performance of these methods largely depends on the assumption of consistency. Moreover, another group of works aim at boosting the conventional annotation models and adding a new learning stage incrementally under the incomplete setting. The techniques utilized in the new learning stage could be multi-task learning [11], ensemble learning [37], and structured output learning [38, 39].

In particular, the structured output learning technique has been attracting much attention since it allows the learning of functions with complex outputs, such as object poses, segmentations, and parse trees. It has been successfully applied to the problems such as object detection, image segmentation and image parsing. It is notable that the structured output learning technique is an efficient scheme to handle the difficulties of incomplete labeling in image annotation tasks. First, it captures interdependencies of multiple labels from the structure in the output space. Secondly, the weak learning manner allows it to explore the potential usages of missing labels, and those missing labels can be captured by latent variables [40].

Specifically, in the celebrated work of [39], McAuley *et al.* propose a model termed OVA-SSVM that combines conventional OVA-SVM with structured output learning method. In the OVA-SSVM model, the structured output learning method is adopted to boost the performance of pre-trained OVA-SVM classifiers under the incomplete setting, and a structured loss function of image classification is adopted to benefit prediction of missing labels.

Promising prediction results are obtained on the classification task conducted on the ImageNet dataset [3] where each training image has a single label. However, due to the limitation of structured loss function used, the OVA-SSVM method may not be well extended to more practical circumstances where each training image has multiple labels.

Therefore, in this chapter, the author puts effort to improve OVA-SSVM in three aspects:

- A new method is proposed to tackle with more general case where each training image has one or more labels. The proposed method is more efficient for the multi-label annotation problem than the OVA-SSVM method
- A novel image specific structured loss term is formulated, which is more appropriate than the classification-style structured loss used in OVA-SSVM, to account for the dependencies of predicted multiple labels of a specific image
- An efficient optimization algorithm with lower complexity is developed by exploiting the properties of the proposed structured loss.

Extensive evaluation on two benchmark annotation datasets with various settings of incompleteness are performed on image annotation tasks. The empirical results demonstrate that the proposed method is significantly better than OVA-SSVM, and achieve competitive annotation performance compared with other state-of-the-art methods designed for incomplete labeling.

3.2 Related work

Generally, image annotation can be viewed as a multi-label classification problem as each image is associated with multiple labels. Several approaches cast the problem of multi-label classification to separately train a discriminative classifier for each label, and use these classifiers to predict the presence/absence of each particular label. Usually, SVM [9], boosting [41], random forest [42] are typically applied to learn the discriminative classifiers. It is worth noting that the OVA-SVM that has been widely used on image classification tasks [43, 44] also shows promising results on image annotation task, and hence, it is usually treated as a baseline for annotation model evaluation. However, these approaches neglect the correlation of different labels that are useful for model learning.

To exploit the dependencies among labels, several efficient schemes are proposed in the multi-label classification framework, e.g., incorporating the correlation of labels from prior label information [45], using hypergraph regularization to constrain the label correlation [2], constructing bi-relational graph to model the label correlation and image similarity [46]. The main shortcoming of multi-label classification methods is that in order to train a reliable annotation model, they require a large set of training images with clean and complete manual annotations.

Besides the multi-label classification methods, several advanced machine learning techniques have been applied, including generative and nearest neighbor (NN)-based methods. The generative methods [5, 22, 47], such as topic model [22], learn the joint distribution of labels and image features, where different hypotheses of probability distribution can be assumed, such as multinomials, separate Bernoullis, mixture of Gaussian. The NN-based methods [12–14, 48] perform image annotation via propagating labels from visually similar images, since such methods generally assume that visually similar images probably share common labels. For the NN-based methods, the primary research issue focuses on designing an efficient label propagation scheme [12, 13, 15] and selecting appropriate visual neighbors [14, 48, 49]. Similar to the multi-label classification methods, to achieve good performance, the generative and NN-based methods also require a large number of well annotated images.

3.3 Image annotation under incomplete labeling

3.3.1 Problem formulation

In this section, the author first introduces the conventional OVA-SVM used for image annotation task, and then describes the OVA-SSVM method that uses structured output learning for boosting the conventional OVA-SVM classifiers under the incomplete setting. Some notations used in the following sections are also defined in this section.

Conventional OVA-SVM [9]

Let $\mathscr{T} = \{(x^1,Y^1),...,(x^N,Y^N)\}$ be an incompletely labeled dataset, where $x^n \in \mathscr{X}$ represents an image feature vector, $Y^n \subseteq \mathscr{Y}$ is a set of labels, where $\mathscr{Y} = \{y_1,...,y_C\}$ is the vocabulary of C labels. Note that Y^n is a subset of the ideally full set Ω^n of ground truth labels for image x^n. The goal is to learn an annotation model that, for an unseen image x, outputs an optimal set $\hat{Y}$ consisting of K distinct labels. Usually, K is set to be a fixed value for different test images, to simplify the problem setting and the model evaluation.

A conventional annotation model consists of learning a series of binary OVA-SVM classifiers that distinguish a single label from all other. In other words, for each class $y_c \in \mathscr{Y}$, OVA-SVM learns a separate parameter vector $\mathbf{w}_{OVA}^{y_c}$. where Fig. 3.2 (a) shows the learning procedure of OVA-SVM and the learnt binary classifiers $\{w_{OVA}^{y_c}\}_{c=1}^{C}$ of all labels. To predict a set of K labels $\hat{Y}$ for an unseen image x, the annotation model simply returns the labels

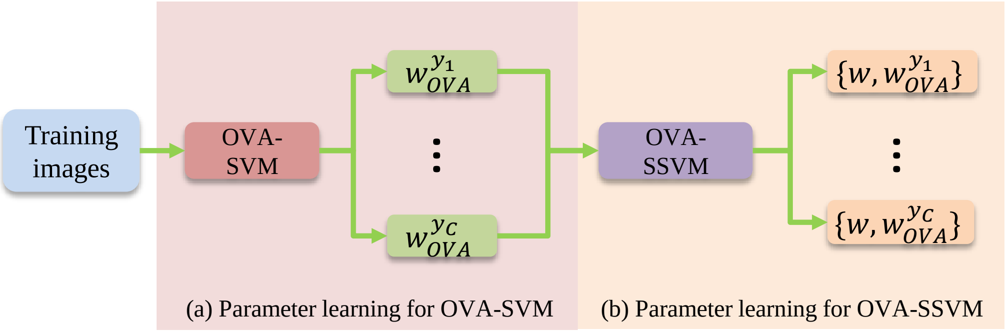


Fig. 3.2 Procedure of parameter learning for OVA-SVM and OVA-SSVM methods.

with the K highest scores performing on classifiers of all labels:

$$\hat{Y} = \arg\max_{Y \subseteq \mathscr{Y}} \sum_{y_c \in Y} x \cdot \mathbf{w}_{OVA}^{y_c}, \tag{3.1}$$

where $Y \subseteq \mathscr{Y}$ represents any possible output set containing K labels. It is worth noting that the annotation model of OVA-SVM classifiers has the following limitations that (1) the one-versus-all learning manner ignores the dependencies of labels, which implies that the OVA-SVM optimizes the prediction of only one single output label, ignoring the "structure" altogether; (2) the performance of OVA-SVM classifiers would drop drastically when incomplete labels for training image are provided.

OVA-SSVM [39]

To overcome the disadvantages of conventional OVA-SVM and to exploit the structured associations in output label set Y, the structured output learning method OVA-SSVM [39] refers to the training set consisting of structured input-output pairs $\mathscr{T} \in (\mathscr{X} \times \mathscr{Y})^N$. The prediction rule of optimal output labels $\hat{Y}$ for an unseen image x is

$$\hat{Y} = \arg\max_{Y \subseteq \mathscr{Y}} \Phi(x,Y) \cdot \mathbf{w} = \arg\max_{Y \in \mathscr{Y}} \sum_{y \in Y} \phi(x,y) \cdot \mathbf{w}, \tag{3.2}$$

where Φ is the joint feature vector that describes the relationship between input x and any structured output Y, ϕ is the joint feature vector for input x and a single label y in Y, and $\mathbf{w}$ is the parameter vector to be learnt. In particular, given a set of pre-trained OVA-SVM classifiers $\{\mathbf{w}_{OVA}^{y_c}\}_{c=1}^{C}$, the joint feature vector $\Phi(x,Y)$ in OVA-SSVM is defined as

$$\Phi(x,Y) = \sum_{y \in Y} x \circ \mathbf{w}_{OVA}^{y}, \tag{3.3}$$

where $x \circ \mathbf{w}_{OVA}^{y}$ represents the Hadamard product of x and $\mathbf{w}_{OVA}^{y}$. Then the annotation model in Equation 3.2 can be formulated as

$$\hat{Y} = \arg\max_{Y \subseteq \mathscr{Y}} \sum_{y \in Y} \langle x \circ \mathbf{w}_{OVA}^{y}, \mathbf{w} \rangle. \tag{3.4}$$

We can learn from Equation 3.4 that OVA-SSVM incrementally learns a single parameter vector $\mathbf{w}$ that re-weights the parameter vectors of existing OVA-SVM classifiers $\{\mathbf{w}_{OVA}^{y_c}\}^{y_c \in \mathscr{Y}}$ (the learning procedures of OVA-SSVM is illustrated in Fig. 3.2 (b)), and incorporates the structured nature of output Y through the joint feature vector $\Phi(x,Y)$.

Moreover, for the setting of incomplete labeling of $\mathscr{T}$, the input-output relationship is not completely characterized by $(x,Y) \in \mathscr{X} \times \mathscr{Y}$. It is rational to introduce a set of unobserved latent variables, $Z = \{Z^1, ..., Z^N\}$, where Z^n represents a set of labels that appear in image x^n but are not annotated in the ground truth. The full set of labels for the image x^n is $\Omega^n = Y^n \cup Z^n$ (note that $Y^n \cap Z^n = \varnothing$). Now the joint feature vector $\Phi(x,\Omega)$ describes the relation among input x, output Y and latent variables Z, and it is defined as

$$\Phi(x,\Omega) = \sum_{y \in Y} x \circ \mathbf{w}_{OVA}^{y} + \sum_{z \in Z} x \circ \mathbf{w}_{OVA}^{z}. \tag{3.5}$$

To train OVA-SSVM, the parameter vector $\mathbf{w}$ is determined by minimizing the regularized risk on the training set $\mathscr{T}$. The risk is measured through a user-provided structured loss function $\Delta(Y,Y^n)$ that quantifies how much the prediction Y differs from the given label set Y^n of image x^n. The resulting convex optimization problem is to minimize an objective function as

$$\begin{gathered} \min_{\mathbf{w},\xi} \frac{\lambda}{2} \|\mathbf{w}\|^2 + \frac{1}{N} \sum_{n=1}^{N} \xi_n \qquad (3.6) \\ s.t \quad \mathbf{w} \cdot \Phi(x^n, \Omega^n) - \mathbf{w} \cdot \Phi(x^n, Y) \geq \Delta(Y, Y^n) - \xi_n, \\ \forall n, Y \subseteq \mathscr{Y}. \end{gathered}$$

The constraints of Equation 3.6 identify the prediction Y with a score $\mathbf{w} \cdot \Phi(x^n, Y)$ that is smaller than the score $\mathbf{w} \cdot \Phi(x^n, \Omega^n)$ of the “full” ground truth Ω^n by a soft margin that equals to the loss $\Delta(Y,Y^n)$ with the slack variable ξ_n. The optimization problem can be solved efficiently using a constraint generation strategy [50]: the constraint can be generated by identifying the most violated (incorrect) prediction $\bar{Y}$ from Y for the current parameter vector

w on x^n. This amounts to solving

$$\bar{Y} = \arg\max_{Y \subseteq \mathscr{Y}} \{\Delta(Y, Y^n) + \mathbf{w} \cdot \Phi(x^n, Y)\}. \tag{3.7}$$

Given the definition of the user-provided structured loss Δ, $\bar{Y}$ of all $x^n \in \mathscr{X}$ can be used to approximate a lower bound of the objective in Equation 3.6. Then the gradient of Equation 3.6 can be computed, and alternately optimize the latent variables Z and the parameter vector **w**. In the next section, the author will introduce the proposed image specific loss term that is elaborately designed for incompletely labeled training data, and derive the corresponding optimization algorithm in the structured output learning framework.

3.3.2 Image specific structured loss

Since the given label set Y^n may not describe all the object in the image x^n, an annotation model should not be penalized for predicting "incorrect" labels that actually describe those objects in x^n. To address this issue, a structured loss function Δ is designed in OVA-SSVM. It argues that, given a set of predicted output labels Y for x^n, it should not penalize the method if one of the predicted labels $y \in Y$ is similar to any of the ground truth labels $y^n \in Y^n$. The loss function is defined as

$$\Delta(Y, Y^n) = \min_{y \in Y} \min_{y^n \in Y^n} d(y, y^n), \tag{3.8}$$

where $d(y, y^n)$ is the error term measuring the difference between label y and y^n. In practice, $d(y, y^n)$ could be a flat error measure: $d(y, y^n) = 0$ if $y = y^n$, and 1 otherwise. Besides, $d(y, y^n)$ could also be a hierarchical error measure that measures the shortest path distance between y^n and y in a taxonomic vocabulary tree.

Actually, there are several limitations of the structured loss of Equation 3.8 for the incomplete setting. Firstly, to predict output labels Y, ensuring that only one of the predicted labels is similar to the ground truth is not enough. In other words, it is expected that each of the predicted labels is similar to any (even all) of the ground truth labels. Secondly, the error measure of $d(y, y^n)$ is either too coarse to quantify the difference of labels (i.e. flat error measure), or too rigorous to require the prior construction of taxonomic tree (i.e. hierarchical error measure). Thirdly, the error measure indicates that the variances of labels are based on the global statistics of training data, whereas for the incomplete setting, it is not sufficient to model the relatedness of missing labels and ground truth labels.

Fig. 3.3 demonstrates two examples of label prediction using the *flat* structured loss in OVA-SSVM. Note that in Fig. 3.3 the fourth column, the image specific loss of each predicted

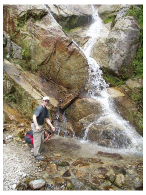
Image x^n	Incomplete labels Y^n	Predicted labels Y		Structured loss Δ	
		OVA-SSVM (Flat)	Proposed	Flat	Image specific
	bloom, leave	bloom, flower, fruit, forest, branch	bloom (0), leave (0), trunk (0.531), flower (0.552), plant (0.765)	0	0.3696
	man, one, rock	man, woman, front, bottle, forest	man (0), rock (0), tee-shirt (0.647), hand(0.685), waterfall (0.689)	0	0.4042

Fig. 3.3 Examples of label prediction using flat loss (OVA-SSVM (Flat)) *vs.* image specific structured loss (proposed method), where the two images are selected from IAPRTC-12 dataset.

label is also presented. The loss values Δ in last two columns are calculated according to Equation 3.8 and Equation 3.9 respectively. It can be observed that, although flat structured loss (in fifth column) is generated to be zero (since the predicted labels “bloom,” “man” match the incomplete ground truth Y^n), the predicted result (in third column) is inferior since it contains several incorrect labels, e.g. {“fruit,” “forest”}, {“woman,” “bottle,” “forest”}. Thus, it implies that numerically minimizing the structured loss of Equation 3.8 could not guarantee all predicted labels to be similar to ground truth labels.

To address the limitations of Equation 3.8, the author assumes that each of the predicted labels is related to *all* of the ground truth labels, and proposes a new structured loss term to capture the variances of labels relying on the specific image content. The proposed image specific loss term is defined as

$$\Delta(Y,Y^n;x^n) = \frac{1}{|Y|}\frac{1}{|Y^n|}\sum_{y\in Y}\sum_{y^n\in Y^n} d(y,y^n;x^n). \tag{3.9}$$

Here the error measure $d(y,y^n;x^n)$ is image specific, representing the difference of label y and y^n particularly on the image x^n. In addition, the structured loss $\Delta(Y,Y^n;x^n)$ ensures that each of the predicted labels in Y to be related to all the ground truth labels in Y^n. Since the incomplete label set Y^n is small, here the structured loss of Equation 3.9 is restricted to moderately consider the dependencies between each of the predicted labels and all labels in Y^n.

Inspired by the works [10, 20], the author casts the calculation of $d(y, y^n; x^n)$ to comparing the relatedness of image x^n to labels y and y^n. In particular, for a given label y_c, let $\mathscr{X}_c^+$ be the set of images that are annotated with label y_c, and the remaining images be $\mathscr{X}_c^- = \mathscr{X} \backslash \mathscr{X}_c^+$. For image x^n in $\mathscr{X}_c^+$, the relatedness of image x^n to label y_c is defined as $R(x^n, y_c) = 1$ since x^n is annotated with y_c. For image x^n that belongs to $\mathscr{X}_c^-$ of y_c, the relatedness score of $R(x^n, y_c)$ is determined by considering three factors: visual similarity, semantic similarity and image-label association in the visual neighborhood. Correspondingly, $R(x^n, y_c)$ consists of

- Visual similarity based relatedness score $R_V(x^n, y_c)$: we compute the visual distance $dist(\cdot)$ (scaled to range $[0, 1]$) of x^n with its nearest neighbor $x^* \in \mathscr{X}_c^+$, and define $R_V(x^n, y_c) = 1 - dist(x^n, x^*)$.
- Semantic similarity based relatedness score $R_S(x^n, y_c)$: we first compute the correlation score between pairwise labels y_i and y_j, $\forall y_i, y_j \in \mathscr{Y}$ as: $co\_occur(y_i, y_j) = \frac{f_{i,j}}{f_i + f_j - f_{i,j}}$, where f_i and f_j are the count of occurrence of labels y_i and y_j, and $f_{i,j}$ is the count of co-occurrence of labels y_i and y_j. Let Y^n be the label set of image x^n and $R_S(x^n, y_c) = \max_{y^n \in Y^n} co\_occur(y_c, y^n)$.
- Reverse nearest neighbors based relatedness score $R_N(x^n, y_c)$: given a fixed value of $M (= 5)$, first the m^{th} nearest neighbor for each image in $\mathscr{X}_c^+$ on the whole training set $\mathscr{X}$ is calculated, and then p_m is set to be the number of images in $\mathscr{X}_c^+$ that have x^n as their m^{th} nearest neighbor. Finally, $R_N(x^n, y_c)$ is defined as $\sum_{m=1}^{M} \frac{p_m}{m} / \sum_{m=1}^{M} p_m + \varepsilon$, where $\varepsilon > 0$ is a small number to avoid division by zero.

Finally, $R(x^n, y_c)$ is defined as the average of these three scores, similar as in [10]:

$$R(x^n, y_c) = \frac{1}{3}(R_V(x^n, y_c) + R_S(x^n, y_c) + R_N(x^n, y_c)). \tag{3.10}$$

Now the error measure $d(y, y^n; x^n)$ can be calculated by comparing the relatedness scores of image x^n to labels y and y^n as

$$d(y, y^n; x^n) = R(x^n, y^n) - R(x^n, y) = 1 - R(x^n, y). \tag{3.11}$$

Recalling that $y^n \in Y^n$ is the ground truth label of x^n, therefore it has highest relatedness score (equals to 1). It can be learnt that the calculation of Equation 3.11 is directly determined by the relatedness score $R(x^n, y)$ of label y to image x^n. If the predicted label y has larger relatedness score to x^n, it would have smaller difference with all the ground truth labels. This is consistent with the proposed structured loss of Equation 3.9, which now can be efficiently measured by the relatedness of predicted labels to the specific image.

When computing the proposed structured loss in Equation 3.11, the complexity of calculating the relatedness scores $R_V(x^n, y_c)$ and $R_S(x^n, y_c)$ is $\mathscr{O}(N^2)$ and $\mathscr{O}(C^2)$, respectively, due to the pairwise neighborhood searching. In addition, calculating $R_N(x^n, y_c)$ also relies on the searched neighborhood, thus $R_N(x^n, y_c)$ can be obtained along with $R_V(x^n, y_c)$. It is worth mention that since usually the number of training samples N is much larger than the label dictionary size C, the time consuming of the Equation 3.11 scales quadratically to N.

Compared with the flat/hierarchical structured loss, the proposed structured loss of Equation 3.9 has several advantages. Firstly, as shown in Fig. 3.3, although the loss values (in last column) are numerically larger than "zero" of flat structured loss (in fifth column), the predicted labels are more similar to the provided incomplete labels. This is because the proposed structured loss moderately considers the predicting labels based on their relatedness to specific image content, and the relatedness measure is elaborately designed and more appropriate than the simple 0-1 measure. Secondly, the proposed structured loss is more flexible to the number of ground truth labels as it accumulatively measures each of the predicted labels to all the ground truth, while the flat structured loss focuses on the most dominant one in the predicted label to a single label of the ground truth. Thirdly, in the proposed method the relatedness measure can be directly and precisely computed from incompletely labeled training images, while prior knowledge of taxonomy or large quantities of fully labeled training data is required in the traditional algorithms.

3.3.3 Optimization method

Given the proposed structured loss function of Equation 3.9, the most violated constraint of prediction $\bar{Y}$ for image x^n can be generated according to Equation 3.7 as in the form of

$$\begin{aligned}\bar{Y} &= \arg\max_{Y\in\mathscr{Y}}\{\frac{1}{|Y|}\frac{1}{|Y^n|}\sum_{y\in Y}\sum_{y^n\in Y^n} d(y, y^n; x^n) + \sum_{y\in Y}\mathbf{w}\cdot\phi(x^n, y)\}\\ &= \arg\max_{Y\in\mathscr{Y}}\{\frac{1}{|Y|}\sum_{y\in Y}(1 - R(x^n, y)) + \sum_{y\in Y}\mathbf{w}\cdot\phi(x^n, y)\},\end{aligned} \tag{3.12}$$

where the calculation of structured loss $\Delta(Y, Y^n; x^n)$ is converted to computing the relatedness scores of predicted label set Y to image x^n, as described in Section 3.3.2.

We can obtain the solution of Y of Equation 3.12 by simply sorting the term $\frac{1}{|Y|}(1 - R(x^n, y_c)) + \mathbf{w}\cdot\phi(x^n, y_c)$ for each label $y_c \in \mathscr{Y}$, and then by choosing the top K labels for $\bar{Y}$.

