Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and Information Systems

School of Computing and Information Systems

2-2011

Mining Social Images with Distance Metric Learning for Automated Image Tagging

Pengcheng WU
Nanyang Technological University

Steven C. H. HOI
Singapore Management University, chhoi@smu.edu.sg

Peilin ZHAO
Nanyang Technological University

Ying HE
Nanyang Technological University

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

Part of the Databases and Information Systems Commons, and the Numerical Analysis and Scientific Computing Commons

Citation

WU, Pengcheng; HOI, Steven C. H.; ZHAO, Peilin; and HE, Ying. Mining Social Images with Distance Metric Learning for Automated Image Tagging. (2011). *WSDM '11: Proceedings of the 4th International Conference on Web Search and Data Mining: Hong Kong, China, February 9-12*. 197-206.
Available at: https://ink.library.smu.edu.sg/sis_research/2352

This Conference Proceeding Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylds@smu.edu.sg.

Mining Social Images with Distance Metric Learning for Automated Image Tagging

Pengcheng Wu, Steven C.H. Hoi, Peilin Zhao, Ying He
School of Computer Engineering
Nanyang Technological University
Singapore, 639798
{wupe0003,chhoi,zhao0106,yhe}@ntu.edu.sg

ABSTRACT

With the popularity of various social media applications, massive social images associated with high quality tags have been made available in many social media web sites nowadays. Mining social images on the web has become an emerging important research topic in web search and data mining. In this paper, we propose a machine learning framework for mining social images and investigate its application to automated image tagging. To effectively discover knowledge from social images that are often associated with multimodal contents (including visual images and textual tags), we propose a novel Unified Distance Metric Learning (UDML) scheme, which not only exploits both visual and textual contents of social images, but also effectively unifies both inductive and transductive metric learning techniques in a systematic learning framework. We further develop an efficient stochastic gradient descent algorithm for solving the UDML optimization task and prove the convergence of the algorithm. By applying the proposed technique to the automated image tagging task in our experiments, we demonstrate that our technique is empirically effective and promising for mining social images towards some real applications.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Experimentation

Keywords

Social images, distance metric learning, inductive learning, transductive learning, automated image tagging

1. INTRODUCTION

Along with the popularity of digital cameras and high quality mobile devices as well as the advances of internet technologies, users can easily upload their images and photos over the World Wide Web (WWW). Moreover, with the great success of social networks and social web sites recently, web users have been highly motivated to share their images with friends and public that allows other users to tag and comment on their image collections. Nowadays, web images, especially social images, which are often of high quality and rich user-generated contents including good quality user tags, are playing a more and more important role in WWW. Mining web and social images thus has become an emerging popular research topic in web search and data mining area.

In this paper, we investigate a machine learning scheme for mining social images, and its application to resolve a challenging task, automated image tagging, which is important and beneficial to many web and multimedia applications. The goal of an automated image tagging task is to assign a set of semantic labels or tags to a novel image with some pre-trained image recognition models. The traditional approach typically has two steps: (1) representing images by extracting visual features [16], and (2) pre-training recognition models by building classification models from a collection of manually-labeled training data [2]. In literature, numerous studies have been devoted to automated image annotation and object recognition tasks [15, 20].

Despite being studied extensively, regular image annotation approaches, which usually work well on small-sized testbeds with high quality labels, often fail to handle large-scale real photo tagging applications. One major challenge faced by large-scale photo annotation is primarily due to the well-known semantic gap between low-level features and high-level semantic concepts. Besides, it is also expensive and time-consuming to collect a large set of manually-labeled training data by conventional methods. Hence, it has become an urgent need to develop new paradigms for automated image tagging.