Solving Equation 3.12 greedily takes $\mathcal{O}(C\log C)$, thus it is faster than the constraints generation method in OVA-SSVM which takes $\mathcal{O}(C^2\log C)$. After the most violated constraint $\bar{Y}$ for each image has been generated, the lower bound of the objective function in Equation 3.6 can be derived as

$$J(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{N}\sum_{n=1}^{N}[\Delta(\bar{Y}-Y^n) + \mathbf{w}\cdot\Phi(x^n,\bar{Y}) - \mathbf{w}\cdot\Phi(x^n,\Omega^n)], \tag{3.13}$$

and the gradient of $J(\mathbf{w})$ with respect to $\mathbf{w}$ is

$$\nabla_{\mathbf{w}}J(\mathbf{w}) = \lambda\mathbf{w} + \frac{1}{N}\sum_{n=1}^{N}[\mathbf{w}\cdot\Phi(x^n,\bar{Y}) - \mathbf{w}\cdot\Phi(x^n,\Omega^n)]\,. \tag{3.14}$$

It can be observed in Equation 3.13 and Equation 3.14 that calculating $J(\mathbf{w})$ and its gradient is involved in computing the joint feature vector $\Phi(x^n,\Omega^n)$ on "full" label set Ω^n of each image. Here the number of labels in Ω^n is fixed to be K so that the missing labels Z^n can be precisely inferred. Thus, $\Phi(x^n,\Omega^n)$ can be efficiently computed according to Equation 3.5 with latent variable Z^n. To learn the parameter vector $\mathbf{w}$ with latent variable Z^n, the previous alternating optimization technique proposed in [38, 39] is adopted. Specifically, the optimization algorithm alternates between optimizing the parameter vector $\mathbf{w}^t$ by initializing the latent variable Z^n for each image in the t^{th} iteration, and re-estimating the latent variable Z^n for the $(t+1)^{th}$ iteration given the learnt parameter vector $\mathbf{w}^t$.

The pseudocode for solving the alternating optimization problem is depicted in Algorithm 1. In practice, the author refers to the implementation in [39] to conduct the optimization algorithm, which utilizes GPU and Nvidia's high-performance BLAS library to accelerate the optimization procedure.

3.4 Experimental results under incomplete labeling

In this section, the effectiveness of the proposed method is evaluated by comparing it with the previous OVA-SSVM and other state-of-the-art annotation methods under incomplete setting.

Algorithm 1: Alternating optimization of proposed method

```
Input: Incompletely labeled training data $\mathscr{T} = \{(x^n, Y^n)\}_{n=1}^{N}$, pre-trained binary classifiers
    $\{\mathbf{w}_{OVA}^{y_c}\}_{c=1}^{C}$
Output: Parameter vector $\mathbf{w}$
Initialize $\mathbf{w}_0 = \mathbf{1}$ for iteration $t = 0$
repeat
    Set $t = t + 1$
    for $n = 1, ..., N$ do
        Assign latent variable $Z_t^n = \{arg\max_{Y \in \mathscr{Y}} \mathbf{w}_{t-1} \cdot \Phi(x^n, Y)\} \backslash Y^n$ for $x^n$ (preserving
        $K - |Y^n|$ missing labels)
    end for
    for $n = 1, ..., N$ do
        Generate the most violated constraint $\bar{Y}_t$ for $x^n$ according to Equation 3.12
    end for
    Compute objective $J_t(\mathbf{w})$ and gradient $\nabla_{\mathbf{w}} J_t(\mathbf{w})$ according to Equation 3.13 and
    Equation 3.14
    Minimize loss of Equation 3.6 to calculate $\mathbf{w}_t$
until Loss in Equation 3.6 is converged
```

3.4.1 Experimental setup

Datasets and Features

Our evaluation experiments are conducted on two publicly available benchmark datasets: IAPRTC-12 [51] and NUS-WIDE [52]. These two datasets are very challenging with significant diversity among the images that are obtained from the social web. Table 3.1 shows the general statistics of these two datasets. The items in the second row are listed in the format "training/test," and items in the third and fourth rows are given in the format "mean/minimum/maximum." It is worth noting that they cover both conditions of large vocabulary size and large number of images.

In the experiments, for IAPRTC-12 dataset, the same multiple features is used as those in [10, 12, 13, 35]. These multiple features consist of global and local features. The global features include histograms in RGB, HSV and LAB color space, and the GIST features; and the local features include the SIFT and hue descriptors obtained densely from multi-scale grid, and from Harris-Laplacian interest points. For NUS-WIDE dataset, besides global GIST features, other five types of SIFT based local features (C-SIFT, Opponent-SIFT, RGB-SIFT, RG-SIFT) are also extracted using the public colorDescriptor tools [53]. The SIFT based features are computed without orientation invariance and the grid has a step size of three. The codebook for each SIFT based feature is generated from 7,000 randomly selected images, and quantized to 4,000 corresponding k-means clusters. For both datasets, L2 normalization for each type of features is firstly performed, and then concatenated to an fused feature vector (37,152-dimensions for IAPRTC-12 and 20,512-dimensions for NUS-WIDE) to represent each image.

Table 3.1 General statistics for the two datasets used for evaluation.

	IAPRTC-12	NUS-WIDE
Number of labels	291	81
Number of images	17,665/1,962	138,563/92,484
Number of labels per image	5.7/1/23	1.8/1/20
Number of images per label	34/153/4,999	2,512/333/16,425

Incomplete setting

We consider the original IAPRTC-12 as fully labeled dataset since the average number of labels per image is more than 5 (5.7 in Table 3.1), which could be sufficient to describe multiple objects in an image. To simulate the incomplete setting, partial labels for each image are randomly deleted, and the deletion process which stands by the principle $\min(1, \lceil M \times (1 - ratio) \rceil)$ ensures that each image preserve at least one label. Here, M denotes the number of original labels of an image, $\lceil \cdot \rceil$ denotes the ceiling function which gives the smallest integer not smaller than the given value, and $ratio$ represents the degree of incompleteness. In the experiments, $ratio = \{10\%, 30\%, 50\%, 70\%, 90\%\}$, and it indicates that the larger the ratio is, the higher the degree of incompleteness would be. Note that $ratio = 0\%$ means the dataset is fully labeled without any incompletion. For NUS-WIDE, as the average number of labels per image is less than 2 (1.8 in Table 3.1), which could be insufficient compared with the situation of IAPRTC-12, thus the NUS-WIDE is considered as an originally incompletely labeled dataset.

Binary classifiers

As the proposed method needs a pre-trained binary classifier for each class as a starting point for structured learning, following the previous works [10, 39], the OVA-SVM classifiers are learned for initialization. In particular, a linear OVA-SVM classifier is trained for each label using Pegasos algorithm [54], and the raw confidence scores predicted by the OVA-SVM classifiers are calibrated to probabilities using Platt algorithm [55]. Finally, the calibrated probability scores are utilized as the initial input to the proposed method.

Evaluation metrics

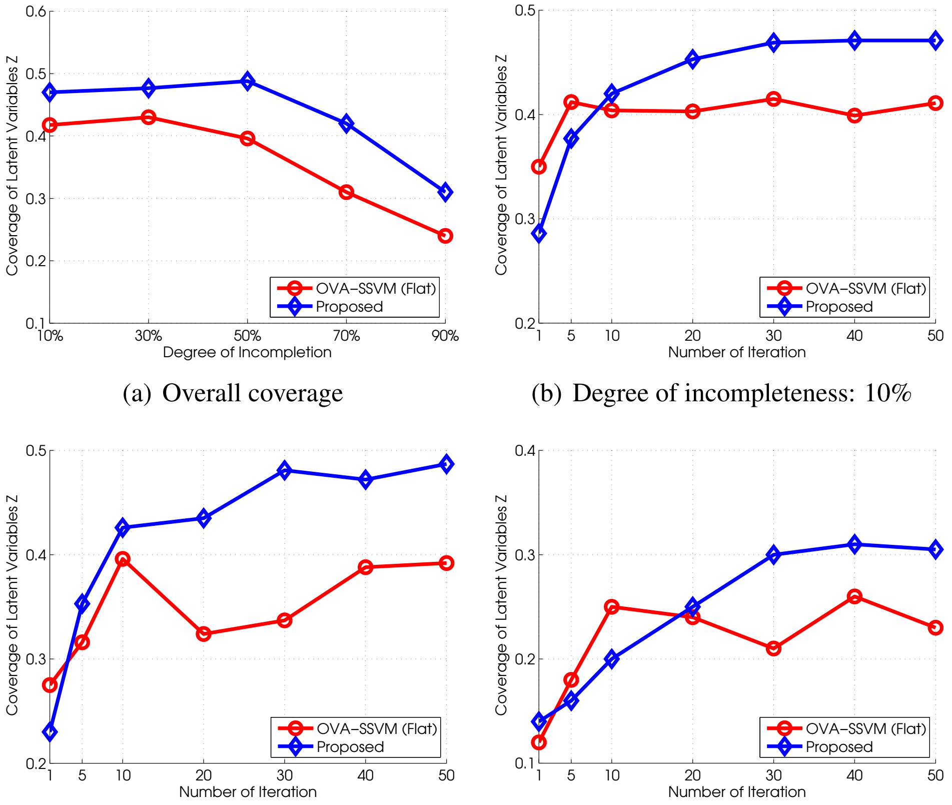
Given an unlabeled test image, the score for each label is firstly computed using the learnt model, and then five top-scored ($K = 5$, $|Y| = 5$) labels are selected according to Equation 3.4. Here two standard criteria are used to evaluate the performance: (1) average precision per label P, (2) average recall per label R. Note that the P and R scores are obtained by first computing precision and recall for each label and then averaging. In addition, as the number of labels in NUS-WIDE dataset is considerably small, another two criteria: *Hamming loss* and *Average AUC* are exported, which take the performance of overall prediction and ranking into account. For all the adopted evaluation metrics except *Hamming loss*, a larger numerical value indicates better performance. All the experiments are conducted on a computer which has Inter Xeon E5520 2.27GHz CUP with 16GB RAM.

3.4.2 Evaluation on IAPRTC-12 dataset

Assessment of assigning latent variables

The author then evaluate the proposed method on image annotation task in this section and the next section. Due to the more manageable size of the two datasets IAPRTC-12 and NUS-WIDE, the proposed image-specific structured loss for each training image can be directly computed in these two datasets. Since these two datasets have no pre-defined semantic hierarchy, thus the hierarchical structured loss is invalid for evaluation. The author first investigates the the efficiency of structured output learning with latent variables in the proposed method and OVA-SSVM (Flat) method. Intuitively, the latent variables should capture those objects that appear in an image, but are not present in the ground truth. Specifically, the author explores how closely the assigned latent variable Z^n matches those labels $\Omega^n \backslash Y^n$ deleted from the originally full annotations of image x^n when the learning algorithm depicted in Algorithm 1 is followed. The author uses a measure termed $Coverage = \frac{1}{N}\sum_{n=1}^{N}\frac{|Z^n \cap (\Omega^n \backslash Y^n)|}{|Z^n|}$ to represent the averaged intersection between Z^n and $\Omega^n \backslash Y^n$ for all training image $x^n \in \mathscr{X}$. Note that higher coverage indicates better assignments of latent variables.

Fig. 3.4 (a) shows the overall coverage of latent variable to the deleted labels in the full annotations with different degree of incompleteness. It can be observed that (1) the coverage of latent variable of both methods increases when the degree of incompleteness decreases, and this is reasonable because the more labels are given for training, the better the model can predict the missing labels; (2) the proposed method consistently obtains higher coverage for missing labels than OVA-SSVM (Flat), which simply uses flat structured loss, as the



(a) Overall coverage

(b) Degree of incompleteness: 10%

(c) Degree of incompleteness: 50%

(d) Degree of incompleteness: 90%

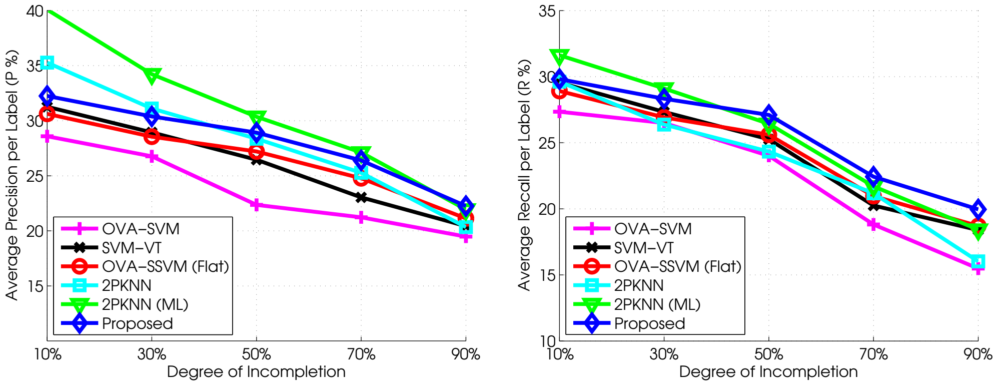
Fig. 3.4 Evaluation of coverage of latent variables with various degree of incompleteness.

image specific structured loss used in the proposed method is more efficient to exploit various contextual information of labels and images under the incomplete setting.

Furthermore, Fig. 3.4 (b)-(d) explicitly demonstrates the changing of coverage of the latent variables through the iterations (as described in Algorithm 1) under different degree of incompleteness: 10%, 50%, 90%. We can learn that the proposed image specific structured loss is appropriate to ensure the proposed method to perform robustly, while OVA-SSVM (Flat) seems to be unstable through the iterations and results in inferior coverage. Especially, when the degree of incompleteness is pretty high (e.g. 70%, 90%), the coverage of the proposed method is significantly better than OVA-SSVM (Flat), showing the superiority of the proposed method under the incomplete setting.

Overall comparison with various degree of incompleteness

To make comprehensive comparison, the labeling results from binary classifiers: OVA-SVM and SVM-VT [10] (OVA-SVM combined with the proposed image specific label relatedness



(a) Comparison of P

(b) Comparison of R

Fig. 3.5 Comparison of annotation performance with various degree of incompleteness.

as depicted in Section 3.3.2 are included, without structured output learning), then the binary classifiers is boosted by structured output learning via flat structured loss (OVA-SSVM (Flat)) and image specific structured loss (the proposed method). In addition, the method termed 2PKNN [14] is also into account, which is a nearest neighbor based method that achieves the state-of-the-art performance for image annotation task with full labeling. The basic 2PKNN method and its metric learning version 2PKNN (ML) under the same settings of incompleteness are evaluated in the experiment.

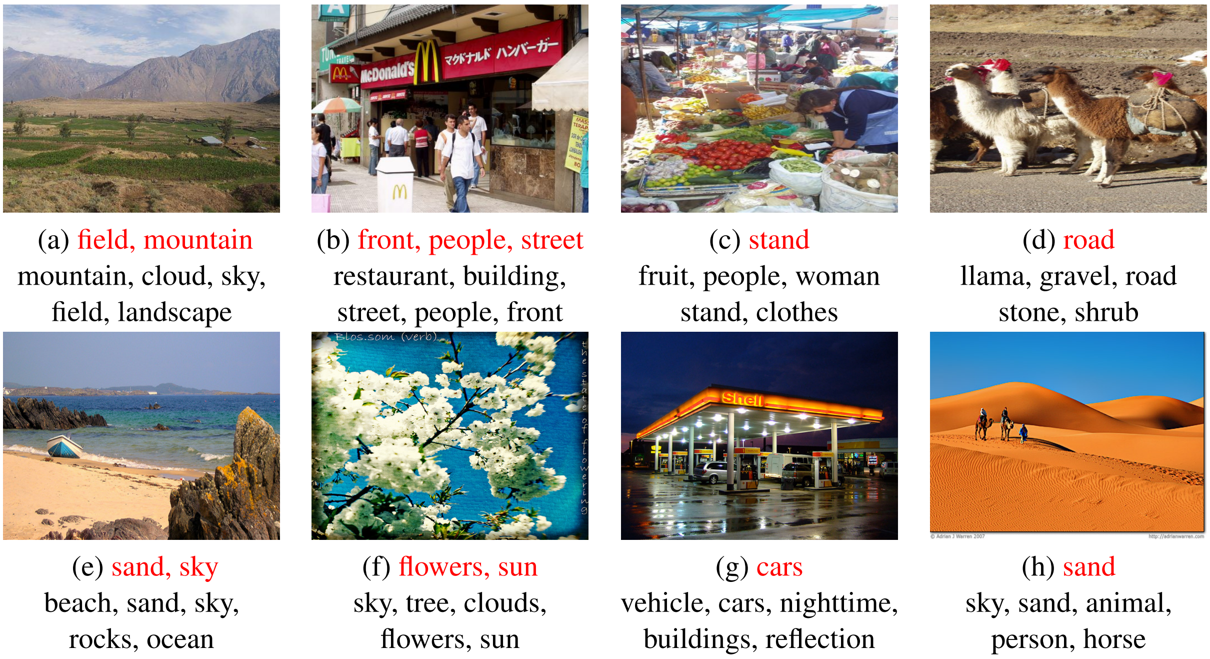
Fig. 3.5 shows the annotation performance of these methods in terms of P and R with various degree of incompleteness. It can be learned that:

- As the degree of incompleteness decreases, the performance of all the methods becomes better. This is because more labels for training data provides more information.

- The proposed method can boost the performance of binary classifiers OVA-SVM and SVM-VT under the incomplete setting, which verifies the efficiency of the incrementally structured output learning.

- Regarding the structured output learning stage, the proposed method performs remarkably better than OVA-SSVM (Flat) which uses the flat structured loss, especially when the degree of incompleteness is considerably high (e.g. 70%, 90%). The reason behind this is that more appropriate structured loss which efficiently accounts for the dependencies among the predicted labels is used under the incomplete setting.

- The 2PKNN method performs better than the proposed method when the degree of incompleteness is not high (e.g. 10%, 30%, 50%), as the elaborately designed neighbor selection scheme in 2PKNN method can efficiently construct a balanced neighborhood in this case, and the metric learning version can further capture the importance of different labels based on measured image similarity. However, when the degree of incompleteness is considerably high (e.g. 70%, 90%), the 2PKNN method performs inferiorly to the proposed method. We believe that large proportions of missing labels from training images prohibit the 2PKNN method obtaining a balanced neighborhood and learning appropriate distance metric for similarity measure. Therefore, it indicates that the proposed method is more robust to tackle the issue of incompleteness.



(a) field, mountain
mountain, cloud, sky,
field, landscape

(b) front, people, street
restaurant, building,
street, people, front

(c) stand
fruit, people, woman
stand, clothes

(d) road
llama, gravel, road
stone, shrub

(e) sand, sky
beach, sand, sky,
rocks, ocean

(f) flowers, sun
sky, tree, clouds,
flowers, sun

(g) cars
vehicle, cars, nighttime,
buildings, reflection

(h) sand
sky, sand, animal,
person, horse

Fig. 3.6 Samples of annotation results of the proposed method on IAPRTC-12 ((a)-(d)) and NUS-WIDE ((e)-(h)), where the red labels are the ground truth and black ones are top five labels predicted by the proposed method.

3.4.3 Evaluation on NUS-WIDE dataset

In the previous section, the author evaluates the performance of the proposed method on a moderate dataset IAPRTC-12 by simulating the cases of different degrees of incompletion. In this section, the author directly utilizes a large scale dataset NUS-WIDE that is originally incompletely labeled dataset for evaluation. To make comprehensive comparisons, besides the four methods compared above, the state-of-the-art annotation methods which are designed for full labeling and incomplete labeling are also considered. The methods for full labeling include JEC [12], Tagprop [13], 2PKNN [14] and M3L [45]. In particular, for Tagprop and 2PKNN, their metric learning version is used, denoting as Tagprop (ML) and 2PKNN (ML).

The methods for incomplete labeling consist of SVM-VT [10], MLR-GL [34], Fasttag [35] and LEML [36]. To make fair comparison, the code provided by these authors are directly used in the experiment and the instructions in the corresponding papers are referred to tune model parameters.

Table 3.2 Annotation performance comparison among different methods on NUS-WIDE dataset, where the best results of previous and the proposed methods are highlighted in bold.

Method	P (%)	R (%)	Average AUC	Hamming loss
JEC [12]	11.9	16.6	0.557	0.083
Tagprop (ML) [13]	13.2	23.8	0.707	0.074
2PKNN (ML) [14]	14.5	23.9	0.726	0.073
OVA-SVM	12.3	22.8	0.782	0.079
M3L [45]	16.1	23.2	0.791	0.071
SVM-VT [10]	16.7	24.3	0.806	0.069
MLR-GL [34]	14.2	23.5	0.722	0.078
Fasttag [35]	**18.4**	21.3	**0.834**	0.067
LEML [36]	17.5	24.6	0.798	0.076
OVA-SSVM (Flat) [39]	16.9	24.1	0.772	0.070
Proposed	17.7	**25.6**	0.819	**0.064**

Table 3.2 shows the annotation performance of various methods depicted above. It can be observed that:

- The proposed method consistently boosts the binary SVM classifiers (OVA-SVM and SVM-VT) and also obtain better performance than OVA-SSVM (Flat).

- The conventional nearest neighbor based methods such as JEC, Tagprop and 2PKNN fail to achieve comparable performance. This reason may be that the noisy and missing label assignments of NUS-WIDE dataset affects the results of selecting representative neighborhood and learning appropriate distance metric.

- The annotation methods including the proposed method designed for incomplete labeling are generally superior to conventional annotation methods with full labeling, which again addresses the significance of tackling the issue of incompleteness of practical annotation data.

- The proposed method performs comparable or better than even the recently proposed methods with incomplete labeling, which corroborates the efficiency of structured output learning on capturing the semantic correlations of labels when labels are incomplete. And it explores more efficient SVM based method combining with structured output learning and provides another aspect for the image annotation problem with incomplete labeling.

Fig. 3.6 gives qualitative samples of annotation results predicted by the proposed method on the two datasets. In particular, for IAPRTC-12 dataset, the original training images are preserved without the deletion process to evaluate the generalization of the proposed method. From the samples it can be learned that, although there are fewer ground truth labels, the proposed method can still make correct prediction to them. In addition, the proposed method can also reflect semantic connectedness among the predicted labels, e.g. {“field,” “landscape”}, {“restaurant,” “building,” “street”}, {“beach,” “sand,” “ocean”}, etc. This further demonstrates the effectiveness of the proposed method using structured output learning.

3.5 Summary

In this chapter, the author considers a special case of weak labeling, which is termed incomplete labeling. The author leverages the structured output learning method to boost the performance of conventional OVA-SVM classifiers, and then formulates an image specific structured loss function which is more appropriate to explore the dependencies of predicted multiple labels. Later, an efficient optimization algorithm with lower computational complexity is developed to learn model parameters. Experimental evaluation demonstrates that the proposed method is efficient to handle the issue of incomplete labeling and performs superior than several existing methods on image annotation task. For the future direction, it is rational to explore more efficient strategy of calculating the image specific structured loss term for large quantities of training data, and to extend the proposed method to a more general case where even some of the incomplete labels are incorrectly assigned to the training samples. This in turn would facilitate the annotation model to be robust against the defection of training data.

Chapter 4

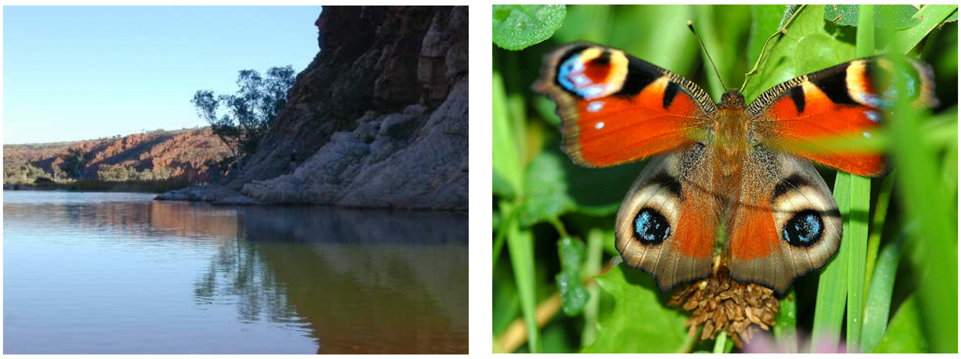
Image-tag re-weighting with defective tag assignments

4.1 Introduction

In Chapter 3, the author discusses a practical issue of *incomplete labelling* where the images in the training set are not completely annotated with all relevant labels. Later in Chapter 3, the author proposes a scheme to explore the image-label associations in the incompletely labeled dataset, and formulate a image-specific structured loss term in the structured output learning method OVA-SSVM. The proposed scheme of modeling the image-label association is efficient and robust when the training dataset is incompletely labeled.