In this paper, we investigate an emerging retrieval-based annotation paradigm [24, 26] for automated photo tagging by mining massive social images freely available on the web. Unlike traditional web images, social images often contain tags and rich user-generated contents, which offer a new opportunity to resolve some long-standing challenges in multimedia, for instance the semantic gap. The idea of the retrieval-based paradigm [24] is to first retrieve a set of k most similar images for a test photo from the social image repository, and then to assign the test photo with a set of t most relevant tags associated with the set of k retrieved social images. Figure 1 shows an example of tagging a novel image by the proposed technique in this paper.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
WSDM'11, February 9–12, 2011, Hong Kong, China.
Copyright 2011 ACM 978-1-4503-0493-1/11/02 ...$10.00.

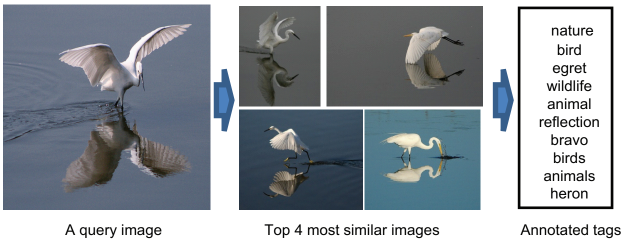


Figure 1: Example of automatically tagging a novel image by the proposed technique in this paper.

The crux of the retrieval-based photo tagging paradigm is to effectively identify and retrieve a set of top k similar photos from social image database, which mainly relies on two key components: (1) a feature representation scheme to extract salient visual features, and (2) a distance measure scheme to compute distances for extracted features. This paper focuses on techniques to tackle the second challenge. In particular, by considering features that are represented in vector space, our goal is to study an effective distance measure scheme for improving the retrieval performance. To this end, we propose to apply Distance Metric Learning (DML) techniques to resolve this challenge.

DML has been actively studied in machine learning and data mining community, which usually assumes the learning task is provided with explicit side information given in the form of either class labels [10] or pairwise constraints [1] where each pairwise constraint indicates whether two examples are similar ("must-link") or dissimilar ("cannot-link"). Although DML has been extensively studied [10, 1, 13, 4], it is not straightforward to directly apply regular DML techniques as side information is not explicitly available in our learning task. Moreover, regular DML techniques may not be very effective for solving our task, primarily because social image data are often associated with rich contents (including textual and visual contents) that differ from typical single-view data used in regular DML methods.

To this end, this paper presents a novel unified distance metric learning (UDML) framework, which aims to learn effective metrics from implicit side information of social images towards the application of automated photo tagging. Unlike the regular DML techniques, the proposed UDML technique aims to optimize metrics by integrating both textual and visual contents smoothly in a unified framework. Besides, this framework also unifies both inductive and transductive metric learning approaches together in a systematic approach.

As a summary, the key contributions of this paper include: (1) a novel unified distance metric learning framework to learn distance metrics from implicit side information of social images; (2) an effective algorithm to solve the unified distance metric learning task; (3) a new solution by applying the UDML technique to a real application of automated photo tagging; (4) extensive experiments to compare our method with a number of state-of-the-art DML algorithms, in which encouraging results were obtained.

The rest of this paper is organized as follows. Section 2 introduces the retrieval-based annotation framework of mining social images for automated photo tagging. Section 3 presents the proposed unified distance metric learning framework and an effective algorithm to learn distance metrics from social images of multi-modal contents. Section 4 gives experimental results and discussions. Section 5 briefly reviews some related work, and Section 6 concludes this work.

2. MINING SOCIAL IMAGES FOR AUTOMATED IMAGE TAGGING

We first introduce a generic retrieval-based annotation framework for mining web/social images for automated image tagging [24], followed by the discussion of some open challenges in this framework.

2.1 Overview of Retrieval based Annotation

The basic assumption of a retrieval based annotation approach towards automated photo tagging is that similar/identical images would share the common/similar tags. Based on this assumption, one can attack automated photo tagging, a long-standing challenging in multimedia and computer vision, by mining a large collection of web/social images. Specifically, Figure 2 shows a diagram to illustrate the process of a retrieval based annotation scheme for mining social images to tackle the automated photo tagging task.

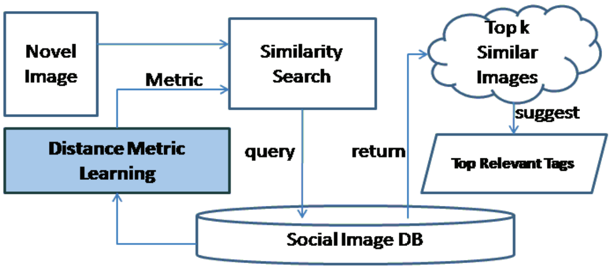


Figure 2: Block diagram for illustrating the process of a retrieval-based annotation approach by mining social images with distance metric learning

As shown in the figure, when a novel image is given, we first conduct a similarity search step to find a subset of top k images most similar to the novel image from a social image database. Once obtaining a subset of top k similar images from the similarity search process, the next step is to summarize the tags associated with these similar images, and recommend the top relevant tags by some approach (e.g. ranking the associated tags by majority voting).

2.2 Open Research Challenges

Despite the simplicity for the above retrieval-based annotation framework, there are some open research challenges that have yet to be solved effectively. One important step of the whole framework is how to perform the similarity search process effectively, which is a key process that significantly affects the performance of the subsequent annotation process. In general, the similarity search process requires a distance metric for distance measure in the retrieval process. Hence, distance metric learning to find an optimal metric is an open challenge in this framework. Besides, there are also some other open issues, such as the efficiency and scalability of the retrieval process that often requires an effective indexing scheme, and an effective tag ranking scheme that ranks the tags associated with the top k similar images. In this paper, we focus on addressing the first challenge of distance metric learning for improving the retrieval process in this framework.

3. UNIFIED DISTANCE METRIC LEARNING FOR MINING SOCIAL IMAGES

3.1 Overview

In this section, we present a novel machine learning approach to learn distance metrics from social images to resolve the automated photo tagging task. Our goal is to attack the challenge of the similarity search process by optimizing the distance metrics from social images.

In particular, given a novel image $\mathbf{x}_q \in \mathbb{R}^d$ that is represented in a d-dimensional space, for any image $\mathbf{x} \in \mathbb{R}^d$ in the database, we consider a family of Mahalanobis distances $d_M(\mathbf{x}_q, \mathbf{x})$ to calculate distance between $\mathbf{x}_q$ and $\mathbf{x}$ as follows:

$$d_M(\mathbf{x}_q, \mathbf{x}) = \|\mathbf{x}_q - \mathbf{x}\|_M^2 = (\mathbf{x}_q - \mathbf{x})^\top M(\mathbf{x}_q - \mathbf{x}) \quad (1)$$

where $M \in \mathbb{R}^{d\times d}$ is any pre-defined positive semi-definite matrix that parameterizes the Mahalanobis distance. For example, if we choose M as an identity matrix, the above formula reduces to (square) Euclidean distance.

Therefore, the goal of distance metric learning is to learn an optimal matrix M from training data such as it can effectively tackle the similarity search process of the retrieval-based photo annotation paradigm. However, unlike conventional DML tasks where side information is often explicitly given a prior (in the forms of either pairwise constraints or class labels), in our problem, side information is only implicitly available in the social image collection.

To facilitate the distance metric learning task, in the following, we first present a simple approach to generate explicit side information from a collection of social images. With the side information, we further present a unified distance metric learning approach that can combine both textual and visual contents smoothly in a systematic learning framework.

3.2 Generation of Side Information

As no explicit side information is given for our DML task, the first step before DML is to derive side information from a collection of N social images $\mathcal{S} = \{s_i | i = 1, \ldots, N\}$. In general, a social image contains rich user-generated contents, including visual images, textual tags, comments, rating, etc. To simplify the discussion, in our approach, we assume each social image s_i consists of two components: visual image and textual tags, i.e., $s_i = (\mathbf{x}_i, \mathbf{t}_i)$, where $\mathbf{x}_i$ denotes the visual features extracted from the social image, and $\mathbf{t}_i$ denotes the tag vector of the social image.