In this chapter, the author posts a more general problem of *defective tag assignments*, where the raw user-provided ground truth tags (labels) assigned to the training images are incomplete or noisy. It is notable that the issue of *incomplete labelling* can be considered as a subcase of *defective tag assignments*. As an example, in the case of *defective tag assignments*, an image whose true tag assignment is (t_1, t_2, t_3) is defectively assigned as (t_1) or (t_2, t_4). Fig. 4.1 shows examples of defectively tagged images from benchmark datasets IAPRTC-12 [51] and MIR Flickr [56]. In Fig. 4.1, italic tags are provided ground truth tags, underlined tags are incorrect noisy tags, and bold tags are potentially missing tags.

Indeed, there are several reasons that cause defective tag assignments. (1) Tag-synonyms: the synonyms of tags can be used interchangeably, e.g. "*lake*" and "*pond*," but users usually will not spend time to enter all the synonyms. (2) Tag-hypernyms: some tags share hierarchical and structural overlap, e.g. "*animal*," "*insect*," "*wildlife*," and users tend to choose general tags. (3) User-unconsciousness: some "obvious" or visually secondary tags might be ignored by user unconsciously, e.g "*tree*," "*plant*." (4) User-misconception:



(a) *cliff, lake, landscape, shade, sky,* ***pond****,* ***rock****,* ***tree****,* ***water***

(b) *nature, macro, flowers, germany, animal, eyes, insect,* ***plant****,* ***leaf****,* ***wildlife***

Fig. 4.1 Example images with defective tag assignments: (a) an image from IAPRTC-12 dataset, (b) an image from MIR Flickr dataset.

visually ambiguous content leads to inappropriate tagging, e.g. "*eyes*." (5) User-arbitrariness, sometimes users choose unrelated tags arbitrarily in order to minimize their efforts, e.g. "*germany*."

To tackle the challenge of defection, the goal of the work in this chapter is to learn a tag completion model from defectively tagged training images to recover the complete tag assignments of training images, and then to apply the learned model on untagged or defectively tagged test images. In other words, the work in this chapter can be treated as a unified framework that tackles both tag completion and denoising. Although several previous studies [57–59] are also declared to be unified frameworks, an inevitable disadvantage of them is the difficulty in balancing the degree of completion and denoising. When recalling missing tags (completion) or rejecting noisy tags (denoising) from an image, the penalties of different tags are equally allocated, leading the result to be biased on either completion or denoising. This potential disadvantage motivates the author to consider specific image-tag penalty term when assigning each tag to each image. Imaging the examples in Fig. 4.1, the cost of recalling tag "*pond*" ought to be different from tag "*car*," as well the penalty of rejecting "*germany*" would vary from "*animal*."

The main contributions of the work in this chapter are summarized as follows:

- An efficient re-weighting scheme is proposed to elaborately assign the penalty of predicting each tag to each image, and the penalty of pairwise image-tag is fully derived from both image-image similarities and tag-tag associations.

- A unified re-weighted empirical loss term is formulated and then is incorporated in two recently proposed tag completion models to tackle the tag completion problem with defective tag assignments.
- Efficient optimization algorithms are developed for the two extended tag completion models.

4.2 Related work

Given the training data with defective tag assignments, one way to complete the missing tags in the test data is to directly apply automatic image annotation techniques such as [10, 12, 13]. Most of these techniques treat image annotation problem as a set of binary classification problem, one for each tag. The main shortcoming of these approaches is that they require training images to be clean without missing or noisy tags. However, the condition of defective tag assignments would potentially result in biased annotation result, hence it is suboptimal to directly utilize these image annotation methods.

Besides these conventional image annotation approaches, several advanced machine learning methods have been proposed for the issues of partially tagged images, tag refinement. For the issue of partially tagged images, the ongoing researches mainly extend conventional multi-label ranking [34], ridge regression [35] or binary SVM classification [10]. These approaches exploits the potential associations between the assigned and unassigned tags.

Regarding tag refinement, several approaches have been proposed to remove noisy tags from correct tags using tag relevance [60, 61], denoising sparse error [59], collaborative propagation among tags [58], etc. For examples, Li *et al.* [60] utilize neighbour voting to learn the relevance of each tag, and then differentiated noisy tags from correct ones; Zhu *et al.* proposed a method to decompose the initial tag matrix into a low-rank refined matrix and a sparse error matrix with low-rank and sparsity constraints. A potential problem of the tag refinement approaches is that they focus more on tag denoising but less on tag completion. Indeed, it is hard to design a unified framework to perform both tag denoising and tag completion, since the difficulty in controlling the degree of denoising may affect the performance of tag completion. Therefore, recently studies such as [57, 62] discuss the tag completion problem and utilize matrix factorization (reconstruction) methodologies by embedding various contextual information such as semantic and visual similarities. Specially, Wu *et al.* [57] use matrix factorization approach to find the optimal tag matrix, which is consistent with both observed tags and visual similarities. Lin *et al.* [62] try to reconstruct the optimal tag matrix from both image-specific and tag-specific sides, with the constraint that an image contains a few objects and a tag represents a few levels of meaning, and usually objects or levels of meaning are redundantly contained or implied in the context. These methods specialized for the tag completion problem usually show better performance than the tag denoising approaches.

4.3 Image-tag re-weighting

The author first describes the proposed re-weighted empirical loss function and then present an efficient scheme to determine the image-tag weights combining both visual and semantic cues.

4.3.1 Re-weighted empirical loss

Let $\mathscr{X} = \{x_1,...,x_n\}$ be a collection of n images and $\mathscr{Y} = \{t_1,...,t_m\}$ be a vocabulary of m tags. The user-annotated dataset $\hat{T}_{n\times m} = \{\hat{Y}_1,...,\hat{Y}_n\}$ is a defective tag matrix, where $\hat{Y}_i \subseteq \mathscr{Y}$ is an initial defective tag vector of image x_i, and $\hat{Y}_{i,j}$ is set to one if tag j is assigned to image i and zero otherwise. Our target is to recover the final complete tag matrix $T_{n\times m}$ from initial defective $\hat{T}$, where the element $T_{i,j} \geq 0$ denotes confidence score of assigning tag t_j to image x_i. The general formulation of previous tag completion works [35, 57–59] corresponds to an empirical loss minimization with a regularization term as follows:

$$\min_{T} \lambda E_p(T) + E_r(T), \tag{4.1}$$

where the regularization term $E_p(T)$ restricts the model to some constraints, such as the visual and semantic consistencies, and $\lambda \geq 0$ controls the trade-off between the empirical loss and the regularization term. Specifically, the empirical loss term $E_r(T)$ represents the deviation of complete tag matrix T from initial defective tag matrix $\hat{T}$, and is generally formulated as

$$E_r(T) = \|T - \alpha\hat{T}\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}(Y_{i,j} - \alpha_{i(j)}\hat{Y}_{i,j})^2. \tag{4.2}$$

Here $\|\cdot\|_F$ denotes Frobenius matrix norm, α is an $n \times n$ diagonal matrix where α_i plays as scaling factor for image x_i. Please note α could also be an $m \times m$ diagonal matrix, where α_j is a scaling factor for tag t_j. With respect to the selection of α, previous tag completion works can be categorized into two types. 1) **Identical-type**, which assumes α to be an identical diagonal matrix. This type indicates that each element in $\hat{T}$ has the equal scaling value. 2) **Scaling-type**, which assumes α to be a diagonal matrix with different diagonal values. This type implies that each row (column) in $\hat{T}$ has equal scaling value of α_i (α_j).

It is indicated from Equation 4.2 that: 1) the complete tag matrix T should be consistent (similar) to the initial defective matrix $\hat{T}$, 2) the penalty of assigning each tag t_j to each image x_i contributes equally to empirical loss $E_r(T)$. However, these two assumptions may be inappropriate to obtain the optimal T under defective setting. Firstly, given the defective tag matrix $\hat{T}$, the constraint of consistency between $\hat{T}$ and T is too rigorous to twist the complete T from defective $\hat{T}$, which leads the final T to be similar with original $\hat{T}$ (see more interpretation in Fig. 4.2). Secondly, the identical penalty assumption of assigning each tag t_j to each image x_i is imprecise and inhibits the model from recalling missing tags and rejecting noisy tags from $\hat{T}$.

To formulate a more appropriate empirical loss $E_r(T)$ under defective tag assignments, a weighted image-tag penalty matrix $Z_{n\times m}$ is introduced, whose element $Z_{i,j}$ represents the penalty of assigning tag j to image i in T. Then, the re-weighted empirical loss can be derived as:

$$E_r(T) = \|(T-\alpha\hat{T})\circ Z\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}(Y_{i,j}-\alpha_{i(j)}\hat{Y}_{i,j})^2 Z_{i,j}. \tag{4.3}$$

Here, $\circ$ represents point-wise product of matrices. It is worth noting that the specific image-tag penalty $Z_{i,j}$ in Z is determined by corresponding image-tag relatedness. Specifically, for tag t_j that is related but missing in image x_i, the penalty $Z_{i,j}$ needs to be smaller so that higher confidence score can be allocated to recall t_j in x_i. Likewise, for tag t_j that is noisy but tagged in image x_i, the penalty $Z_{i,j}$ ought to be smaller so that lower confidence score can be allocated to reject t_j in x_i.

Compared with traditional empirical loss in Equation 4.2, the re-weighted empirical loss in Equation 4.3 indeed have two advantages:

- It allows the complete tag matrix T to be moderately inconsistent to the initial defective matrix $\hat{T}$, and it is possible to recall untagged related tags which are missing in $\hat{T}$ and to remove tagged noisy tags from $\hat{T}$, since the weighted penalty Z could balance the empirical loss.

- Each element $Z_{i,j}$ in Z is flexible to explicitly incorporate both visual and semantic cues to model the specific image-tag relatedness, while the identical penalty in Equation 4.2 would neglect the discrimination of pairwise image-tag and limitedly formulate a suboptimal empirical loss term.

Fig. 4.2 illustrates the tag completion results with/without image-tag re-weighting in identical-type model (α is identical diagonal matrix). In Fig. 4.2, $\hat{T}$ is initial defective tags for training, T is final tag completion result. Tags in T is in the format of "tag ($Z_{i,j}$)(confidence

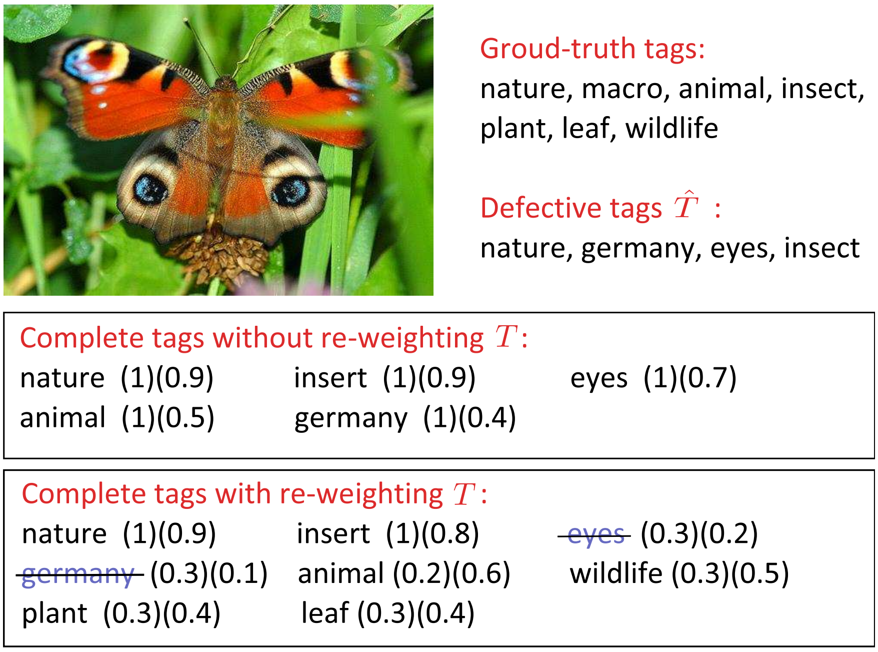


Fig. 4.2 Tag completion example with/without image-tag re-weighting.

score).” The empirical loss without re-weighting (Equation 4.2) is: $0.1^2+0.1^2+0.3^2+0.5^2+0.6^2=0.72$, with re-weighting (Equation 4.3) is: $0.1^2+0.2^2+0.8^2\times0.3+0.9^2\times0.3+0.6^2\times0.2+0.5^2\times0.3+0.4^2\times0.3+0.3^2\times0.4=0.716$. Note that the rejected noisy tags “*eyes*,” “*germany*” also contribute to the re-weighted loss in Equation 4.3. With similar empirical loss values (0.72 *vs.* 0.716), the recovered tag matrix T without re-weighting (top box in Fig. 4.2) twists initial $\hat{T}$ slightly, since using identical penalty it is hard to decline the confidence scores of noisy tags (e.g. “*eyes*,” “*germany*”) and to promote missing tags (e.g. “*plant*,” “*leaf*”) while keeping low empirical loss. However, after importing the re-weighted image-tag penalties, T with re-weighting (bottom box in Fig. 4.2) is able to assign appropriate confidence scores relying on image-tag relatedness while maintaining low empirical loss values.

4.3.2 Determining image-tag penalty

The author proposes an efficient approach to determine $Z_{i,j}$ for pairwise image-tag entry in $\hat{T}$ incorporating both visual and semantic similarities. For a given tag $t_j, \forall j \in \{1,...,m\}$, let $\mathscr{X}_{t_j}^+$ be a set of images that are annotated with tag t_j, and the remaining samples as $\mathscr{X}_{t_j}^- = \mathscr{X} \setminus \mathscr{X}_{t_j}$, and $Z_{i,j}$ for $x_i \in \mathscr{X}_{t_j}^+$ and $x_i \in \mathscr{X}_{t_j}^-, i \in \{1,...,n\}$ are considered separately.

- Visual similarity based relatedness $R_V(\cdot)$. The visual distance score $dist(\cdot)$ is computed (scaled to range $[0,1]$) of x_i with all images $x_k \in \mathscr{X}_{t_j}$. Specially, for $x_i \in$

$\mathscr{X}_{t_j}^+$, $R_V(x_i,t_j) = 1 - \min dist(x_i,x_k), x_k \in \mathscr{X}_j^+ \backslash x_k$, for $x_i \in \mathscr{X}_{t_j}^-$, $R_V(x_i,t_j) = 1 - \min dist(x_i,x_k), x_k \in \mathscr{X}_{t_j}^+$.

- Semantic similarity based relatedness $R_S(\cdot)$. Firstly, computing the correlation score between pairwise tags t_i and t_j as: $co\_occur(t_i,t_j) = \frac{f_{i,j}}{f_i+f_j-f_{i,j}}$, where f_i and f_j are the occurrence of tags t_i and t_j, and $f_{i,j}$ is the co-occurrence of tags t_i and t_j. Let $\hat{Y}_i$ be the tag set of image x_i and $R_S(x_i,t_j) = \max_{t_k \in \hat{Y}_i} co\_occur(t_j,t_k)$.

Based on the two similarity measures above, the relatedness $R(x_i,t_j)$ is then defined as a weighted linear combination of the two similarity components above:

$$R(x_i,t_j) = \pi R_V(x_i,t_j) + (1-\pi) R_S(x_i,t_j), \tag{4.4}$$

where $R_V(\cdot)$ and $R_S(\cdot)$ are normalized similarity scores from image-specific and tag-specific measures, and π is a weight parameter in $[0,1]$ which is validated during parameter tuning. In practice, the author also use the relatedness term (Equation 3.10) defined in Chapter 3 to calculate $R(x_i,t_j)$, which can be considered as a more general form to account for the relatedness of images and tags. Therefore, for each tag t_j, the image-tag penalty $Z_{i,j}$ for the i-th image based on $R(x_i,t_j)$ is computed as:

$$Z_{i,j} = \begin{cases} R(x_i,t_j) & \text{if } x_i \in \mathscr{X}_{t_j}^+ \\ \frac{1}{1+\exp(R(x_i,t_j))} & \text{if } x_i \in \mathscr{X}_{t_j}^- \end{cases} \tag{4.5}$$

From Equation 4.5 it can be learned that missing and noisy tags have smaller penalties $Z_{i,j}$ than other tags depending on their relatedness value $R(x_i,t_j)$. Thus these tags are more likely to be predicted with higher confidence scores, and they are able to be recalled (missing tags) or rejected (noisy tags) in recovered complete tag matrix T. Similar as the discussion in Chapter 3, Z can be calculated off-line in practice from the initial defective tag matrix $\hat{T}$ and treated as the input of the proposed methods during optimization.

4.4 Tag completion via image-tag re-weighting

In this section, two typical tag completion models are selected: tag matrix completion (TMC) [57] from identical-type models and Fasttag [35] from scaling-type models. Then the proposed image-tag re-weighting scheme is embedded and the parameter optimization algorithms are derived for these two types of models.

4.4.1 TMC with image-tag re-weighting

Given visual features of image data $V_{n\times d}$, tag correlation matrix $R_{m\times m}$ and observed defective tag matrix $\hat{T}_{n\times m}$, the TMC model reconstructs the complete tag matrix $T_{n\times m}$ under three constraints:

1. The complete tag matrix T should be similar to the observed defective matrix $\hat{T}$, and the empirical loss term is formulated as $\|T-\hat{T}\|_F^2$.

2. The complete tag matrix T should be consistent with the visual similarity, and this constraint is formulated as $\|TT^\top - Vdiag(w)V^\top\|_F^2$, where $w_{d\times 1}$ is a non-negative vector representing the weight of each visual feature dimension.

3. The complete matrix T should also be consistent with tag correlation matrix R, which can be formulated as $\|TT^\top - R\|_F^2$. It is indicated that TMC is an identical-type model, as α is set to be an identical diagonal matrix in the first constraint.

When embedding the proposed re-weighting scheme in TMC model, the first constraint $\|T-\hat{T}\|_F^2$ is replaced with the re-weighted empirical loss term in Equation 4.3. Thus, the objective of the TMC model with image-tag re-weighting (TMC-RW) is to minimize the loss function $\mathscr{L}(T,w)$, which is derived as,

$$\begin{aligned}\min_{T,w}\mathscr{L}(T,w) = \min_{T,w} & \underbrace{\|(T-\hat{T})\circ Z\|_F^2}_{re-weighted\ loss} + \eta\|TT^\top - R\|_F^2 \\ + \lambda\|TT^\top - & Vdiag(w)V^\top\|_F^2 + \mu\|T\|_1 + \gamma\|w\|_1, \end{aligned} \tag{4.6}$$

where the $L1$ regularizer for both T and w are to generate sparse solutions for them. The subgradient descent based algorithm proposed in [57] is adopted to optimize T and w with non-negative constraints. Defining $\mathscr{L}(T,w) = \mathscr{A}(T,w) + \mu\|T\|_1 + \gamma\|w\|_1$, $G = TT^\top - Vdiag(w)V^\top$, $H = TT^\top - R$, the subgradients for $\mathscr{A}(T,w)$ are derived as

$$\begin{aligned}\bigtriangledown_T\mathscr{A}(T,w) &= 2\lambda GT + 2\eta TH + 2(T-\hat{T})\circ Z, \\ \bigtriangledown_w\mathscr{A}(T,w) &= -2\lambda diag(V^\top GV).\end{aligned} \tag{4.7}$$

The composite function optimization is used similar as in [57], and the final update rules for T and w in each iteration are derived as following:

$$\begin{aligned}T \leftarrow & [T - \zeta_t \bigtriangledown_T \mathscr{A}(T,w) - \mu\zeta_t 1_{n\times m}]_+, \\ w \leftarrow & [w - \zeta_t \bigtriangledown_w \mathscr{A}(T,w) - \gamma\zeta_t 1_{d\times 1}]_+,\end{aligned} \tag{4.8}$$

where ζ_t is the step size and $[f]_+ = \max(0, f)$ projects the negative entries in T and w to be zero in iteration t.

4.4.2 Fasttag with image-tag re-weighting

To keep the notations consistent with Fasttag, the visual features of images data is denoted as $V_{d\times n} = \{x_1, ..., x_n\}$ and observed defective tag matrix is denoted as $\hat{T}_{m\times n} = \{\hat{Y}_1, ..., \hat{Y}_n\}$ for each image in each *column*. Fasttag assumes that the observed defective tag matrix $\hat{T}$ is a "corrupted" version of the complete tag matrix $T_{m\times n}$, and introduces an enrichment mapping matrix $B_{m\times m}$ to reconstruct T from $B\hat{T}$ by estimating which tags are likely to be recalled or rejected in $\hat{T}$. The original empirical loss is formulated as $||T - B\hat{T}||_F^2$, which implies that Fasttag is a scaling-type model as each image has identical value of α. Moreover, besides the enriched mapping to reconstruct complete tag matrix T, T can also be reconstructed from the image features V. Then a co-regularizer term is introduced to force the two reconstructions from both tag-side and image-side with small difference $||B\hat{T} - WV||_F^2$, where each row of $W_{m\times d}$ contains the weights of a linear classifiers for each tag. After embedding the re-weighted empirical loss term, the objective of Fasttag with image-tag re-weighting (Fasttag-RW) is to minimize the loss function $\mathscr{L}(W, B)$ as

$$\min_{B,W} \underbrace{||(T - B\hat{T}) \circ Z||_F^2}_{re-weighted\ loss} + \gamma||B\hat{T} - WV||_F^2 + \lambda||W||^2, \tag{4.9}$$

where $L2$ regularization on W is to reduce complexity and avoid overfitting. In Equation 4.9, T is not directly optimized, and instead, the backtracked corruption method [35] is used to approximate the re-weighted empirical loss. In particular, some tags are removed randomly with independent probability $p \in (0, 1)$ from $\hat{T}$ to generate a further corrupted version $\bar{T} = \{\bar{Y}_1, ..., \bar{Y}_n\}$ from $\hat{T}$. This backtracked corruption procedures are repeated several times, and the expected corruption value to approximate the re-weighted empirical loss is used. Finally, the empirical loss can be approximated as

$$\begin{aligned} ||(T - B\hat{T}) \circ Z||_F^2 &\approx \mathbb{E}\left[||(\hat{T} - B\bar{T}) \circ Z||_F^2\right]_{p(\bar{T}|\hat{T})} \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[||(\hat{Y}_i - B\bar{Y}_i) \circ z_i||_F^2\right]_{p(\bar{Y}_i|\hat{Y}_i)}, \end{aligned} \tag{4.10}$$

where z_i is the i-th column in Z for the i-th image. When expanding Equation 4.10, referring to $P = \sum_{i=1}^{n} \hat{Y}_i \mathbb{E}[\bar{Y}_i \circ z_i]^\top$ and $Q = \sum_{i=1}^{n} \mathbb{E}[\bar{Y}_i \circ z_i \bar{Y}_i^\top]$, the re-weighted empirical loss in Equation

4.10 can be rewritten as

$$\|(T - B\hat{T}) \circ Z\|_F^2 \approx \frac{1}{n} trace(BQB^\top - 2PB^\top + \hat{T} \circ Z\hat{T}^\top). \tag{4.11}$$

Here P and Q can be computed as

$$\begin{aligned} &P = (1-p)\hat{T} \circ Z\hat{T}^\top, \\ &Q = (1-p)^2\hat{T} \circ Z\hat{T}^\top + p(1-p)\delta(\hat{T} \circ Z\hat{T}^\top), \end{aligned} \tag{4.12}$$

where $\delta(\cdot)$ is an operator to set all entries except for the diagonal to zero. By substituting Equation 4.11 and Equation 4.12 in Equation 4.9, the block-coordinate descent algorithm can be used to optimize B and W alternatingly. Finally, the close form solutions for B and W are derived as

$$\begin{aligned} B =&(P + \gamma W V \hat{T}^\top)(Q + \gamma \hat{T}\hat{T}^\top)^{-1}, \\ W =&\gamma B\hat{T}V^\top(\gamma VV^\top + n\lambda 1_{d \times d})^{-1}. \end{aligned} \tag{4.13}$$

The deriving details for parameter optimization in the proposed two re-weighted based models are depicted in Appendix B.