The basic idea of our side information generation approach is to extract side information in terms of "triplet" format, i.e., $(\mathbf{x}, \mathbf{x}_+, \mathbf{x}_-)$, which indicates that image $\mathbf{x}$ and image $\mathbf{x}_+$ are similar/relevant to each other, while image $\mathbf{x}$ and image $\mathbf{x}_-$ are dissimilar/irrelevant. To this purpose, we randomly pick a social image from the collection of social images as a query image $\mathbf{q}_i = (\mathbf{x}_{q_i}, \mathbf{t}_{q_i})$, and then generate a subset of triplets $\mathcal{P}_i$ with respect to $\mathbf{q}_i$ as follows:

$$\mathcal{P}_i = \{(\mathbf{x}_{q_i}, \mathbf{x}_{k_i^+}, \mathbf{x}_{k_i^-}) | \forall \mathbf{x}_{k_i^+} \in \mathcal{R}_k(\mathbf{t}_{q_i}), \forall \mathbf{x}_{k_i^-} \in \bar{\mathcal{R}}_k(\mathbf{t}_{q_i})\} \quad (2)$$

where $\mathcal{R}_k(\mathbf{t}_{q_i})$ denotes the set of top k social images that are most relevant with respect to a text-based query $\mathbf{t}_{q_i}$, and similarly $\bar{\mathcal{R}}_k(\mathbf{t}_{q_i})$ denotes the set of top k least relevant social images. Finally, we repeat the generation process N_Q times, and form a set of side information $\{\mathcal{P}_i, i = 1 \ldots, N_Q\}$, which will be used as input training data for our distance metric learning task.

3.3 Formulation

We now present the formulation of the proposed distance metric learning method. The basic idea of the proposed unified DML method is to combine the ideas of both inductive and transductive learning principles for DML in order to fuse both textual and visual contents of social images smoothly in a systematic optimization framework. Below we first present two kinds of different objective functions for our DML tasks, respectively, and then show the final formulation of the unified distance metric learning method.

3.3.1 Inductive metric learning by maximizing margin

First of all, following the similar idea of large margin learning principle [25], we consider the following inductive learning formulation for optimizing distance metric from side information:

$$\min_{M\succeq 0} \qquad J_1(M) \triangleq \frac{1}{N_p} \sum_{i=1}^{N_Q} \sum_{\forall(\mathbf{x}_{q_i}, \mathbf{x}_{k_i^+}, \mathbf{x}_{k_i^-}) \in \mathcal{P}_i} \ell(M; (\mathbf{x}_{q_i}, \mathbf{x}_{k_i^+}, \mathbf{x}_{k_i^-})) \quad (3)$$

where N_p denotes the total number of triplets, and ℓ is a typical hinge loss function defined as:

$$\begin{aligned} &\ell(M; (\mathbf{x}_{q_i}, \mathbf{x}_{k_i^+}, \mathbf{x}_{k_i^-})) \\ &\qquad = \max\{0, 1 - [d_M(\mathbf{x}_{q_i}, \mathbf{x}_{k_i^-}) - d_M(\mathbf{x}_{q_i}, \mathbf{x}_{k_i^+})]\} \quad (4) \end{aligned}$$

The above loss function indicates that we should optimize the metric by penalizing (1) large distance between two similar images, and (2) small distance between two dissimilar images. This clearly reflects the intuition of large margin learning principle.

3.3.2 Transductive fusion of text and visual contents

Second, we also consider a transductive approach to integrate with both textual tags and visual contents of social images for learning distance metric as follows:

$$\min_{M\succeq 0} \qquad J_2(M) \triangleq \sum_{i,j} w_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_M^2 \quad (5)$$

where w_{ij} is the cosine similarity between the two textual tag vectors of the two social images, i.e., $w_{ij} = \cos(\mathbf{t}_i, \mathbf{t}_j)$. The above formulation indicates that if two social images share similar textual tags, we expect to force their visual distance to be small.