4.4.3 Discussion

Once the penalty matrix Z is computed, the proposed two re-weighted models TMC-RW and Fasttag-RW maintain the same level of computational complexity since the close-form solutions for parameters in Equation 4.8 and Equation 4.13 have the same forms as TMC and Fasttag models and only the products of matrices are involved in the re-weighted penalty Z. Specifically, the convergence rate for the subgradient descent method adopted in TMC-RW is $\mathcal{O}(1/\sqrt{t})$, where t is the number of iterations. And the space requirement of TMC-RW is $\mathcal{O}(nm)$. Regarding Fasttag-RW, the solving the ridge regression in Equation 4.13 requires $\mathcal{O}(nm^3)$, as usually we have the number of training samples $n \gg m$, thus the training time of Fasttag-RW scales linearly to n. Similar as the case of Fasttag in [35], Fasttag-RW converges with a few bootstrap iterations (usually ranges from 2 to 8).

4.5 Experiments

4.5.1 Experimental configuration

Datasets and Features. In the experiments, three benchmark datasets ESP Game [63], IAPRTC-12 [51] and MIR Flickr [56] are used, which are very challenging with large tag vocabulary and significant diversity among visual content. Table 4.1 shows general statistics of these datasets. In Table 4.1, the counts of images are in the formatted as "training/test" and the counts are in the format as "mean/minimum/maximum" in the forth and fifth columns. The original multiple features including global descriptors (RGB, HSV, LAB, Gist) and local descriptors (SIFT, Hue) were extracted in [13], and widely utilized in subsequential works [10, 34, 35]. Here the multiple feature mapping and random projection scheme in [35] is adopted, dimensionality reduction is applied to all original features and they are concatenate to a 12,912-dimensional merged feature to represent each image.

Table 4.1 General statistics of the three datasets.

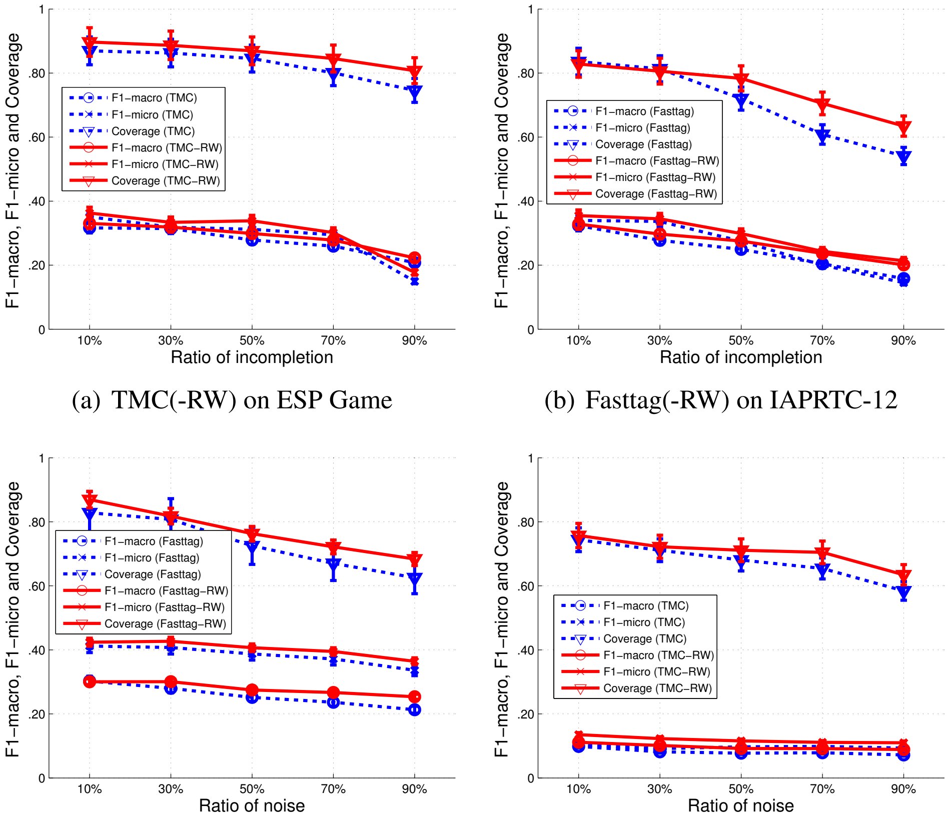
Dataset	Images	Labels	Labels per image	Images per label
ESP Game	18,689/2,081	268	4.7/1/15	326/172/4,553
IAPRTC-12	17,665/1,962	291	5.7/1/23	34/153/4,999
MIR Flickr	9,359/9,335	457	4.55/1/45	145/50/1,483

Defective tag assignments. To simulate the situation of defective tag assignments, experiments are conducted in two settings: incomplete setting and noisy setting. (1) For incomplete setting, partial tags are randomly deleted for each images , the deletion process stands by the principle $\min(1, \lceil N \times (1 - ratio) \rceil)$ ensuring that each images preserves at least one tag. (2) For noisy setting, untagged tags are randomly added for each images, the addition process follows the principle $N + \min(1, \lceil N \times ratio \rceil)$, ensuring that each image is corrupted by at least one noisy tag. Here N denotes the counts of originally tagged tags in one image, $\lceil \cdot \rceil$ denotes the ceiling function which gives the smallest integer not smaller than the given value, $ratio$ represents the degree of incompletion or noise. In the experiments, $ratio = \{10\%, 30\%, 50\%, 70\%, 90\%\}$. It indicates that the larger the ratio, the higher the degree of incompletion or noise. Note that $ratio = 0\%$ means the dataset is fully labeled without any incompletion or noise.

Evaluation metrics. Three standard evaluation criteria used in previous tag completion works [34, 35, 57] and tag refinement works [60, 61] are adopted: (1) *F1-macro*, which is the averaged F1 score ($F1 = 2\frac{Precision*Recall}{Precision+Recall}$) of all tags, (2) *F1-micro*, which is the averaged F1 score of all images, (3) *Coverage* ($Coverage = \frac{N_+}{N_{total}}$), where N_+ is the count of recalled tags by model and N_{total} is the quantity of tags in vocabulary. To calculate these criteria, all images are predicted with top five tags which have highest confidence score. It is notable that for all three criteria, larger numerical value indicates better performance. All the experiments are conducted on a computer which has Inter Xeon E5520 2.27GHz CUP, 16GB RAM.

4.5.2 Evaluation on various degree of defection

Firstly, the proposed re-weighting based TMC-RW and Fast-RW models are compared with previous TMC and Fasttag on various degree of incomplete and noisy settings. To obtain the best parameters for these models, 30% of initial training samples in each dataset are randomly split as validation set, and perform model selection. Experiments are repeated five times for each model, and report the evaluation results averaged over the five trials.



(a) TMC(-RW) on ESP Game

(b) Fasttag(-RW) on IAPRTC-12

(c) Fasttag(-RW) on IAPRTC-12

(d) TMC(-RW) on MIR Flickr

Fig. 4.3 Tag completion performance of proposed methods with various degree of incomplete and noisy settings.

Fig. 4.3 compares tag completion performance of the proposed re-weighting based models and the previous models without re-weighting scheme. It can be observed that in most conditions, varying the degree of incomplete and noisy settings, the tag completion results are constantly improved on all three evaluation metrics when the re-weighting scheme is taken into account. In particular, both TMC-RW and Fasttag-RW models are robust to serious condition of defection. For degree of incompletion and noise from 50% to 90%, TMC-RW and Fasttag-RW drop slightly on three measures while TMC and Fasttag decrease drastically. The reason behind this is that when the degree of defection is serious, using equal penalty for image-tag is too coarse to distinguish missing and noisy tags from other tags, whereas the re-weighted penalty is more reasonable and appropriate to represent the relatedness of pairwise image-tag since the defection can be compensated by utilizing both visual similarities and tag correlations.

ESP Game	IAPRTC-12	MIR Flickr	ESP Game	IAPRTC-12	MIR Flickr
Ground truth : boat, coin, money, old, silver	**Ground truth :** horizon, orange, sky, sun	**Ground truth :** sunset, italy, italia, roma, rome	**Ground truth :** grass, green, house, sky, tree	**Ground truth :** building, car, street	**Ground truth :** yellow, flowers, orange
Incompletion 70%: boat, money	**Incompletion 90%:** sun	**Incompletion 50%:** sunset, roma	**Noise 90%:** bikini, cat, game, green, house, light, sky, tree, grass	**Noise 70%:** building, car, field, shirt, street	**Noise 30%:** yellow, flowers, orange, newyork
Fasttag: coin, metal, head, old, round	**TMC:** cloud, sun, orange, painting, sunset	**Fasttag:** sunset, sky, blue, england, olympus	**TMC:** sand, boat, forest, tree, green	**Fasttag:** field, street, bay, car, road	**TMC:** light, flowers, nyc, yellow, holland
Fasttag-RW: coin(0.13), money(0.91), old(0.32), round(0.45), silver(0.25)	**TMC-RW:** sunset(0.11), horizon(0.35), sun(0.89), orange(0.27), landscape(0.37)	**Fasttag-RW:** sunset(0.88), italy(0.16), italia(0.21), roma(0.83), cityscape (0.35)	**TMC-RW:** grass(0.97), green(0.95), house(0.72), man(0.38), tree(0.81)	**Fasttag-RW:** building(0.95), car(0.81), pavement(0.46), sky(0.41), street (0.8)	**TMC-RW:** macro (0.24), flower(0.16), yellow(0.85), flowers(0.96), orange(0.79)

Fig. 4.4 Tag completion results of exemplar images on three datasets with various degree of incomplete (left three) and noisy (right three) settings.

Fig. 4.4 demonstrates tag completion results of exemplary images using the proposed methods with various degree of incomplete and noisy conditions. In Fig. 4.4, the fourth row is initial defective tags, last two rows are predicted tags from tag completion models. For the proposed models (last row), weighted penalty $Z_{i,j}$ of predicted tags are also presented. It can be learned from Fig. 4.4 that using the proposed re-weighting scheme, the proposed models are able to improve defective tags which are initially assigned, to recall related missing tags and to reject existed noisy tags. Moreover, the weighted penalties for predicted tags based on the principle of the proposed re-weighting scheme, which allocates smaller penalties to both missing and noisy tags than correctly labeled tags.

4.6 Summary

In this chapter, the author considers the practical issue of training dataset with defective tag assignments. The author proposes an image-tag re-weighting scheme and formulates a re-weighted empirical loss term which is more efficient for optimization in tag completion problems under defective tag assignments. In particular, the weighted penalty of pairwise image-tag is derived from image-image similarities and tag-tag associations, which ensures that missing and noisy tags have discriminant importance rather than identical importance in the previous works. Experimental evaluations verify that using the proposed re-weighting scheme achieves superior tag completion results on various degree of incomplete and noisy settings. In addition, the proposed re-weighting scheme can also benefit image annotation performance which assumes complete tag assignments.

Chapter 5

Multimodal learning for images, texts and their semantics

5.1 Introduction

In the previous chapters, the author has discussed the topic model used in the traditional image annotation problems and also the practical issues of *incomplete labeling* and *defective tag assignments* of the training image datasets. Actually, a promising application of image annotation could be searching for a set of images using a query string or multiple keywords (e.g., as is done in Google Image search). This application can be considered as a reverse process of image annotation, which is called "keyword based image retrieval." The "keyword based image retrieval" involves different modalities of data: images and texts.

Recently, jointly modeling different modalities of images and texts has been an active research area, as there has been massive explosion of multimedia content such as images and texts from multiple input channels on the Internet. Usually, the images are usually associated with texts. Fig. 5.1 illustrates concrete examples of images with the associated texts. The associated text are used to describe the semantic content of these images. In particular, the term "keywords" refers to high level labels or categories, and the terms "text" and "tag" are textual descriptions taking from the surrounding web pages and the user-provided tags respectively. It is worth noting that the semantics of an image may be given by one or more *keywords* and the *text/tags* are typically noisy and may not explicitly mention the ground truth *keywords*.

Indeed, there are two types of cross-modal retrieval scenarios relying on the different modalities of images and texts: (1) "Text2Img": automatically associating one or more semantically relevant images given a piece of text (i.e., a label, phrase or caption) [64–

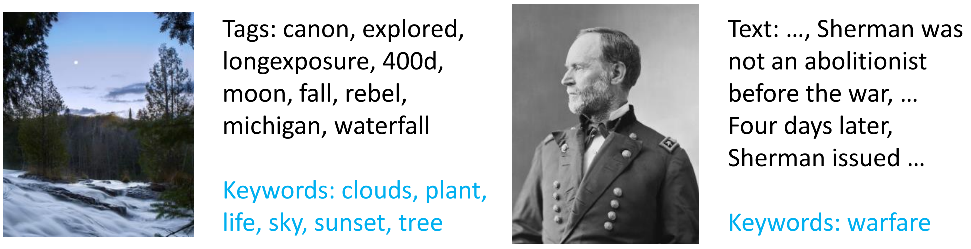

Tags: canon, explored, longexposure, 400d, moon, fall, rebel, michigan, waterfall
Keywords: clouds, plant, life, sky, sunset, tree
Text: ..., Sherman was not an abolitionist before the war, ... Four days later, Sherman issued ...
Keywords: warfare

Fig. 5.1 Examples of Internet images with associated texts.

66]; (2) “Img2Text”: automatically annotating images using semantic labels [12, 13, 67] or describing images with phrases [66]. Due to the distinct statistical properties from the different modalities of image and text, one of the primary issues involved in the cross-modal retrieval is learning useful representations about such data by fusing the modalities into a joint representation. Indeed, the modalities of image and text may have different kinds of representations and correlation structures. For example, an image is usually represented by real-values and dense feature descriptors, whereas a text is represented as discrete sparse word count vectors. However, many existing methods only consider the direct correlation between the original representations of images and corresponding texts, leading to inefficient latent subspaces for representing data of both modalities. Moreover, these methods also have difficulties in discovering the highly non-linear relationship across different modalities.

In this chapter, the author investigates the problem of modeling images and their associated text for cross-modal retrieval tasks such as Text2Img and Img2Text. To make the data from images and texts modalities comparable, the previous cross-modal retrieval methods directly learn two projection matrices to map the raw features of the two modalities into a common subspace, in which cross-modal data matching can be performed. However, the different feature representations and correlation structures of different modalities inhibit these methods from efficiently modeling the relationships across modalities through a common subspace.

To handle the diversities of different modalities, the author aims to learn a more efficient latent subspace from the original representations of different modalities and proposes a unified framework (see Fig. 5.2) for cross-modal retrieval problems. The proposed framework consists of *coupled dictionary learning* and *coupled feature mapping*. Firstly, the original multimodal data transformed into sparse representations via coupled dictionary learning, with the guarantee that the generated representations are homogeneous. Then, these representations are acquired by coupled linear regression method where data from each modality are simultaneously projected into a common subspace of keywords. For the testing stage (i.e. Img2Text or Text2Img), the sparse representation of a given query from one modality is

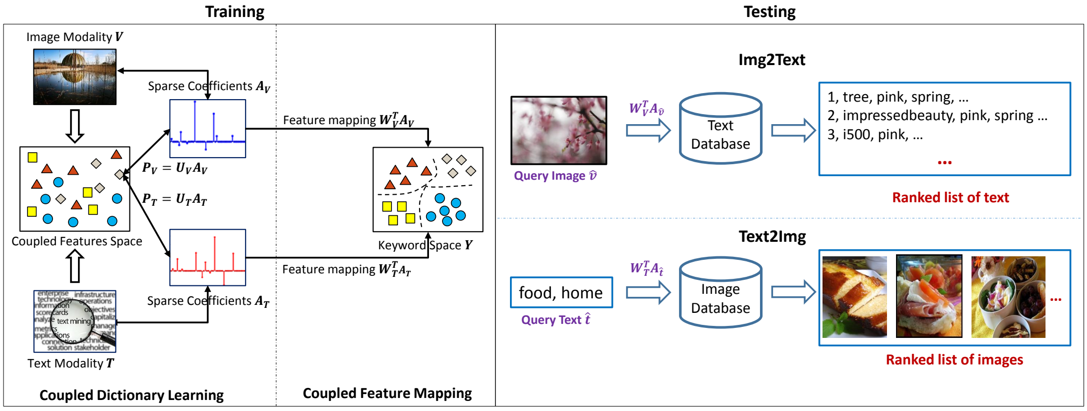


Fig. 5.2 Illustration of the proposed framework for cross-modal retrieval, where $\{\mathbf{A}_V, \mathbf{A}_T\}$ and $\{\mathbf{W}_V, \mathbf{W}_T\}$ are the sparse representations and projection matrices for image and text modalities, respectively.

generated using the learned coupled dictionary of this modality, then the sparse representation is mapped into the common (keyword) subspace using the learned projection matrix. Finally, in the common subspace, the best match is selected from the other modality as the output.

The main contributions of the work in this chapter can be summarized as follows:

- A novel framework is proposed to integrate the schemes of coupled dictionary learning and coupled feature mapping for the cross-modal retrieval problem. The proposed framework cannot only handle the diversity of different modal data, but also improve the efficiency of subspace learning and coupled feature selection.

- The previous scheme of coupled dictionary learning developed for single modality is extended to the case of multi-modal data in this work, and an efficient iterative algorithm of the dictionary learning is derived to solve the corresponding minimization problem.

- The proposed framework is evaluated with a variety of cross-modal retrieval tasks on three challenging datasets, and the experimental results show that the proposed framework outperforms several relevant state-of-the-art approaches.

5.2 Related work

The problems of image and text retrieval are well-studied research topics [68]. Some of the earliest research is based on retrieval of unimodal data, i.e., both query and retrieved data belong to the same modality (either image or text). For example, given a text query, the unimodal methods directly match the query with textual meta-data on the web instead of examining the corresponding image. Usually, these unimodal approaches cannot be applied to cross-modal retrieval.

Recently, a number of successful approaches [67, 69–71] designed for cross-modal retrieval are proposed to jointly model images and associated texts. The major goal of these works is learning a common latent space for image and text modalities, in which both modalities are projected into the same space for cross-modal retrieval. Specifically, several approaches rely on Canonical Correlation Analysis (CCA), a classic technique that maps two views, such as visual and textual features, into a common latent space where the correlation between the two views is maximized. Hwang and Grauman [71] have modeled the relative importance of words based on the order from the user provided annotations to improve cross-modal retrieval results. Ballan *et al.* [67] have used kernel CCA (KCCA) to develop cross-view approach to associate images and text. Gong *et al.* [70] have leveraged CCA for cross-modal retrieval under the multi-view learning framework, in which the image and the text views are linked to the underlying third view of semantic labels.

Except for the CCA based approaches, there are some other methods for the cross-modal retrieval. Chen *et al.* [72] have applied the Partial Least Squares (PLS) to cross-modal document retrieval. They use PLS to transform the visual features into the text space, then learn a semantic space to measure the similarity between two different modalities. Recently, Sharma *et al.* [73] made a comprehensive analysis for the multi-view learning framework to deal with cross-modal retrieval problem. They extend traditional discriminative methods, i.e., Linear Discriminant Analysis (LDA), Marginal Fisher Analysis (MFA), to the multi-view counterpart: Generalized Multi-view LDA (GMLDA) and Generalized Multi-view MFA (GMMFA). Wang *et al.* [74] proposed a generic framework to jointly perform common subspace learning and coupled feature selection from different modalities of data. They unified coupled linear regression, L_{21} norm and trace norm regularization terms into the generic framework and achieved the state-of-the-art performance for cross-modal retrieval task.

It can be learned from these methods that learning the common latent space is crucial for matching data of different modalities. However, when learning the latent space, these methods only considered the direct correlation between image and text modalities, ignoring the intrinsic diversity of representations and correlational structures in them. Therefore, these methods have difficulty in incorporating the highly non-linear relationship between the low-level features across different modalities. To overcome this, the author takes advantage of dictionary learning to obtain the sparse representations of different modalities in a coupled learning manner. The sparse representations are homogenous for different modalities and can incorporate the relationship across modalities, ensuring to learn a more representative latent space. The details of the proposed framework for cross-modal retrieval will be described in the next section.

5.3 Proposed framework

In this section, the author will first present the details of the two procedures in the training stage as shown in Fig. 5.2, and then develop iterative algorithm for learning parameters in each procedure. Finally, the testing stage of the proposed framework will be described.

5.3.1 Coupled dictionary learning

Let $\mathscr{D} = \{I_1, ..., I_n\}$ be a collection of n images, each image $I_i = (\mathbf{v}_i, \mathbf{t}_i)$ consists of features from two modalities: d_1 dimensional visual feature vector and d_2 dimensional textual feature vector. Here, the data from image modality is denoted as $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_n] \in \mathbb{R}^{d_1 \times n}$ and data from text modality as $\mathbf{T} = [\mathbf{t}_1, ..., \mathbf{t}_n] \in \mathbb{R}^{d_2 \times n}$, respectively. To handle the diversity of different modalities $\mathbf{V}$ and $\mathbf{T}$, the sparse representation from dictionary learning for each modality is used since it has been shown to be very effective in data representation or reconstruction tasks. It is worth noting that the dictionary learning for different modalities is coupled and an associate function $f(\cdot)$ is imposed to relate the sparse representations of different modalities. The coupled dictionary learning can be formulated as the following minimization problem:

$$\begin{aligned} \min_{\mathbf{D}_V, \mathbf{D}_T, \mathbf{A}_V, \mathbf{A}_T} \quad & \|\mathbf{V} - \mathbf{D}_V \mathbf{A}_V\|_F^2 + \|\mathbf{T} - \mathbf{D}_T \mathbf{A}_T\|_F^2 + \sigma(\|\mathbf{A}_V\|_1 + \|\mathbf{A}_T\|_1) + f(\mathbf{A}_V, \mathbf{A}_T) \\ s.t. \quad & \|\mathbf{d}_{v,i}\|_2 \leq 1, \|\mathbf{d}_{t,i}\|_2 \leq 1, \forall i. \end{aligned} \tag{5.1}$$

In Equation 5.1, the original features $\mathbf{V}$ and $\mathbf{T}$ of the two modalities are reconstructed from the sparse representations $\mathbf{A}_V \in \mathbb{R}^{k_1 \times n}$ and $\mathbf{A}_T \in \mathbb{R}^{k_2 \times n}$, respectively. Here $\mathbf{A}_V$ and $\mathbf{A}_T$ rely on the learnt dictionaries $\mathbf{D}_V \in \mathbb{R}^{d_1 \times k_1}$ and $\mathbf{D}_T \in \mathbb{R}^{d_2 \times k_2}$, where k_1 and k_2 are dictionary size for $\mathbf{D}_V$ and $\mathbf{D}_T$, respectively. And $f(\mathbf{A}_V, \mathbf{A}_T)$ is an associating function defining the cross-modal relationship in terms of $\mathbf{A}_V$ and $\mathbf{A}_T$. Once the relationship between $\mathbf{A}_V$ and $\mathbf{A}_T$ is captured, the dictionaries $\mathbf{D}_V$ and $\mathbf{D}_T$ can be updated accordingly.

Regarding $f(\mathbf{A}_V, \mathbf{A}_T)$, following the scheme in [75], a k_c dimensional common feature space $\mathscr{P}$ is introduced to relate the sparse representations of different modalities. It is restricted that $k_c = k_1 = k_2$ so that different modalities of data should be comparable in $\mathscr{P}$. In addition, $\mathbf{P}_V = \mathbf{U}_V \mathbf{A}_V$ is the projected data from sparse representations $\mathbf{A}_V$ of image modality, where $\mathbf{U}_V \in \mathbb{R}^{k_c \times k_1}$ is the projection matrix. The same remarks are applied to $\mathbf{P}_T$ and $\mathbf{U}_T$.

And it is required that each pair of samples from $\mathbf{A}_V$ and $\mathbf{A}_T$ can be projected to the same instance of $\mathbf{P}_V$ and $\mathbf{P}_T$ in space $\mathscr{P}$ (see the illustration in Fig. 5.2). To be more precise, for an instance $\mathbf{p} = \mathbf{U}_V \alpha_V$ (or $\mathbf{p} = \mathbf{U}_T \alpha_T$) in $\mathscr{P}$ which is projected from the sparse representation α_V (or α_T) of image (or text) modality, the corresponding sparse representation of text (or image) modality $\alpha_T = \mathbf{U}_T^{-1}\mathbf{p}$ (or $\alpha_V = \mathbf{U}_V^{-1}\mathbf{p}$) can be reconstructed. Finally, the definition of $f(\mathbf{A}_V, \mathbf{A}_T)$ can be formulated as

$$f(\mathbf{A}_V, \mathbf{A}_T) = \gamma(\|\mathbf{A}_V - \mathbf{U}_V^{-1}\mathbf{P}_T\|_F^2 + \|\mathbf{A}_T - \mathbf{U}_T^{-1}\mathbf{P}_V\|_F^2) + \mu(\|\mathbf{U}_V^{-1}\|_F^2 + \|\mathbf{U}_T^{-1}\|_F^2), \quad (5.2)$$

where additional regularized constraints on $\mathbf{U}_V^{-1}$ and $\mathbf{U}_T^{-1}$ ensure numerical stability and avoid over-fitting. It can be learnt that the constraints in Equation 5.2 exhibits the capabilities in recovering the sparse representations in one modality using data projected from the other, and hence, the relationship across different modalities can be efficiently incorporated in the sparse representations.

By substituting Equation 5.2 into Equation 5.1, the objective function for the coupled dictionary learning procedure can be obtained. Although the objective function is not jointly convex with respect to dictionaries $\{\mathbf{D}_V, \mathbf{D}_T\}$, sparse representations $\{\mathbf{A}_V, \mathbf{A}_T\}$ and projection matrices $\{\mathbf{U}_V, \mathbf{U}_T\}$, it is convex with respect to each of them when fixing the other variables. They can be iteratively updated in an alternating manner. The proposed iterative algorithm is outlined in Algorithm 2 and the detailed optimization is provided in Appendix C.

5.3.2 Coupled feature mapping

Let $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n] \in \mathbb{R}^{c \times n}$ be a keyword matrix for all images in $\mathscr{D}$. The coupled feature mapping procedure aims to learn two projection matrices $\mathbf{W}_V \in \mathbb{R}^{k_1 \times c}$ and $\mathbf{W}_T \in \mathbb{R}^{k_2 \times c}$, which map the sparse representations $\mathbf{A}_V$ and $\mathbf{A}_T$ of the two modalities into the common space defined by the keywords. Note that the roles of $\mathbf{W}_V$ and $\mathbf{W}_T$ are different from $\mathbf{U}_V$ and $\mathbf{U}_T$ of coupled dictionary learning procedure: the former bridges the sparse representations to the keyword space, while the latter performs on the intermediate space $\mathscr{P}$.

In this work, the author leverages a coupled linear regression method and develops two schemes (a simple scheme and an advanced scheme) for coupled feature mapping, based on how the projection matrices are learnt. Specifically, for the simple scheme, the ridge regression method is adopted to minimize projection errors of the sparse representations of each modality to the keyword space. The generic minimization problem of the simple scheme is derived as:

$$\min_{\mathbf{W}_V,\mathbf{W}_T} \|\mathbf{W}_V^\top \mathbf{A}_V - \mathbf{Y}\|_F^2 + \|\mathbf{W}_T^\top \mathbf{A}_T - \mathbf{Y}\|_F^2 + \lambda(\|\mathbf{W}_V\|_F^2 + \|\mathbf{W}_T\|_F^2). \tag{5.3}$$

The analytical solutions of $\mathbf{W}_V$ and $\mathbf{W}_T$ can be derived as:

$$\begin{aligned} \mathbf{W}_V &= \mathbf{Y}\mathbf{A}_V^\top(\mathbf{A}_V\mathbf{A}_V^\top + \lambda I)^{-1}, \\ \mathbf{W}_T &= \mathbf{Y}\mathbf{A}_T^\top(\mathbf{A}_T\mathbf{A}_T^\top + \lambda I)^{-1}. \end{aligned} \tag{5.4}$$

For the advanced scheme, the coupled feature mapping method in [74] is adopted to further incorporate coupled feature selection. Intuitively, relevant and discriminative features from the sparse representations can be selected to enhance the relevance across different modalities. Therefore, for the advanced scheme, it has similar objective function as in [74]:

$$\begin{aligned} \min_{\mathbf{W}_V,\mathbf{W}_T} & \frac{1}{2}(\|\mathbf{W}_V^\top \mathbf{A}_V - \mathbf{Y}\|_F^2 + \|\mathbf{W}_T^\top \mathbf{A}_T - \mathbf{Y}\|_F^2) \\ & + \lambda_1(\|\mathbf{W}_V\|_{21} + \|\mathbf{W}_T\|_{21}) + \lambda_2\|[\mathbf{W}_V^\top \mathbf{A}_V \mathbf{W}_T^\top \mathbf{A}_T]\|_*, \end{aligned} \tag{5.5}$$

where the L_{21} norm[3] $\|\cdot\|_{21}$ is used to select features from coupled feature space jointly, and the trace norm[4] $\|\cdot\|_*$ encodes the correlations of different modalities with a low-rank constraint. It is worth noting that the advanced scheme in Equation 5.5 differs from [74], i.e., the coupled feature selection is performed on sparse representations $\mathbf{A}_V$ and $\mathbf{A}_T$ to leverage the merit from coupled dictionary learning, rather than on original features of both modalities in [74].

5.3.3 Test phase for cross-modal retrieval

Algorithm 2 summarizes the iterative algorithm used in the above two procedures of the training stage. Once the training stage is complete, the dictionaries $\{\mathbf{D}_V, \mathbf{D}_T\}$, sparse representations $\{\mathbf{A}_V, \mathbf{A}_T\}$, projection matrices $\{\mathbf{W}_V, \mathbf{W}_T\}$ of two modalities can be obtained successively. In the testing phase, given a test image $\hat{I} = (\hat{\mathbf{v}}, \hat{\mathbf{t}})$, the sparse representations $\hat{\mathbf{A}}_{\hat{v}}$ and $\hat{\mathbf{A}}_{\hat{t}}$ based on learnt dictionaries $\mathbf{D}_V$ and $\mathbf{D}_T$ is firstly generated, respectively. Then $\hat{\mathbf{A}}_{\hat{v}}$ and $\hat{\mathbf{A}}_{\hat{t}}$ are projected into the keyword space through projection matrices $\mathbf{W}_V$ and $\mathbf{W}_T$. To

[3] For a matrix $\mathbf{M} \in \mathbb{R}^{n\times m}$, its i-th row and j-th column are denoted by $\mathbf{M}^{(i)}$, $\mathbf{M}_j$ respectively. The L_{21} norm of $\mathbf{M}$ is $\|\mathbf{M}\|_{21} = \sum_{i=1}^n \|\mathbf{M}^{(i)}\|_2$.

[4] The trace norm of $\mathbf{M}$ is defined as $\|\mathbf{M}\|_* = \sum_{i=1}^{\min(m,n)} \tau_i$, where τ_i denotes the i-th singular value of M.

Algorithm 2: Iterative Algorithm for the proposed framework

```
Input: Image feature matrix $\mathbf{V}$ with associated text feature matrix $\mathbf{T}$, parameters $\sigma$, $\gamma$ and $\mu$.
1: Initialize $\{\mathbf{D}_V^0, \mathbf{D}_T^0\}$ and $\{\mathbf{A}_V^0, \mathbf{A}_T^0\}$ by [75], and $\{\mathbf{U}_V^0, \mathbf{U}_T^0\}$ as $\mathbf{I}$ for iteration $i = 0$.
2: Let $\mathbf{P}_V^0 \leftarrow \mathbf{U}_V^0\mathbf{A}_V^0$ and $\mathbf{P}_T^0 \leftarrow \mathbf{U}_T^0\mathbf{A}_T^0$.
3: repeat
4:     Update $\mathbf{D}_V^{i+1}$, $\mathbf{D}_T^{i+1}$ with $\mathbf{A}_V^i$, $\mathbf{A}_T^i$ and $\mathbf{U}_V^i$, $\mathbf{A}_T^i$.
5:     Update $\mathbf{A}_V^{i+1}$, $\mathbf{A}_T^{i+1}$ with $\mathbf{D}_V^{i+1}$, $\mathbf{D}_T^{i+1}$ and $\mathbf{U}_V^i$, $\mathbf{A}_T^i$.
6:     Update $\mathbf{U}_V^{i+1}$, $\mathbf{U}_T^{i+1}$ with $\mathbf{D}_V^{i+1}$, $\mathbf{D}_T^{i+1}$ and $\mathbf{A}_V^{i+1}$, $\mathbf{A}_T^{i+1}$.
7:     Update $\mathbf{P}_V^{i+1} \leftarrow \mathbf{U}_V^{i+1}\mathbf{A}_V^{i+1}$ and $\mathbf{P}_T^{i+1} \leftarrow \mathbf{U}_T^{i+1}\mathbf{A}_T^{i+1}$.
8:     Set $i = i + 1$.
9: until Objective function of Equation 5.1 converges.
Output: Dictionaries $\{\mathbf{D}_V, \mathbf{D}_T\}$, sparse representations $\{\mathbf{A}_V, \mathbf{A}_T\}$.
Input: Sparse representations $\{\mathbf{A}_V, \mathbf{A}_T\}$, keywords matrix $\mathbf{Y}$, parameters $\lambda$ (or $\{\lambda_1, \lambda_2\}$).
10: Compute $\{\mathbf{W}_V, \mathbf{W}_T\}$ according to Equation 5.4 or solutions of Equation 5.5.
Output: Projection matrices $\{\mathbf{W}_V, \mathbf{W}_T\}$.
```

perform cross-modal retrieval, one modal data (i.e., image modality) of $\hat{I}$ can be taken as the query to retrieve the other modality (i.e., text modality).

5.3.4 Discussion

The time consuming for training the proposed model includes sparse representation generation for the coupled dictionary learning procedure and linear regression for the feature mapping procedure. Typically, the computational complexity of the sparse representation generation is $\mathcal{O}(ndk_c^2)$, and the computational complexity of the linear regression is $\mathcal{O}(nck_c^3)$. Note that, usually we have the number of training samples $n \gg d$ and $n \gg k_c$, thus the total time complexity of training the proposed model is linear to n, which is scalable and efficient for large-scale datasets. For the testing time, given a test sample of one modality, generating its sparse representation and matching the query requires $\mathcal{O}(dck_c)$ and $\mathcal{O}(c^2)$, respectively. Thus the testing time is not as important as that of the training time.

5.4 Experimental results

5.4.1 Experimental setting

The proposed framework is applied to two cross-modal retrieval tasks: *Text2Img* and *Img2Text*. Given an image (or a text) query, the goal is to find the nearest neighbors from the text (or the image) database. The two proposed schemes (simple and advanced) are compared with several related methods on two publicly available datasets: Pascal Voc [71] and Wiki [69] datasets, in which each image is assigned with only one keyword. In addition, a more specific task called *Img2Keyword* is taken into account. The goal of *Img2Keyword* to predict the results from the keyword space rather than retrieving a sample from the text modality for a given image. It can be considered as a traditional image annotation task and here it is used to evaluate the efficiency of the proposed feature mapping into the common subspace defined by keywords. For this task, a challenging dataset MIRFlickr-25K is adopted, in which each image is assigned with at least one keyword. Table 5.1 lists some of the general statistics of the three datasets used in this work.

Table 5.1 General statistics of three datasets used in this work.

Dataset	Images	Keywords	Image feature	Text feature
Pascal Voc	2808, 2841	20	512-dim Gist	399-dim word frequency histogram
Wiki	1300, 1566	10	128-dim SIFT	10-dim LDA features
MIRFilckr-25K	12500, 12500	38	7500-dim multiple features[5]	457-dim word frequency histogram

To evaluate the performance of the proposed schemes, the standard measure of mean average precision (MAP) in [69] is used for *Img2Text* and *Text2Img* tasks. To compute MAP, the average precision (AP) of the retrieval result of each query is calculated, then average the AP values over all queries in the query set is computed. Moreover, to evaluate the *Img2Keyword* task, the standard measures of image annotation tasks are adopted: average precision per keyword (P), average recall per keyword (R) and $F1$ measure ($F1 = 2 \times \frac{P \times R}{P+R}$). The values of P and R are computed by predicting the top five keywords for each test image on MIRFlickr-25K dataset. Notable that for all the measures, larger numerical value indicates better performance. In the training stage, the parameters σ, γ and μ is empirically set to 0.01, 0.0001 and 0.001 for the coupled dictionary learning procedure, and the parameters λ, λ_1 and λ_2 to 0.01, 0.1 and 0.001 for the feature mapping procedure, respectively. In testing phase, the cosine distance is adopted to measure the similarity of features and select the matches. All the experiments are conducted on a computer which has Inter Xeon E5520 2.27GHz CUP with 16GB RAM.

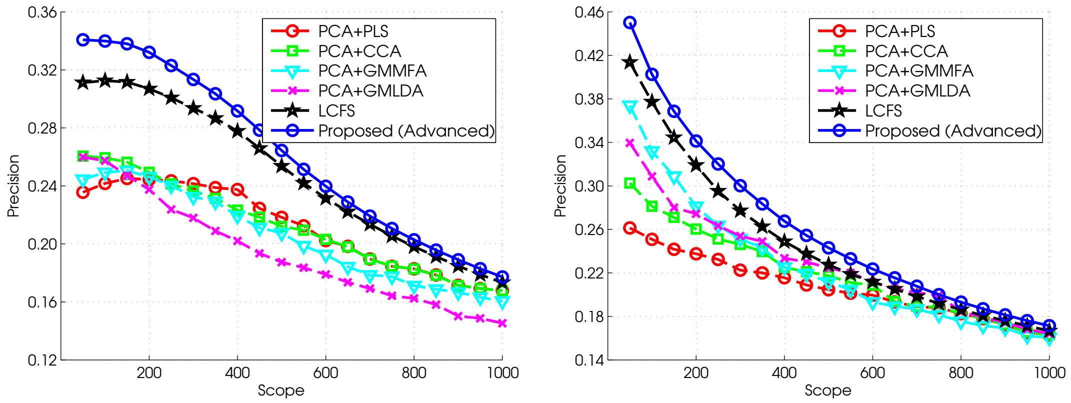
[5] The multiple features including various local and global features, such as SIFT, HUE, RGB, HSV, LAB, GIST and etc. They can be downloaded from http://lear.inrialpes.fr/people/guillaumin/data.php.

5.4.2 Results on Pascal Voc and Wiki datasets

The two proposed schemes is firstly compared with common subspace learning methods PLS [72], CCA [71], GMMFA [73], GMLDA [73] and LCFS [74] and the results for *Img2Text* and *Text2Img* tasks are reported. Using the settings of LCFS method, Principal Component Analysis (PCA) is performed on the original features of Pascal Voc dataset to remove redundant features for PLS, CCA, GMMFA, GMLDA methods while preserving the original features for LCFS method and the proposed two schemes. Since the Wiki dataset has low dimensional features, theses features are directly used without the process of dimension reduction.

Table 5.2 MAP scores of different methods on Pascal Voc dataset. Previous and proposed best results are highlighted in bold.

Method	Img2Text	Text2Img	Average
PCA+PLS	0.2757	0.1997	0.2377
PCA+CCA	0.2655	0.2215	0.2435
PCA+GMMFA	0.3090	0.2308	0.2699
PCA+GMLDA	0.2418	0.2038	0.2228
LCFS	0.3438	0.2674	0.3056
Proposed (Simple)	0.3262	**0.2997**	0.3129
Proposed (Advanced)	**0.3741**	0.2944	**0.3342**



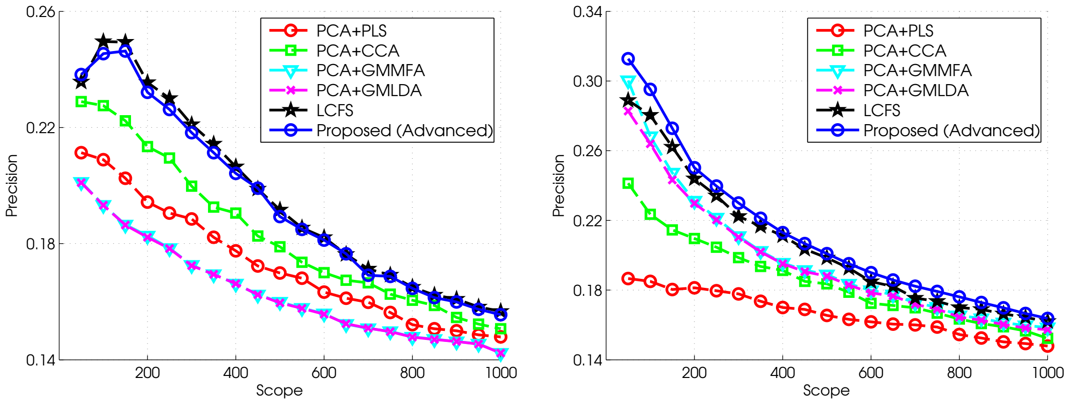
(a) Img2Text on Pascal Voc

(b) Text2Img on Pascal Voc

Fig. 5.3 Performance of different methods on the Pascal Voc dataset, based on precision-scope curve for $K = 50$ to 1000: (a) image query to match text, (b) text query to match image.

Table 5.2 shows the results of different methods in terms of MAP on Pascal Voc dataset. It can be observed that the proposed simple scheme significantly outperforms the previous methods. It shows that the coupled dictionary learning algorithm has the advantage of outputting sparse representation that preserves the relationship among different modalities. Moreover, the advanced scheme makes further improvement over the simple one, showing that coupled feature selection can be done in the sparse representations and benefits the learned common space for feature mapping.

The corresponding precision-scope curves on Pascal Voc dataset are plotted in Fig. 5.3 for both forms of cross-modal retrieval tasks, i.e., Image query to match text (a) and Text query to match image (b). The scope (i.e., the top K retrieved samples) for the precision-scope varies from 50 to 1,000. It can be observed that compared with the previous methods, the proposed method obtains better results for both tasks.



(a) MAP on Pascal Voc

(b) MAP on Wiki

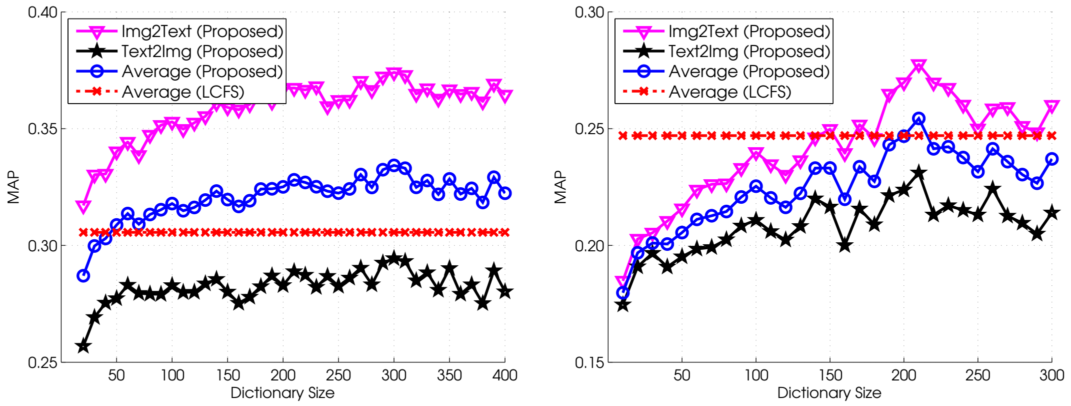
Fig. 5.4 Performance of different methods on the Wiki dataset, based on precision-scope curve for $K = 50$ to 1000: (a) image query to match text, (b) text query to match image.

Table 5.3 shows the MAP scores of different approaches on the Wiki dataset. On average, the proposed schemes achieves higher MAP scores than the other methods but perform worse than some methods on *Img2Text* task. As discussed in [74], it is challenging to improve on this dataset due to the low dimensions of image and text features. Nevertheless, the proposed schemes consistently outperform previous methods especially on *Text2Img* task. To further understand the reason, the performance is compared by changing the dictionary size of the coupled dictionary learning step in the advanced scheme. Fig. 5.5 shows the MAP scores with different dictionary size on the two datasets. It can be learned that, larger dictionary size generally has better capability for sparse representation and the best MAP scores on the two datasets are achieved with dictionary size 300 and 210, respectively. Therefore, the limitation of low dimensional (10-dim) text features (see Table 5.1) of the Wiki dataset can be tackled by the coupled dictionary learning procedure where more efficient high dimensional (210-dim) sparse features are used, benefitting the coupled feature selection in the feature mapping procedure. However, the decay on *Img2Text* indicates that the sparse representations of different modalities may need to be further balanced.