We can further simplify the above formulation. In particular, we note that each valid metric M can be decomposed into a linear mapping $A : \mathbb{R}^d \mapsto \mathbb{R}^r$ where $A = [\mathbf{a}_1, \ldots, \mathbf{a}_r] \in \mathbb{R}^{d\times r}$ such that $M = AA^\top$. With this representation, we can rewrite the distance measure as:

$$\begin{aligned} d_M(\mathbf{x}_q, \mathbf{x}) &= \|\mathbf{x}_q - \mathbf{x}\|_M^2 = (\mathbf{x}_q - \mathbf{x})^\top AA^\top(\mathbf{x}_q - \mathbf{x}) \\ &= \|A^\top(\mathbf{x}_q - \mathbf{x})\| \qquad (6) \end{aligned}$$

As a result, we can rewrite the formulation of the above objective function as:

$$\begin{aligned} J_2(M) &= \sum_{i,j} w_{ij}\|\mathbf{x}_i - \mathbf{x}_j\|_M^2 = \sum_{k=1}^{r} \mathbf{a}_k^\top X(D-W)X^\top \mathbf{a}_k \\ &= \sum_{k=1}^{r} \mathbf{a}_k^\top XLX^\top \mathbf{a}_k = tr(A^\top XLX^\top A) \\ &= tr(XLX^\top AA^\top) = tr(XLX^\top M) \end{aligned} \quad (7)$$

where D is a diagonal matrix whose diagonal elements are the sums of the row entries of matrix W, and $L = D - W$ is known as the Laplacian matrix.

3.3.3 Unified distance metric learning

Finally, by unifying both the inductive formulation and the transductive formulation together, we can achieve the following formulation of unified distance metric learning:

$$\min_{M \succeq 0} \quad J(M) \triangleq \frac{1}{2} tr(M^\top M) + CJ_1(M) + \lambda J_2(M) \quad (8)$$

where C and λ are parameters to trade off between inductive and transductive objective functions, and the first regularization term is introduced to penalize the norm of the metric to prevent some values of the metric dominating all the other elements.

Since each part of the objective function is convex, the above formulation of the unified distance metric learning (UDML) problem is a convex optimization task. More exactly, it belongs to semi-definite programming (SDP), which in general can be resolved by some existing convex optimization techniques. Since it is often highly intensive to solve an SDP task by a generic SDP solver, it is not efficient and scalable to directly apply existing SDP solvers for our application. To develop an efficient and scalable solution, below we present an efficient algorithm to resolve the optimization of the unified distance metric learning.

3.4 Algorithm

The key challenge of the UDML optimization is to optimize the metric with respect to the inductive maximal margin learning term, which is related to a large set of triplets that can be potentially huge since a large amount of side information is available in practice. To overcome this challenge, we propose a stochastic gradient descent algorithm that resolves the optimization iteratively by randomly sampling a subset of active triplets at every optimization iteration.

Formally, for a particular iteration, we randomly choose a subset of triplets from the whole set of triplets:

$$\mathcal{A}_t = \{(\mathbf{x}_{q_i}, \mathbf{x}_{k_i^+}, \mathbf{x}_{k_i^-})|\ i \in [Q]\} \quad (9)$$

which satisfies $|A_t| = N_a \ll N_p$. Further, from $\mathcal{A}_t$, we can derive an active set of triplets whose values of the loss function are nonzero, i.e.,

$$\mathcal{A}_t^+ = \{(\mathbf{x}_{q_i}, \mathbf{x}_{k_i^+}, \mathbf{x}_{k_i^-}) \in A_t|\ \ell(M; (\mathbf{x}_{q_i}, \mathbf{x}_{k_i^+}, \mathbf{x}_{k_i^-})) > 0\}. \quad (10)$$