The corresponding precision-scope curves on Wiki dataset are plotted in Fig. 5.4, and it can be seen that for both forms of cross-modal retrieval tasks, the proposed method finds more correct matches in the top K documents compared with its several counterparts.

Table 5.3 MAP scores of different methods on Wiki dataset. Previous and proposed best results are highlighted in bold.

Method	Img2Text	Text2Img	Average
PLS	0.2402	0.1663	0.2032
CCA	0.2549	0.1846	0.2198
GMMFA	0.2750	0.2139	0.2445
GMLDA	0.2751	0.2098	0.2425
LCFS	**0.2798**	0.2141	0.2470
Proposed (Simple)	0.2717	0.2233	0.2475
Proposed (Advanced)	0.2776	**0.2311**	**0.2544**

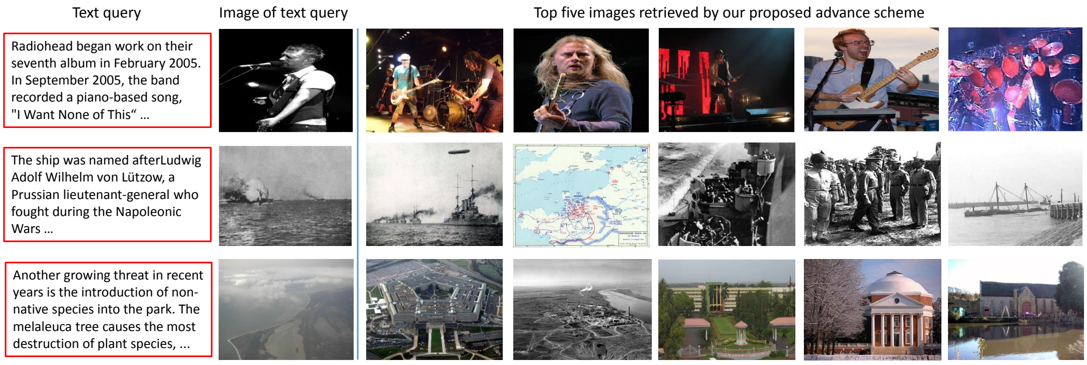


(a) MAP on Pascal Voc

(b) MAP on Wiki

Fig. 5.5 The MAP with different dictionary size of the proposed scheme (advanced) on the Pascal Voc dataset (a) and Wiki dataset (b), where the best MAP score (Average) of LCFS method is also provided for guides in red dotted line.

Specifically, for *Text2Img* task, Fig. 5.6 shows two examples of text queries and the top five images retrieved using the advanced scheme. In each example, the text query and its corresponding image are shown at the left, and the top five images retrieved are listed in the following columns. It can be observed that the advanced scheme finds the closet matches of the image modality at the semantic level for both text queries. And the retrieved images are all belonging to the same keyword of the text queries (“music” at the top, “warfare” at the bottom).



Radiohead began work on their seventh album in February 2005. In September 2005, the band recorded a piano-based song, "I Want None of This" ...

The ship was named afterLudwig Adolf Wilhelm von Lützow, a Prussian lieutenant-general who fought during the Napoleonic Wars ...

Another growing threat in recent years is the introduction of non-native species into the park. The melaleuca tree causes the most destruction of plant species, ...

Fig. 5.6 Two examples of the *Text2Img* task obtained by the proposed advanced scheme on Wiki dataset.

5.4.3 Results on MIRFlickr-25K dataset

The comparison is then made for the *Img2Keyword* task. To make comprehensive comparison, several state-of-the-art image annotation methods are also taken into account, which can also be evaluated for the *Img2Keyword* task. These methods include: (1) JEC [12] , Tagprop [13] and Fasttag [35], which only use image modality for learning model; (2) Multi-kernel SVMs [65], which applies different kernel functions to train SVM classifiers for image and text modalities; (3) Kernel CCA (KCCA) [67] and LCFS [74], which are common subspace learning methods. Since KCCA does not directly map image modal data into the keyword space, it is used for nearest neighbor selection then combine it with the nearest neighbor based tag propagation scheme Tagprop (this has been reported with promising result in [67]). For the proposed scheme, the dictionary size is empirically set to be 350 for coupled dictionary learning.

The overall performance on *Img2Keyword* task is shown in Table 5.4. It can be observed that: 1) using an additional text modality improves the accuracy of *Img2Keyword* task than only using the image modality; 2) both of the proposed schemes outperform the subspace learning method LCFS, indicating that the coupled dictionary learning procedure in the proposed framework is efficient to handle the diversity of different modalities and the learned sparse representations is more powerful than the original features for subspace learning; 3) the advanced scheme achieves the highest performance and generally outperforms the state-of-the-art method KCCA+Tagprop, showing that the learned sparse representations

Table 5.4 Comparison of *Img2Keyword* in terms of P, R and F1 measure for the proposed schemes and related methods on MIRFlickr-25K dataset. Previous and proposed best results are highlighted in bold.

Method	P	R	F1
JEC	0.3290	0.1733	0.2270
Tagprop	0.4521	0.3021	0.3622
Fasttag	0.4498	0.3637	0.4021
Multi-kernel SVMs	0.5162	0.3658	0.4282
KCCA+Tagprop	0.5469	0.3542	0.4299
LCFS	0.3253	0.3122	0.3186
Proposed (Simple)	0.4933	0.3445	0.4056
Proposed (Advanced)	**0.5890**	**0.3779**	**0.4604**

are powerful for subspace learning and coupled feature selection is crucial in enhancing the relationships across different modalities.

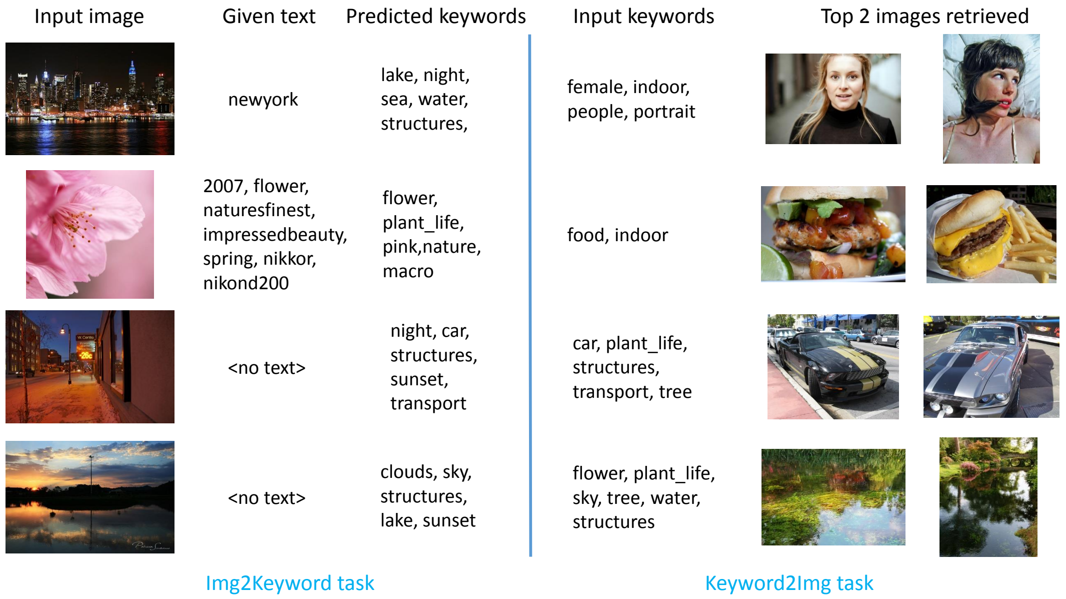

Input image
Given text
Predicted keywords
Input keywords
Top 2 images retrieved
newyork
lake, night, sea, water, structures,
female, indoor, people, portrait
2007, flower, naturesfinest, impressedbeauty, spring, nikkor, nikond200
flower, plant_life, pink,nature, macro
food, indoor
<no text>
night, car, structures, sunset, transport
car, plant_life, structures, transport, tree
<no text>
clouds, sky, structures, lake, sunset
flower, plant_life, sky, tree, water, structures
Img2Keyword task
Keyword2Img task

Fig. 5.7 Examples of the *Img2Keyword* and *Keyword2Img* tasks obtained by the proposed advanced scheme on MIRFlickr-25K dataset.

Fig. 5.7 shows several *Img2Keyword* and *Keyword2Img* examples obtained by the proposed advanced scheme on MIRFlickr-25K dataset. In each case, the query image or text is shown at the left, and the retrieved texts or images are listed at the following columns. It can be observed that the proposed advanced scheme can efficiently find the closest matches across different modalities of data.

5.5 Summary

In this chapter, the author has proposed a novel framework to solve the problem of cross-modal retrieval. The framework consists of two procedures: coupled dictionary learning and coupled feature mapping. The former procedure is imposed to obtain homogeneous sparse representations, which is to handle the diversities of different modalities and incorporate the relationships across them. The latter procedure is utilized to learn projection matrices based on the sparse representations of different modalities, which can efficiently map the data of different modalities into the keywords space for comparison. The author has developed efficient iterative algorithm to solve the derived minimization problem in the proposed framework. Experimental results on a variety of cross-modal retrieval tasks shows the superiority of the proposed framework. In the future, the author will explore the potentiality of the proposed framework in fill in missing modalities given the observed one, to deal with the issue of absence of some modalities.

Chapter 6

Conclusion

To make conclusion, the main contributions in the thesis are depicted in Section 6.1. Then the future work are discussed in Section 6.2.

6.1 Contributions

Modeling the topic correlation of images and texts via latent topic model: The author has made the first trial to extend the LDA based topic model for image annotation problem, and then extends the popular LDA based topic models such as corrLDA and sLDA-bin to the CTM based models: corrCTM and sCTM-bin, which can incorporate topic correlation. A general mean-field variational algorithm for parameter estimation is derived in these CTM based models. Informative comparison of annotation performance between proposed CTM based and previous LDA based models are performed, which shows the superiority of the proposed models.

Exploring image-label associations in weakly labeled datasets: For the issue of incomplete labeling, the author has put effort to improve OVA-SSVM in three aspects:

1. A method that can tackle with more general case where each training image has one or more labels is proposed. The proposed method is more efficient for the multi-label annotation problem than the OVA-SSVM method.

2. A novel image specific structured loss term is formulate, which is more appropriate than the classification-style structured loss used in OVA-SSVM, to account for the dependencies of predicted multiple labels of a specific image.

3. The author also has developed an efficient optimization algorithm with lower complexity by exploiting the properties of the proposed structured loss.

Finally, experiments are conducted to compare the proposed method with OVA-SSVM on image classification task, observing that proposed method performs on par with OVA-SSVM and can automatically capture the semantic hierarchies without pre-definition (used in OVA-SSVM). Extensive evaluation on two benchmark annotation datasets is conducted with a variety of the setting of incompleteness for image annotation task. The empirical results demonstrate that the proposed method is significantly better than OVA-SSVM, and achieves competitive annotation performance compared with other state-of-the-art methods designed for incomplete labelling.

For the issue of defective tag assignments, the author has proposed an efficient re-weighting scheme to elaborately assign the penalty of predicting each tag to each image, and the penalty of pairwise image-tag is fully derived from both image-image similarities and tag-tag associations. Then, a unified re-weighted empirical loss term is formulated by utilizing the re-weighted penalty term. Finally, the author has extended two recently proposed tag completion models to incorporate the re-weighted empirical loss term, developing the corresponding optimization algorithms to perform tag completion with defective tag assignments.

Multimodal learning for images, texts and their semantics: The author has proposed a novel framework that unifies coupled dictionary learning and coupled feature mapping for the cross-modal retrieval problem. The proposed framework can not only handle the diversity of different modal data, but also improve the efficiency of subspace learning and coupled feature selection. The previous sparse representation method of coupled dictionary learning developed for unimodal data is extended to the case of multi-modal data, and an efficient iterative algorithm is developed to solve the complex minimization problem in the proposed framework. The proposed framework is evaluated with a variety of cross-modal retrieval tasks on three challenging datasets, and the experimental results show that the proposed framework outperforms several relevant state-of-the-art approaches.

6.2 Future work

In the conventional image annotation problem, as discussed in Chapter 2, the target is to train an annotation model from a collections of annotated training images with a fixed size of pre-defined labels. However, the corpus of labels can only support general visual knowledge and does not cover the semantics of the entire world. The expansibility of the annotation model is limited since it cannot predict newly discovered labels that are not in the pre-defined corpus. In this case, the annotation model should incrementally learn the new visual knowledge and predict new labels. Thus, the incremental learning scheme could be employed to enable the model to tackle with image with new visual knowledge. And it is necessary to design a framework in which the topic representation can be efficiently used with large size of labels, e.g., tens of thousands of labels extracted from social media, which is an essentially difficult challenge.

In the Chapter 3 and Chapter 4, the author has investigated the practical issues of incomplete labeling and defective tag assignments in the training images, and it can be found in the experiments that the quality of the manually annotated labels of the training images highly impacts the performance of the annotation systems. And the author has developed a data-driven based scheme to compute the image-label associations from the training corpus. However, it is still difficult to capture the true associations while maintaining quality and consistency of image labeling. Indeed, the prior knowledge such as ontologies and semantic hierarchy of labels can be utilized when computing the associations of image and labels. It is necessary to develop more appropriate scheme to incorporate the prior knowledge and consider the scalability of the scheme when performing on large quantities of training images.

In Chapter 5, the author has considered the cross-modal retrieval task and focused on eliminating the diversity of two different modalities of images and texts. A potential problem of the proposed framework is that it consists of two loosely connected steps. The objective functions in the two steps are formulated separately and the parameters in the two functions are not jointly learned. Indeed, the proposal can be further improved into a compact framework and all the parameters can be learned jointly, which may reasonably leads to better performance. Moreover, besides the two modalities of images and texts, various information provided by other modalities (e.g., data of various sensors) can be integrated to extend the cross-modal system to a multi-modal scenario. For examples, the different information sources such as GPS and inertial sensors can be utilized, which may provide additional information. In addition, the audio information along with the visual frames in the videos can also be leveraged to enrich the relationship among different modalities.

Acknowledgements

First of all, I would like to express my sincere gratitude to my advisor Prof. Taniguchi for the continuous support of my Ph.D study and living in Japan, for his insightful guidance, invaluable advice and encouragement. His guidance helped me in all the time of research and learning the Japanese culture.

I would like to thank the rest of my thesis committee: Prof. Uchida, Associate Prof. Nagahara and Associate Prof. Shimada, for their encouragement, and insightful comments. They gave me a lot of beneficial suggestions on my research and writing the thesis.

I sincerely appreciate the support from the Chinese Scholarship Council (CSC) and the concern from Consulate General of The Peoples Republic of China in Fukuoka. Thanks to the financial support from the CSC, I could concentrate on my research during my Ph.D study. In addition, I also appreciate the travel grants offered by Google Inc. and other non-profit organizations such as Telecommunications Advanced Foundation and NEC C&C Foundation, with their support I can attend prestigious international conferences several times for academic exchange.

I am also grateful to all my colleagues in LIMU and to the secretary Kiyoko Furuta. I stay three and a half years of wonderful and memorable time with them. Last but not least, I would like to express my deep thanks to my parents for all their endless love and encouragement through all the time of my life.

References

[1] Matthieu Guillaumin. *Exploiting Multimodal Data for Image Understanding*. PhD thesis, Université de Grenoble, 2010.

[2] G. Chen, J. Zhang, F. Wang, C. Zhang, and Y. Gao. Efficient multi-label classiffication with hypergraph regularization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1658–1665, 2009.

[3] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[4] Jia Deng. *Large Scale Visual Recognition*. PhD thesis, Stanford University, 2012.

[5] Yu Xiang, Xiangdong Zhou, Tat-Seng Chua, and Chong-Wah Ngo. A revisit of generative model for automatic image annotation using markov random fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1153–1160, 2009.

[6] D.M.Blei, A.Y.Ng, and M.I.Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, pages 993–1022, 2003.

[7] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57, 1999.

[8] D.Putthividhya, H.T.Attias, and S.S.Nagarajan. Supervised topic model for automatic image annotation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 321–326, 2010.

[9] G. Carneiro, A.B. Chan, P.J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 29:394–410, 2007.

[10] Y. Verma and C. V. Jawahar. Exploring svm for image annotation in presence of confusing labels. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 325–330, 2013.

[11] Alexander Binder, Wojciech Samek, Klaus-Robert Müller, and Motoaki Kawanabe. Enhanced representation and multi-task learning for image annotation. *Computer Vision and Image Understanding (CVIU)*, 117(5):466–478, 2013.

[12] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *European Conference on Computer Vision (ECCV)*, pages 316–329, 2008.

[13] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 309–316, 2009.

[14] Yashaswi Verma and C.V. Jawahar. Image annotation using metric learning in semantic neighbourhoods. In *European Conference on Computer Vision (ECCV)*, volume 7574, pages 836–849, 2012.

[15] Zijia Lin, Guiguang Ding, and Mingqing Hu. Image auto-annotation via tag-dependent random search over range-constrained visual neighbours. *Multimedia Tools and Applications*, pages 1–26, 2014.

[16] D.M.Blei and J.D.Lafferty. Correlated topic models. In *International Conference on Machine Learning (ICML)*, pages 113–120, 2006.

[17] Xing Xu, A. Shimada, and R.-i. Taniguchi. Correlated topic model for image annotation. In *the 19th Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV)*, pages 201–208, 2013.

[18] Xing Xu, A. Shimada, and R.-i. Taniguchi. Latent topic model for image annotation by modeling topic correlation. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 110–115, 2013.

[19] Xing Xu, Atsushi Shimada, and Rin-ichiro Taniguch. Exploring image specific structured loss for image annotation with incomplete labelling. In *Asian Conference on Computer Vision (ACCV)*, pages 704–719, 2015.

[20] Xing Xu, A. Shimada, and R.-i. Taniguchi. Tag completion with defective tag assignments via image-tag re-weighting. In *IEEE International Conference on Multimedia Expo (ICME)*, pages 482–487, 2014.

[21] Xing Xu, A. Shimada, R.-i. Taniguchi, and Li He. Coupled dictionary learning and feature mapping for cross-modal retrieval. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2015.

[22] D.M.Blei and M.I.Jordan. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 127–134, 2003.

[23] Duangmanee Putthividhya, Hagai T.Attias, and Srikantan S.Nagarajan. Topic regression multi-modal latent dirichlet allocation for image annotation. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 841–848, 2010.

[24] C.Wang, D.M.Blei, and Fei fei Li. Simultaneous image classification and annotation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 929–936, 2009.

[25] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 289–296, 1999.

[26] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 524–531, 2005.

[27] Hideki Nakayama. *Linear Distance Metric Learning for Large-scale Generic Image Recognition*. PhD thesis, University of Tokyo, 2011.

[28] Tomasz J. Malisiewicz and Jonathan C. Huang. Detecting objects via multiple segmentations and latent topic models. Technical report, Carnegie Mellon University, 2006.

[29] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision (IJCV)*, 77:157–173, 2008.

[30] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010.

[31] P. Duygulu, K. Barnard, J.F.G. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision (ECCV)*, pages 97–112, 2002.

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

[33] D. Grangier and S. Bengio. A discriminative kernel-based model to rank images from text queries. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 6:10–23, 2008.

[34] S.S. Bucak, R. Jin, and A.K. Jain. Multi-label learning with incomplete class assignments. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1352–1359, 2011.

[35] M. Chen, A. Zheng, and K. Weinberger. Fast image tagging. In *International Conference on Machine Learning (ICML)*, pages 1274–1282, 2013.

[36] Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit S. Dhillon. Large-scale multi-label learning with missing labels. In *International Conference on Machine Learning (ICML)*, pages 1236–1243, 2013.

[37] I. Dimitrovski, D. Kocev, S. Loskovska, and S. Deroski. Detection of visual concepts and annotation of images using ensembles of trees for hierarchical multi-label classification. In *CLEF Workshop On Recognizing Patterns in Signals, Speech, Images and Videos*, pages 152–161, 2010.

[38] X. Lou and F.A. Hamprecht. Structured learning from partial annotations. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1519–1526, 2012.

[39] J. McAuley, A. Ramisa, and T. Caetano. Optimization of robust loss functions for weakly-labeled image taxonomies. *International Journal of Computer Vision (IJCV)*, pages 257–265, 2011.

[40] Chun-Nam John Yu and Thorsten Joachims. Learning structural svms with latent variables. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 1169–1176, 2009.

[41] F. Wu, Y. Han, Q. Tian, and Y. Zhuang. Multi-label boosting for image annotation by structural grouping sparsity. In *Proceedings of the International Conference on Multimedia (MM)*, pages 15–24, 2010.

[42] Hao Fu, Qian Zhang, and Guoping Qiu. Random forest for image annotation. In *European Conference on Computer Vision (ECCV)*, volume 7577, pages 86–99, 2012.

[43] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, and K Yu. Largescale image classification: fast feature extraction and svm training. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1689–1696, 2011.

[44] J. Sanchez and F. Perronnin. High-dimensional signature compression for large-scale image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1665–1672, 2011.

[45] B. Hariharan, L. Zelnik-manor, S.V.N. Vishwanathan, and M. Varma. Large scale max-margin multi-label classification with priors. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 1428–1435, 2010.

[46] H. Wang, H. Huang, and C Ding. Image annotation using birelational graph of images and semantic labels. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 793–800, 2011.

[47] S.L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1002–1009, 2004.

[48] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *European Conference on Computer Vision (ECCV)*, pages 561–568, 2012.

[49] Xing Xu, Atsushi Shimada, and Rin-ichiro Taniguchi. Image annotation by learning label-specific distance metrics. In *International Conference on Image Analysis and Processing (ICIAP)*, pages 101–110, 2013.

[50] Tsochantaridis I., Joachims T. Hofmann T., and Altun Y. Support vector machine learning for interdependent and structured output spaces. In *Proceeding of the 21th International Conference on Machine Learning (ICML)*, pages 104–111, 2004.

[51] M. Grubinger. *Analysis and Evaluation of Visual Information Systems Performance*. PhD thesis, Victoria University, 2007.

[52] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 48:1–48:9, 2009.

[53] Theo Gevers Koen E. A. van de Sande and Cees G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 32 (9):1582–1596, 2010.

[54] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th International Conference on Machine learning (ICML)*, pages 1453–1460, 2007.

[55] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advanced in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

[56] M. J. Huiskes and M. S. Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international Conference on Multimedia Information Retrieval*, pages 39–43, 2008.

[57] L. Wu, R. Jin, and A.K. Jain. Tag completion for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(3):716–727, 2013.

[58] D. Liu, S. Yan, X. Hua, and H. Zhang. Image retagging using collaborative tag propagation. *IEEE Transactions on Multimedia (TMM)*, 13:702–712, 2011.

[59] G. Zhu, S. Yan, and Y. Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *Proceedings of the International Conference on Multimedia (ACM MM)*, pages 461–470, 2010.

[60] X. Li, C. G M Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia (TMM)*, 11:1310–1322, 2009.

[61] T. Uricchio, M. Ballan, L.and Bertini, and A. Del Bimbo. An evaluation of nearest-neighbor methods for tag refinement. In *IEEE International Conference on Multimedia Expo (ICME)*, pages 142–147, 2013.

[62] Z. Lin, G. Ding, M. Hu, J. Wang, and X. Ye. Image tag completion via image-specific and tag-specific linear sparse reconstructions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 729–736, 2013.

[63] L. Von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (ACM SIGCHI)*, pages 319–326, 2004.

[64] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing System (NIPS)*, pages 1321–1328, 2011.

[65] J. Verbeek, M. Guillaumin, T. Mensink, and C. Schmid. Image annotation with tagprop on the mirflickr set. In *Proceedings of the International Conference on Multimedia Information Retrieval (ACM MIR)*, MIR '10, pages 537–546, 2010.

[66] Y. Verma and CV. Jawahar. Im2text and text2im: Associating images and texts for cross-modal retrieval. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 431–436, 2014.

[67] L. Ballan, T. Uricchio, L. Seidenari, and A. Del Bimbo. A cross-media model for automatic image annotation. In *International Conference on Multimedia Retrieval (ICMR)*, pages 1–6, 2014.

[68] R. Datta, D. Joshi, J. Li, and J. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computer Survey*, 40:1–60, 2008.

[69] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the International Conference on Multimedia (ACM MM)*, pages 251–260, 2010.

[70] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision (IJCV)*, 106:210–233, 2014.

[71] S. Hwang and K. Grauman. Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *International Journal of Computer Vision (IJCV)*, 100:134–153, 2012.

[72] Y. Chen, L. Wang, W. Wang, and Z. Zhang. Continuum regression for cross-modal multimedia retrieval. In *International Conference on Image Processing (ICIP)*, pages 1153–1156, 2012.

[73] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs. Generalized multiview analysis: a discriminative latent space. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2160–2167, 2012.

[74] K. Wang, R. He, W. Wang, L. Wang, and T. Tan. Learning coupled feature spaces for cross-modal matching. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2088–2095, 2013.

[75] D. Huang and Y. Wang. Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2496–2503, 2013.

[76] D.M.Blei and J.D.Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 1:17–35, 2007.

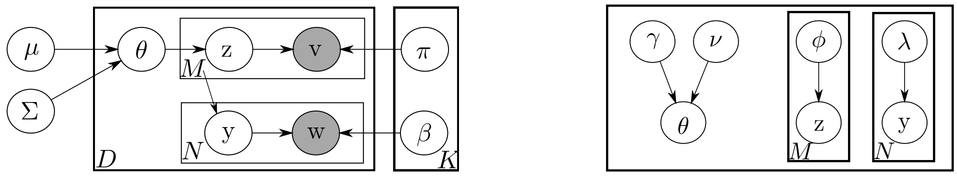
[77] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *International Conference on Machine Learning (ICML)*, pages 689–696, 2009.

Appendix A

This appendix depicts the mathematical derivation details of variational inference and parameter estimation of the proposed CTM based models: corrCTM and sCTM-bin in Chapter 2.

A.1 Derivation of optimization algorithm in corrCTM

As shown in Fig. A.1(a), the corrCTM model assumes that topic-feature matrix π and topic-word matrix β shares the same dimensional topic space K, to ensure the same sets of topics are used to generate corresponding between feature and word modalities. In addition, under corrCTM each word w directly shares a hidden topic variable with a randomly selected image feature v. The corrCTM model specifies the following joint distribution for an image in terms of features $\mathbf{v}$, words $\mathbf{w}$, topic proportion θ and topic assignment z:



(a) corrCTM

(b) variational approximation

Fig. A.1 Illustrations of the proposed corrCTM model: (a) graphical representation, (b) variational approximation for the latent variables.

$$p(\mathbf{v},\mathbf{w},\theta,z,y|\mu,\Sigma,\pi,\beta) = p(\theta|\mu,\Sigma)\left(\prod_{m=1}^{M} p(z_m|\theta)p(v_m|z_m,\pi)\right)\left(\prod_{n=1}^{N} p(y_n|M)p(w_n|y_n,z,\beta)\right). \tag{A.1}$$

Note that μ, Σ, π, β are model parameters we want to estimate. An intuitive way is using Maximum Likelihood Estimation (MLE). Therefore, we marginalize over the hidden variables θ, z, y for observation variables v, w. We then rewrite joint log probability in terms of μ, Σ, π and β as:

$$\log p(\mathbf{v},\mathbf{w}|\mu,\Sigma,\pi,\beta) = \int_{\theta}\sum_{z}\sum_{y} p(\mathbf{v},\mathbf{w},\theta,z,y|\mu,\Sigma,\pi,\beta)d\theta. \tag{A.2}$$

A.1.1 Variational inference

Unfortunately, Equation A.2 is intractable due to the coupling between θ and π in the summation over topic z. We use Jensen's inequality to bound the log probability $\log p(\mathbf{v},\mathbf{w}|\mu,\Sigma,\pi,\beta)$:

$$\begin{aligned}\log p(\mathbf{v},\mathbf{w}|\mu,\Sigma,\pi,\beta) &\geq \mathbb{E}_q[\log p(\mathbf{v},\mathbf{w},\theta,z,y|\mu,\Sigma,\pi,\beta)] + \mathbb{H}(q) \\ &= \mathscr{L}(\gamma,\nu,\phi,\lambda;\mu,\Sigma,\pi,\beta),\end{aligned} \tag{A.3}$$

where $\mathbb{H}(q)$ denotes entropy of variational distributions, $\mathbb{H}(q) = -\mathbb{E}_q[\log q(\theta,z,y|\gamma,\nu,\phi,\lambda)]$. The negative expectation is taken with respect to variational distributions of latent variables in Fig. A.1(a), and from variational approximation in Fig A.1(b) we can factorize $\mathbb{E}_q[\log q(\theta,z,y|\gamma,\nu,\phi,\lambda)]$ over introduced variational variables:

$$q(\theta,z,y|\gamma,\nu,\phi,\lambda) = q(\theta|\gamma,\nu)\left(\prod_{m=1}^{M} q(z_m|\phi_m)\right)\left(\prod_{n=1}^{N} q(y_n|\lambda_n)\right), \tag{A.4}$$

where the variational distribution of $\theta_{1:K}$ are K independent univariate Normals $\{\gamma_{1:K}, \nu_{1:K}\}$. The variational distributions of $z_{1:M}$ are specified by ϕ_m over K topics, and variational distributions of $y_{1:N}$ are specified by λ_n over M features.

Finally, the lower boundary $\mathscr{L}(\gamma,\nu,\phi,\lambda;\mu,\Sigma,\pi,\beta)$ in Equation A.3 can be factorized as:

$$\begin{aligned}\mathscr{L}(\gamma,\nu,\phi,\lambda;\mu,\Sigma,\pi,\beta) &= \sum_{m=1}^{M}\left(\mathbb{E}_q[\log p(z_m|\theta)] + \mathbb{E}_q[\log p(v_m|z_m,\pi)]\right) \\ &+\mathbb{E}_q[\log p(\theta|\mu,\Sigma)] + \sum_{n=1}^{N}\left(\mathbb{E}_q[\log p(y_n|M)] + \mathbb{E}_q[\log p(w_n|y_n,\beta)]\right) \\ &- \mathbb{E}_q[\log q(\theta|\gamma,\nu)] - \sum_{m=1}^{M}\mathbb{E}_q[\log q(z_m|\phi_m)] - \sum_{n=1}^{N}\mathbb{E}_q[\log q(y_n|\lambda)].\end{aligned} \tag{A.5}$$

The expectation terms in Equation A.5 which are different from corrLDA [22] are $\mathbb{E}_q[\log p(\theta|\mu,\Sigma)]$ and $\mathbb{E}_q[\log p(z_m|\theta)]$. Since η derives from multivariate Normals,

$$\mathbb{E}_q[\log p(\theta|\mu,\Sigma)] = (1/2)\log|\Sigma^{-1}| - (K/2)\log 2\pi - (1/2)\mathbb{E}_q[(\theta-\mu)^T\Sigma^{-1}(\theta-\mu)]. \tag{A.6}$$

Learned from [16, 76], the lower bound of non-conjugate logistic normal $f(\theta_i) = \exp\theta_i / \sum_j \exp\theta_j$ can be preserved with a Taylor expansion,

$$\begin{aligned}
\mathbb{E}_q[\log p(z_m|\theta)] &= \mathbb{E}_q[\theta^T z_m] - \mathbb{E}_q\left[\log(\sum_{i=1}^{K}\exp\theta_i)\right] \\
&\geq \mathbb{E}_q[\theta^T z_m] - \zeta^{-1}\left(\sum_{i=1}^{K}\mathbb{E}_q[\exp\theta_i]\right) - 1 + \log(\zeta) \\
&= \sum_{i=1}^{K}\gamma_i\phi_{mi} - \zeta^{-1}\left(\sum_{i=1}^{K}\mathbb{E}_q[\exp\theta_i]\right) - 1 + \log(\zeta), \qquad \text{(A.7)}
\end{aligned}$$

where ζ is a new variational variable introduced, and $\mathbb{E}_q[\exp\theta_i] = \exp\{\gamma_i + \nu_i^2/2\}$ for $i \in \{1, ..., K\}$. Finally, we can maximize Equation A.5 with respect to variational parameters $\gamma_{1:K}$, $\nu_{1:K}$, $\phi_{1:M}$ and $\lambda_{1:N}$. We use a coordinate ascent algorithm, iteratively maximizing the bound with respect to each parameter.

Table A.1 A summary of update rules in corrCTM for all parameters: the former five are variational parameters, and the latter four are model parameters.

Parameters	Update Rule
γ	$\partial\mathscr{L}(\gamma)/\partial\gamma = -\Sigma^{-1}(\gamma-\mu) + \sum_{m=1}^{M}\phi_{m,1:K} - (M/\zeta)\exp(\gamma+\nu^2/2)$
ν	$\partial\mathscr{L}/\partial\nu_i^2 = -\Sigma_{ii}^{-1}/2 - (M/2\zeta)\exp(\gamma+\nu_i^2/2) + 1/(2\nu_i^2)$
ζ_i	$= \sum_{i=1}^{K}\exp\{\gamma_i + \nu_i^2/2\}$
ϕ_{im}	$\propto \pi_{i,v_m}\exp\left(\gamma_i + \sum_{n=1}^{N}\lambda_{nm}\log\beta_{i,w_n}\right)$
λ	$\propto \exp\left(\sum_{i=1}^{K}\phi_{mi}\log\beta_{i,w_n}\right)$
μ	$= (1/D)\sum_{d=1}^{D}\gamma_d$
Σ	$= (1/D)\sum_{d=1}^{D}\left(I\,\nu_d^2 + (\gamma_d-\hat{\mu})(\gamma_d-\hat{\mu})^T\right)$
π_{mi}	$\propto \sum_{d=1}^{D}\sum_{m=1}^{M_d}\phi_{dmi}v_{dm}^{j}$
β_{nm}	$\propto \sum_{d=1}^{D}\sum_{n=1}^{N_d}w_{dn}^{j}\sum_{m=1}^{M_d}\phi_{dmi}\lambda_{dnm}$

A.1.2 Parameter estimation

Specifically, to estimate model parameters $\{\mu, \Sigma, \pi, \beta\}$, the objective function is the likelihood bound given by summing Equation A.5 over all training image-annotation pairs,

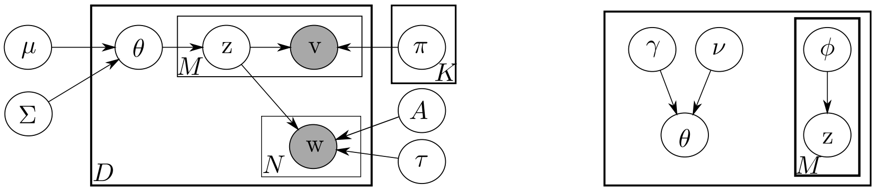
$$\mathbb{L}(\mu, \Sigma, \pi, \beta; \mathbf{v}_{1:D}, \mathbf{w}_{1:D}) \geq \sum_{d=1}^{D}\mathbb{E}_{q_d}[\log p(\theta_d, z_d, y_d|\mu, \Sigma, \pi, \beta)] + \mathbb{H}(q_d). \qquad \text{(A.8)}$$

We use variational EM algorithm to maximize this objective function. In E-step, we maximize the bound with respect to variational parameters $\{\gamma, \nu, \phi, \lambda, \zeta\}$ by performing variational inference for each document, depicted in Section A.1.1. In M-step, we maximize the bound with respect to model parameters $\{\mu,\Sigma,\pi,\beta\}$, given the variational distribution computed in E-step.

Here we supply a complete update rules for all variational parameters and model parameters in Table A.1. In practice, variational updates for γ, ν are not closed-form, we use LBFGS algorithm to update γ and constrained Newton's method to update ν .

A.2 Derivation of optimization algorithm in sCTM-bin

In sCTM-bin, each annotation word is treated as discrete response variable. In the generative process in Fig A.2(a), the occurrence of annotation words w is a multivariate Bernoulli of all words in dictionary. In addition, a logistic regression model from empirical topic proportions $\bar{z}$ of visual features v is used to define the probability of each word.



(a) corrCTM

(b) variational approximation

Fig. A.2 Illustrations of the proposed sCTM-bin model: (a) graphical representation, (b) variational approximation for the latent variables.

A.2.1 Variational inference

Following the similar methodology from corrCTM in Section A.1, the joint log likelihood of image-annotation pairs in one image can be formulated as,

$$\begin{aligned} \log p(\mathbf{v},\mathbf{w}|\mu,\Sigma,\pi,A,\tau) &\geq \mathbb{E}_q[\log p(\mathbf{v},\mathbf{w},\theta,z|\mu,\Sigma,A,\tau)]+\mathbb{H}(q) \\ &= \mathscr{L}(\gamma,\nu,\phi;\mu,\Sigma,\pi,A,\tau), \end{aligned} \tag{A.9}$$

where $\mathbb{H}(q)$ denotes entropy of variational distributions, $\mathbb{H}(q) = -\mathbb{E}_q[\log q(\theta,z|\gamma,\nu,\phi)]$. The negative expectation is taken with respect to variational distributions of latent variables in Fig A.2(a), and from variational approximation in Fig. A.2(b) we can factorize

$\mathbb{E}_q[\log q(\theta, z|\gamma, \nu, \phi)]$ over introduced variational variables:

$$q(\theta, z|\gamma, \nu, \phi) = q(\theta|\gamma, \nu)\left(\prod_{m=1}^{M} q(z_m|\phi_m)\right), \tag{A.10}$$

where the variational distribution of $\theta_{1:K}$ are K independent univariate Normals $\{\gamma_{1:K}, \nu_{1:K}\}$. The variational distributions of $z_{1:M}$ are specified by ϕ_m over K topics.

Finally, the lower boundary $\mathscr{L}(\gamma, \nu, \phi; \mu, \Sigma, \pi, A, \tau)$ in Equation A.9 can be factorized as:

$$\begin{aligned}\mathscr{L}(\gamma, \nu, \phi; \mu, \Sigma, \pi, A, \tau) &= \sum_{m=1}^{M} \left(\mathbb{E}_q[\log p(z_m|\theta)] + \mathbb{E}_q[\log p(v_m|z_m, \pi)]\right) \\ &+ \mathbb{E}_q[\log p(\theta|\mu, \Sigma)] + \sum_{n=1}^{N} \mathbb{E}_q[\log p(w_n|A, \tau)] \\ &- \mathbb{E}_q[\log q(\theta|\gamma, \nu)] - \sum_{m=1}^{M} \mathbb{E}_q[\log q(z_m|\phi_m)].\end{aligned} \tag{A.11}$$

The expectation terms in Equation A.11 which are different form sLDA-bin [8] are also $\mathbb{E}_q[\log p(\theta|\mu, \Sigma)]$ and $\mathbb{E}_q[\log p(z_m|\theta)]$. We can use Equation A.6, A.7 to formulate the lower boundary $\mathscr{L}(\gamma, \nu, \phi; \mu, \Sigma, \pi, A, \tau)$.

A.2.2 Parameter estimation

To estimate model parameters $\{\mu, \Sigma, \pi, A, \tau\}$, the objective function of likelihood bound over all training image-annotation pairs,

$$\mathbb{L}(\mu, \Sigma, \pi, A, \tau; \mathbf{v}_{1:D}, \mathbf{w}_{1:D}) \geq \sum_{d=1}^{D} \mathbb{E}_{q_d}[\log p(\theta_d, z_d|\mu, \Sigma, \pi, A, \tau)] + \mathbb{H}(q_d). \tag{A.12}$$

We use variational EM algorithm to maximize this objective function. In E-step, we maximize the bound with respect to variational parameters $\{\gamma, \nu, \phi\}$ by performing variational inference for each document. In M-step, we maximize the bound with respect to model parameters $\{\mu, \Sigma, \pi, A, \tau\}$, given the variational distribution computed in E-step.

Table A.2 summarizes a complete update rules for all variational parameters and model parameters. Here, ξ is an introduced variational parameter as the same as in sLDA-bin, $\lambda(\xi_i) = \tanh(\xi_i/2)/(4\xi_i)$.

Table A.2 A summary of update rules in sCTM-bin for all parameters: the top panel is variational parameters, and the bottom panel are model parameters.

Parameters	Update Rule
γ	$\partial \mathscr{L} / \partial \gamma = -\Sigma^{-1}(\gamma - \mu) + \sum_{m=1}^{M} \phi_{m,1:K} - (M/\zeta)\exp(\gamma + \nu^2/2)$
ν	$\partial L / \partial \nu_i^2 = -\Sigma_{ii}^{-1}/2 - (M/2\zeta)\exp(\gamma + \nu_i^2/2) + 1/(2\nu_i^2)$
ξ_i^2	$= A_i^T E[\overline{\mathbf{z}}\overline{\mathbf{z}}^T] A_i + 2\tau_i A_i E[\overline{\mathbf{z}}] + \tau_i^2$
ϕ_{im}	$\propto \pi_{i,v_m} \exp \left[\gamma_i + \sum_{j=1}^{V_t} \left(\frac{\lambda(\xi_j)}{2M} A_j - \frac{\lambda(\xi_j)}{M^2} \mathrm{diag}(A_j A_j^T) \ + \ 2 A_j A_j^T \sum_{n \neq m} \phi_{ni} \right) \right]$
μ	$= (1/D) \sum_{d=1}^{D} \gamma_d$
Σ	$= (1/D) \sum_{d=1}^{D} \left(I \ \nu_d^2 + (\gamma_d - \hat{\mu})(\gamma_d - \hat{\mu})^T \right)$
π_{mi}	$\propto \sum_{d=1}^{D} \sum_{m=1}^{M_d} \phi_{dmi} v_{dm}^j$
A_i	$= \left(2 \sum_{d=1}^{D} \lambda(\xi_i^d) E(\overline{\mathbf{z}}_d \overline{\mathbf{z}}_d^T) \right)^{-1} \left(\sum_{d=1}^{D} (w_i^d - 1/2 - 2\lambda(\xi_i^d)\tau_i) E[\overline{\mathbf{z}}_d] \right)$
τ_i	$= \left(\sum_{d=1}^{D} (w_i^d - 1/2 - 2\lambda(\zeta_i^d) A_i^T E[\overline{\mathbf{z}}_d]) \right) / \sum_{d=1}^{D} 2\lambda(\zeta_i^d)$

Appendix B

This appendix presents the mathematical derivation details of optimization algorithms utilized in the proposed TMC-RW and Fasttag-RW models in Chapter 4.

B.1 Optimization in TMC-RW

The objective of TMC model with image-tag re-weighting (TMC-RW) is to minimize the loss function $\mathscr{L}(T,w)$, which is derived as

$$\begin{aligned}\min_{T,w} \mathscr{L}(T,w) &= \min_{T,w} \underbrace{\|(T-\hat{T})\circ Z\|_F^2}_{re-weighted\ loss} + \eta\|TT^\top - R\|_F^2 + \lambda\|TT^\top - Vdiag(w)V^\top\|_F^2 \\ &\quad +\mu\|T\|_1 + \gamma\|w\|_1, \end{aligned} \tag{B.1}$$

where the $L1$ regularizer for both T and w is to generate sparse solutions for them. We adopt the subgradient-based and composite function optimization strategy proposed in [57] to optimize T and w with non-negative sparse constraints.

Defining $\mathscr{L}(T,w) = \mathscr{A}(T,w) + \mu\|T\|_1 + \gamma\|w\|_1$, $G = TT^\top - Vdiag(w)V^\top$, $H = TT^\top - R$, the subgradients of T and w for $\mathscr{A}(T,w)$ are derived as

$$\begin{aligned}\bigtriangledown_T \mathscr{A}(T,w) &= 2\lambda GT + 2\eta TH + 2(T-\hat{T})\circ Z, \\ \bigtriangledown_w \mathscr{A}(T,w) &= -2\lambda diag(V^\top GV), \end{aligned} \tag{B.2}$$

For only $\mathscr{A}(T,w)$ without $L1$ regularizer to T and w, the update solutions for T and w are as follows:

$$\begin{aligned}T_{t+1} &= [T_t - \eta_t \bigtriangledown_T \mathscr{A}(T_t,w_t)]_+, \\ w_{t+1} &= [w_t - \eta_t \bigtriangledown_T \mathscr{A}(T_t,w_t)]_+, \end{aligned} \tag{B.3}$$

where η_t is the step size of iteration t, $[f]_+ = \max(0,f)$ projects learnt w to be non-negative.

Now with $L1$ regularizer to T and w, the iterative update rules of T and w can be derived according to a composite function optimization method as

$$\begin{aligned} T_{t+1} &= arg\min_{T} \frac{1}{2} \|T - \hat{T}_{t+1}\|_F^2 + \eta_t \mu \|T\|_1, \\ w_{t+1} &= arg\min_{w} \frac{1}{2} \|w - \hat{w}_{t+1}\|_F^2 + \eta_t \gamma \|w\|_1, \end{aligned} \tag{B.4}$$

where variables $\hat{T}_{t+1}$ and $\hat{w}_{t+1}$ are intermediate results without $L1$ regularizer, and they can be derived as

$$\begin{aligned} \hat{T}_{t+1} &= T_t - \eta_t \bigtriangledown_T \mathscr{A}(T_t, w_t), \\ \hat{w}_{t+1} &= [w_t - \eta_t \bigtriangledown_T \mathscr{A}(T_t, w_t)]_+. \end{aligned} \tag{B.5}$$

Given Equation B.2 and Equation B.5, the final update solutions for T and w can be formulated as

$$\begin{aligned} T_{t+1} &= [\hat{T}_{t+1} - \eta_t \mu 1_{n \times m}]_+, \\ w_{t+1} &= [\hat{w}_{t+1} - \eta_t \gamma 1_{1 \times d}]_+. \end{aligned} \tag{B.6}$$

Algorithm 3 summarizes the entire procedure for sub-gradient based optimization in TMC-RW model.

Algorithm 3: Sub-gradient descent based optimization in TMC-RW

```
Input:
    Initial defective tag matrix $\hat{T}_{n \times m}$, image feature matrix $V_{n \times d}$, tag correlation matrix
    $R_{m \times m}$, image-tag re-weighted penalty matrix $Z_{n \times m}$, coefficients $\eta$, $\lambda$, $\mu$ and $\gamma$, converge
    threshold $\varepsilon$.
Output: recovered complete tag matrix $T$, image feature weight $w$.
Initialize $T_1 = \hat{T}$, $w_1 = 1_{1 \times d}$, $R = \hat{T}^\top \hat{T}$, iteration $t = 0$, $\varepsilon = 10^{-5}$.
repeat
    Set $t = t + 1$, step size $\eta_t = 1/t$.
    Update $T$ according to Equation B.6.
    Update $w$ according to Equation B.6.
until Convergence is reached to $\|\mathscr{L}(T_{t+1}, w_{t+1}) - \mathscr{L}(T, w)\| \leq \varepsilon \|\mathscr{L}(T, w)\|$.
```

B.2 Optimization algorithm in Fasttag-RW

The objective of Fasttag with image-tag re-weighting (Fasttag-RW) is to minimize the loss function $\mathscr{L}(W,B)$, which is formulated as,

$$\min_{W,B} \mathscr{L}(W,B) = \min_{W,B} \underbrace{\|(T - B\hat{T}) \circ Z\|_F^2}_{re-weighted\ loss} + \gamma \|B\hat{T} - WV\|_F^2 + \lambda \|W\|^2, \tag{B.7}$$

where $L2$ regularization on W is to reduce complexity and avoid overfitting.

In Equation B.7, we do not directly optimize T, and instead, we use the backtracked corruption method to approximate the re-weighted empirical loss. In particular, we randomly remove some tags with independent probability $p \in (0,1)$ from $\hat{T}$ to generate a further corrupted version $\bar{T} = \{\bar{Y}_1, ..., \bar{Y}_n\}$ from $\hat{T}$, and we repeat this backtracked corruption procedures several times, and use the expected corruption value to approximate re-weighted empirical loss. Finally, the empirical loss can be approximated as

$$\begin{aligned} \|(T - B\hat{T}) \circ Z\|_F^2 &\approx \mathbb{E}\left[\|(\hat{T} - B\bar{T}) \circ Z\|_F^2\right]_{p(\bar{T}|\hat{T})} \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\|(\hat{Y}_i - B\bar{Y}_i) \circ z_i\|_F^2\right]_{p(\bar{Y}_i|\hat{Y}_i)}, \end{aligned} \tag{B.8}$$

where z_i is the i-th column in Z for the i-th image. When expanding Equation B.8, we define $P = \sum_{i=1}^{n} \hat{Y}_i \mathbb{E}[\bar{Y}_i \circ z_i]^\top$ and $Q = \sum_{i=1}^{n} \mathbb{E}[\bar{Y}_i \circ z_i \bar{Y}_i^\top]$. Then the re-weighted empirical loss in Equation B.8 can be rewritten as

$$\|(T - B\hat{T}) \circ Z\|_F^2 \approx \frac{1}{n} trace(BQB^\top - 2PB^\top + \hat{T} \circ Z\hat{T}^\top). \tag{B.9}$$

Especially, the expected value of backtracked corruptions $\mathbb{E}[\bar{Y}_i \circ z_i] = (1-p)\hat{Y}_i$, and the variance matrix $\mathbb{V}[\bar{Y}_i \circ z_i \bar{Y}_i^\top] = p(1-p)\delta(\hat{Y}_i \circ z_i \hat{Y}_i^\top)$, where $\delta(\cdot)$ is an operator to set all elements but the diagonal to zero. Thus we can compute matrices P and Q as

$$\begin{aligned} P &= (1-p)\hat{T} \circ Z\hat{T}^\top \\ Q &= (1-p)^2 \hat{T} \circ Z\hat{T}^\top + p(1-p)\delta(\hat{T} \circ Z\hat{T}^\top). \end{aligned} \tag{B.10}$$

Now, by substituting Equation B.9 and Equation B.10 into Equation B.7, we can use block-coordinate descent algorithm to optimize B and W alternatingly. The gradient of

Algorithm 4: Iterative optimization in Fasttag-RW

```
Input:
    Initial defective tag matrix $\hat{T}_{m\times n}$, image feature matrix $V_{d\times n}$, image-tag re-weighted
    penalty matrix $Z_{m\times n}$, corruption probability $p$, coefficients $\gamma$, $\lambda$.
Output: enrichment mapping $B$, linear weights $W$.
1: Initialize $B = 1_{m\times m}$, $W = 1_{d\times m}$, $p = 0$.
2: repeat
3:     Set $p = p + 0.1$, $p \leq 1$.
4:     Fix $W$, calculate $B$ using Equation B.12.
5:     Fix $B$, calculate $W$ using Equation B.12.
6:     Calculate complete tag matrix $T = B\hat{T}$ or $T = WV$.
7: until Convergence is reached until $T$ no longer improves evaluation metrics F1-macro,
   F1-micro and Coverage.
```

$\mathscr{L}(W,B)$ with respect to B and W can be derived as

$$\begin{aligned} \nabla_B \mathscr{L}(B,W) &= B(Q + \gamma\hat{T}\hat{T}^\top) - P - \gamma W V \hat{T}^\top, \\ \nabla_W \mathscr{L}(B,W) &= W(\gamma V V^\top + n\lambda 1_{d\times d}) - \gamma B\hat{T}V^\top. \end{aligned} \tag{B.11}$$

By setting the gradients in Equation B.11 to be zero, we can formulate the close form solutions for B and W as

$$\begin{aligned} B &= (P + \gamma W V \hat{T}^\top)(Q + \gamma\hat{T}\hat{T}^\top)^{-1}, \\ W &= \gamma B\hat{T}V^\top(\gamma V V^\top + n\lambda 1_{d\times d})^{-1}. \end{aligned} \tag{B.12}$$

The iterative optimization algorithm to calculate the optimal B and W is shown in Algorithm 4.

Appendix C

This appendix provides the mathematical derivation details of the iterative algorithm for parameter learning in the proposed framework for cross-modal retrieval problem in Chapter 5. The author will first depict the parameter learning for coupled dictionary learning procedure and then for coupled feature mapping procedure.

C.1 Parameter learning for coupled dictionary learning

Given the data from two different modalities: image modality $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_n] \in \mathbb{R}^{d_1 \times n}$ and text modality $\mathbf{T} = [\mathbf{t}_1, ..., \mathbf{t}_n] \in \mathbb{R}^{d_2 \times n}$, the coupled dictionary learning procedure in the proposed framework aims to get homogenous sparse representations $\mathbf{A}_V$ and $\mathbf{A}_T$ for the two modalities based on their dictionaries $\mathbf{D}_V$ and $\mathbf{D}_T$, respectively. Based on the coupled learning framework in [75], an associate function $f(\cdot)$ (as interpreted in Chapter 5) is introduced to account for the relationships across the different modalities. The final minimization problem involved in the coupled dictionary learning procedure is formulated as:

$$
\begin{aligned}
\min_{\mathbf{D}_V, \mathbf{D}_T, \mathbf{A}_V, \mathbf{A}_T, \mathbf{U}_V, \mathbf{U}_T} & \|\mathbf{V} - \mathbf{D}_V \mathbf{A}_V\|_F^2 + \|\mathbf{T} - \mathbf{D}_T \mathbf{A}_T\|_F^2 + \sigma(\|\mathbf{A}_V\|_1 + \|\mathbf{A}_T\|_1) \\
& + \gamma(\|\mathbf{A}_V - \mathbf{U}_V^{-1}\mathbf{P}_T\|_F^2 + \|\mathbf{A}_T - \mathbf{U}_T^{-1}\mathbf{P}_V\|_F^2) + \mu(\|\mathbf{U}_V^{-1}\|_F^2 + \|\mathbf{U}_T^{-1}\|_F^2) \\
& s.t. \quad \|\mathbf{d}_{v,i}\|_2 \leq 1, \|\mathbf{d}_{t,i}\|_2 \leq 1, \forall i.
\end{aligned}
\tag{C.1}
$$

Similar as in [75], the objective function of Equation C.1 is convex with respect to dictionaries $\{\mathbf{D}_V, \mathbf{D}_T\}$, sparse representations $\{\mathbf{A}_V, \mathbf{A}_T\}$ and projection matrices $\{\mathbf{U}_V, \mathbf{U}_T\}$ when fixing the other variables. In practice, we can iteratively update these parameters in each iteration by an alternating manner.

C.1.1 Updating $\mathbf{D}_V$ and $\mathbf{D}_T$

When updating the dictionaries $\mathbf{D}_V$ and $\mathbf{D}_T$ of the two modalities during each iteration, we treat the sparse representations $\{\mathbf{A}_V, \mathbf{A}_T\}$ and projection matrices $\{\mathbf{U}_V, \mathbf{U}_T\}$ as constants. Then Equation C.1 can be simplified as:

$$\begin{aligned}
&\min_{\mathbf{D}_V} ||\mathbf{V} - \mathbf{D}_V \mathbf{A}_V||_F^2, \quad st. \quad ||\mathbf{d}_{\mathbf{v},i}||_2 \leq 1, \forall i, \\
&\min_{\mathbf{D}_T} ||\mathbf{T} - \mathbf{D}_T \mathbf{A}_T||_F^2, \quad st. \quad ||\mathbf{d}_{\mathbf{t},i}||_2 \leq 1, \forall i,
\end{aligned} \tag{C.2}$$

Equation C.2 is a typical form of quadratically constrained quadratic program (QCQP) with respect to $\mathbf{D}_V$ and $\mathbf{D}_T$, and it can be efficiently solved using Lagrange dual techniques.

C.1.2 Updating $\mathbf{A}_V$ and $\mathbf{A}_T$

Similarly, we calculate the solutions of sparse representations $\mathbf{A}_V$ and $\mathbf{A}_T$ while fixing projection matrices $\{\mathbf{U}_V, \mathbf{U}_T\}$ and dictionaries $\mathbf{D}_V$ and $\mathbf{D}_T$. Thus, we convert Equation C.1 into the following problem:

$$\begin{aligned}
&\min_{\mathbf{A}_V} ||\mathbf{V} - \mathbf{D}_V \mathbf{A}_V||_F^2 + \sigma ||\mathbf{A}_V||_1 + \gamma ||\mathbf{A}_V - \mathbf{U}_V^{-1} \mathbf{P}_T||_F^2, \\
&\min_{\mathbf{A}_T} ||\mathbf{T} - \mathbf{D}_T \mathbf{A}_T||_F^2 + \sigma ||\mathbf{A}_T||_1 + \gamma ||\mathbf{A}_T - \mathbf{U}_T^{-1} \mathbf{P}_V||_F^2.
\end{aligned} \tag{C.3}$$

Here Equation C.3 can be further rewritten as the form of standard sparse coding with respect to $\mathbf{A}_V$ and $\mathbf{A}_T$, as follows:

$$\begin{aligned}
&\min_{\mathbf{A}_V} ||\bar{\mathbf{V}} - \bar{\mathbf{D}}_V \mathbf{A}_V||_F^2 + \sigma ||\mathbf{A}_V||_1, \\
&\min_{\mathbf{A}_T} ||\bar{\mathbf{T}} - \bar{\mathbf{D}}_T \mathbf{A}_T||_F^2 + \sigma ||\mathbf{A}_T||_1,
\end{aligned} \tag{C.4}$$

where $\bar{\mathbf{D}}_V = \begin{bmatrix} \mathbf{V} \\ \sqrt{\gamma} \mathbf{U}_V^{-1} \mathbf{P}_T \end{bmatrix}$, $\bar{\mathbf{D}}_T = \begin{bmatrix} \mathbf{T} \\ \sqrt{\gamma} \mathbf{U}_T^{-1} \mathbf{P}_V \end{bmatrix}$, and $\bar{\mathbf{D}}_V = \begin{bmatrix} \mathbf{D}_V \\ \sqrt{\gamma} \mathbf{I} \end{bmatrix}$, $\bar{\mathbf{D}}_T = \begin{bmatrix} \mathbf{D}_T \\ \sqrt{\gamma} \mathbf{I} \end{bmatrix}$. In practice, we use the open source tool of SPAMS [77] to compute $\mathbf{A}_V$ and $\mathbf{A}_T$.

C.1.3 Updating $\mathbf{U}_V$ and $\mathbf{U}_T$

Using aforementioned strategy we can derive the following formulation to update dictionaries $\mathbf{U}_V$ and $\mathbf{U}_T$:

$$\begin{aligned}
&\min_{\mathbf{U}_V^{-1}} \|\mathbf{A}_V - \mathbf{U}_V^{-1}\mathbf{P}_T\|_F^2 + \mu\|\mathbf{U}_V^{-1}\|_F^2, \\
&\min_{\mathbf{U}_T^{-1}} \|\mathbf{A}_T - \mathbf{U}_T^{-1}\mathbf{P}_V\|_F^2 + \mu\|\mathbf{U}_T^{-1}\|_F^2,
\end{aligned} \tag{C.5}$$

which are standard ridge regression problems with respect to $\mathbf{U}_V$ and $\mathbf{U}_T$. Therefore, we can derive the close-form solutions as:

$$\begin{aligned}
\mathbf{U}_V^{-1} &= \mathbf{A}_V\mathbf{P}_T^\top[\mathbf{P}_T\mathbf{P}_T^\top + (\mu/\gamma)I]^{-1}, \\
\mathbf{U}_T^{-1} &= \mathbf{A}_T\mathbf{P}_V^\top[\mathbf{P}_V\mathbf{P}_V^\top + (\mu/\gamma)I]^{-1}.
\end{aligned} \tag{C.6}$$

In summary, we can update these parameters according to the derived solutions iteratively until the object function of Equation C.1 is converged. Once we get the optimized values of sparse representations $\mathbf{A}_V$ and $\mathbf{A}_T$ of the two modalities, we can use them as the input for the latter coupled feature mapping procedure.

C.2 Parameter learning for coupled feature mapping

After obtaining the sparse representations $\mathbf{A}_V$ and $\mathbf{A}_T$ for the two modalities from the coupled dictionary learning procedure above, given the keyword matrix $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n] \in \mathbb{R}^{c\times n}$, the coupled feature mapping procedure aims to learn two projection matrices $\mathbf{W}_V \in \mathbb{R}^{k_1\times c}$ and $\mathbf{W}_T \in \mathbb{R}^{k_2\times c}$, which map the sparse representations $\mathbf{A}_V$ and $\mathbf{A}_T$ of the two modalities into the common space defined by the keywords.

We leverage the coupled linear regression method and develop two schemes for the coupled feature mapping procedure: the simple scheme and the advanced scheme.

C.2.1 Simple scheme

For the simple scheme, we adopt the ridge regression method to ensure that the projection errors from the sparse representations of each modality to the keyword space are minimized. The generic minimization problem of the simple scheme is derived as:

$$\min_{\mathbf{W}_V,\mathbf{W}_T} \|\mathbf{W}_V^\top\mathbf{A}_V - \mathbf{Y}\|_F^2 + \|\mathbf{W}_T^\top\mathbf{A}_T - \mathbf{Y}\|_F^2 + \lambda(\|\mathbf{W}_V\|_F^2 + \|\mathbf{W}_T\|_F^2), \tag{C.7}$$

where λ is the regularization parameter. The analytical solutions of $\mathbf{W}_V$ and $\mathbf{W}_T$ can be derived as:

$$\begin{aligned} \mathbf{W}_V &= \mathbf{Y}\mathbf{A}_V^\top(\mathbf{A}_V\mathbf{A}_V^\top + \lambda I)^{-1}, \\ \mathbf{W}_T &= \mathbf{Y}\mathbf{A}_T^\top(\mathbf{A}_T\mathbf{A}_T^\top + \lambda I)^{-1}. \end{aligned} \tag{C.8}$$

C.2.2 Advanced scheme

For the advanced scheme, we follow the previous coupled feature mapping method [74] to further incorporate coupled feature selection. Thus, we have the similar objective function as in [74] as:

$$\min_{\mathbf{W}_V,\mathbf{W}_T} \frac{1}{2}(\|\mathbf{W}_V^\top\mathbf{A}_V - \mathbf{Y}\|_F^2 + \|\mathbf{W}_T^\top\mathbf{A}_T - \mathbf{Y}\|_F^2) + \lambda_1(\|\mathbf{W}_V\|_{21} + \|\mathbf{W}_T\|_{21}) + \lambda_2\|[\mathbf{W}_V^\top\mathbf{A}_V \ \ \mathbf{W}_T^\top\mathbf{A}_T]\|_*,$$

where the L_{21} norm $\|\cdot\|_{21}$ is used to select features from coupled feature space simultaneously, and the trace norm $\|\cdot\|_*$ encodes the correlations of different modalities with a low-rank constraint. Note that in Equation C.9, we use sparse representations $\mathbf{A}_V$ and $\mathbf{A}_T$ for feature mapping, while in [74] the original features of $\mathbf{A}_V$ and $\mathbf{A}_T$ are utilized.

Following the proposal in [74], the Equation C.9 can be rewritten as:

$$\begin{aligned} \min_{\mathbf{W}_V,\mathbf{W}_T} \min_{\mathbf{S}\geq 0} \quad & \frac{1}{2}(\|\mathbf{W}_V^\top\mathbf{A}_V - \mathbf{Y}\|_F^2 + \|\mathbf{W}_T^\top\mathbf{A}_T - \mathbf{Y}\|_F^2) + \lambda_1(\|\mathbf{W}_V\|_{21} + \|\mathbf{W}_T\|_{21}) \\ & + \frac{\lambda_2}{2}tr([\mathbf{A}_V^\top\mathbf{W}_V \ \ \mathbf{A}_T^\top\mathbf{W}_T]^\top\mathbf{S}^{-1}[\mathbf{A}_V^\top\mathbf{W}_V \ \ \mathbf{A}_T^\top\mathbf{W}_T]) + \frac{\lambda_2}{2}tr(\mathbf{S}), \end{aligned} \tag{C.9}$$

where $\mathbf{S}$ is the approximately variational formulation for the trace norm in Equation C.9, as:

$$\mathbf{S} = (\mathbf{A}_V^\top\mathbf{W}_V\mathbf{W}_V^\top\mathbf{A}_V + \mathbf{A}_T^\top\mathbf{W}_T\mathbf{W}_T^\top\mathbf{A}_T + \delta\mathbf{I})^{\frac{1}{2}}. \tag{C.10}$$

Furthermore, the Equation C.9 can be reformulated by decomposing the L_{21} norm, as:

$$\begin{aligned} \min_{\mathbf{W}_V,\mathbf{W}_T} \min_{\mathbf{S}\geq 0} \quad & \frac{1}{2}(\|\mathbf{W}_V^\top\mathbf{A}_V - \mathbf{Y}\|_F^2 + \|\mathbf{W}_T^\top\mathbf{A}_T - \mathbf{Y}\|_F^2) + \lambda_1[tr(\mathbf{W}_V^\top\mathbf{P}\mathbf{W}_V) + tr(\mathbf{W}_T^\top\mathbf{Q}\mathbf{W}_T)] \\ & + \frac{\lambda_2}{2}[tr(\mathbf{A}_V^\top\mathbf{W}_V\mathbf{S}^{-1}\mathbf{A}_V^\top\mathbf{W}_V) + tr(\mathbf{A}_T^\top\mathbf{W}_T\mathbf{S}^{-1}\mathbf{A}_T^\top\mathbf{W}_T)] + \frac{\lambda_2}{2}tr(\mathbf{S}). \end{aligned} \tag{C.11}$$

Given $\mathbf{S}$, optimizing the objective function in Equation C.11 for $\mathbf{W}_V$ and $\mathbf{W}_T$ can be decomposed to two problems of half-quadratic analysis:

$$\begin{aligned}
&\min_{\mathbf{W}_V} \frac{1}{2}||\mathbf{A}_V^\top \mathbf{W}_V - \mathbf{Y}||_F^2 + \lambda_1 tr(\mathbf{W}_V^\top \mathbf{P} \mathbf{W}_V) + \frac{\lambda_2}{2} tr(\mathbf{A}_V^\top \mathbf{W}_V \mathbf{S}^{-1} \mathbf{A}_V^\top \mathbf{W}_V),\\
&\min_{\mathbf{W}_T} \frac{1}{2}||\mathbf{A}_T^\top \mathbf{W}_T - \mathbf{Y}||_F^2 + \lambda_1 tr(\mathbf{W}_T^\top \mathbf{Q} \mathbf{W}_T) + \frac{\lambda_2}{2} tr(\mathbf{A}_T^\top \mathbf{W}_T \mathbf{S}^{-1} \mathbf{A}_T^\top \mathbf{W}_T),
\end{aligned} \tag{C.12}$$

where $\mathbf{p} = \frac{1}{2\sqrt{||\mathbf{w}_v^i||_2^2 + \varepsilon}}$ and $\mathbf{q} = \frac{1}{2\sqrt{||\mathbf{w}_t^i||_2^2 + \varepsilon}}$ are auxiliary vectors of the two L_{21} norms respectively, and ε is a small constant value for smoothing. In addition, $\mathbf{P} = Diag(\mathbf{p})$ and $\mathbf{Q} = Diag(\mathbf{q})$.

Therefore, with regarding to $\mathbf{W}_V$ and $\mathbf{W}_T$, their optimal solutions of Equation C.12 can be computed by solving the following two linear system problems:

$$\begin{aligned}
&(\mathbf{A}_V \mathbf{A}_V^\top + \lambda_1 \mathbf{P} + \lambda_2 \mathbf{A}_V \mathbf{S}^{-1} \mathbf{A}_V^\top) \mathbf{W}_V = \mathbf{A}_V \mathbf{Y},\\
&(\mathbf{A}_T \mathbf{A}_T^\top + \lambda_1 \mathbf{Q} + \lambda_2 \mathbf{A}_T \mathbf{S}^{-1} \mathbf{A}_T^\top) \mathbf{W}_T = \mathbf{A}_T \mathbf{Y}.
\end{aligned} \tag{C.13}$$

In summary, the alternating minimization procedure for Equation C.9 is summarized in Algorithm 5.

Algorithm 5: Alternating minimization for the advanced scheme

```
Input: Sparse representations $\mathbf{A}_V$ and $\mathbf{A}_T$, keyword matrix $\mathbf{Y}$, parameters $\lambda_1$, $\lambda_2$, $\delta$ and $\varepsilon$.
Output: Projection matrices $\mathbf{W}_V$ and $\mathbf{W}_T$.
Initialize $\{\mathbf{W}_V$ and $\mathbf{W}_T\}$ as zero matrices for iteration $t = 0$.
repeat
    $t = t + 1$.
    Compute $\mathbf{S}^t$ according to Equation C.10.
    Compute $\mathbf{P}^t$ and $\mathbf{Q}^t$.
    Compute $\mathbf{W}_V^t$ and $\mathbf{W}_T^t$ according to the solutions of linear system problems in
    Equation C.13.
until The value of Equation C.9 converges.
```