Based on the set of triplets $\mathcal{A}_t$, we can rewrite the objective function as follows:

$$\begin{aligned} J(M; \mathcal{A}_t) &= \frac{1}{2} tr(M^\top M) + \lambda tr(XLX^\top M) \\ &+ \frac{C}{N_a} \sum_{(\mathbf{x}_{q_i}, \mathbf{x}_{k_i^+}, \mathbf{x}_{k_i^-}) \in \mathcal{A}_t} \ell(M; (\mathbf{x}_{q_i}, \mathbf{x}_{k_i^+}, \mathbf{x}_{k_i^-})) \end{aligned} \quad (11)$$

```
Algorithm 1: The Stochastic Gradient Descent Algorithm for Unified Distance Metric Learning. (UDML)
INPUT: parameter $C, \lambda$ and the number of iterations $T$
PROCEDURE
Choose $M_1$ s.t. $\|M_1\| \leq \sqrt{2C}$
for $t = 1, 2, \ldots, T$ do
    Randomly choose a set $\mathcal{A}_t$, s.t. $|\mathcal{A}_t| = N_a$
    Set $\mathcal{A}_t^+ = \{(\mathbf{x}_{k_i^+}, \mathbf{x}_{k_i^-}) \in \mathcal{A}_t|\ \ell(M_t; (\mathbf{x}_{q_i}, \mathbf{x}_{k_i^+}, \mathbf{x}_{k_i^-})) > 0\}$
    Set a learning rate $\eta_t = \frac{1}{t}$
    Set $M_{t+1/2} = M_t - \eta_t[\partial J(M_t; \mathcal{A}_t)/\partial M]$
    Set $M_{t+1} = \min\{1, \frac{\sqrt{2C}}{\|M_{t+1/2}\|_F}\}M_{t+1/2}$
end for
Project $M_{T+1}^{psd} = PSD(M_{T+1})$
OUTPUT: $M_{T+1}^{psd}$
END
```

Figure 3: The Stochastic Gradient Descent Algorithm for Unified Distance Metric Learning.

To minimize the objective function, we adopt the gradient descent approach, which needs to compute the sub-gradient of the above objective function as follows:

$$\begin{aligned} &\frac{\partial J(M; \mathcal{A}_t)}{\partial M} \\ &= M + \lambda XLX^\top \\ &- \frac{C}{N_a} \sum_{(\mathbf{x}_{q_i}, \mathbf{x}_{k_i^+}, \mathbf{x}_{k_i^-}) \in A_t^+} [(\mathbf{x}_{k_i^-} - \mathbf{x}_{q_i})(\mathbf{x}_{k_i^-} - \mathbf{x}_{q_i})^\top - (\mathbf{x}_{k_i^+} - \mathbf{x}_{q_i})(\mathbf{x}_{k_i^+} - \mathbf{x}_{q_i})^\top] \end{aligned}$$

We repeat the above stochastic gradient descent approach until the algorithm converges. Figure 3 summarizes the details of the proposed stochastic gradient descent algorithm for UDML. In the algorithm, at the end of each gradient descent step, we perform a scaling process by forcing the solution $M_{t+1} \leq \sqrt{2C}$ below:

$$M_{t+1} = \min\Big\{1, \frac{\sqrt{2C}}{\|M_{t+1/2}\|_F}\Big\} M_{t+1/2} \quad (12)$$

The detailed reason for the above scaling step will be discussed in the subsequent analysis section (referred to Lemma 1 and 2). Besides, to further improve efficiency, we do not force the PSD constraint at every gradient descent step. At the end of the entire algorithm, to ensure that the final solution M is a valid metric, we perform a projection of the final matrix M_{T+1} onto the PSD domain: $M_{T+1}^{psd} = PSD(M_{T+1})$.

3.5 Convergence Analysis

Below we theoretically analyze the convergence of the proposed algorithm. Our proofs and analysis mainly follow the principles and theory of online convex optimization [11, 18].

Firstly, we present a lemma, which provides an upper bound for the norm of the optimal solution M, and explains why performing the scaling step in the algorithm.

LEMMA 1. *The optimal solution of optimization problem (8) is in the convex close set $\mathcal{B}_M = \{M| \|M\|_F \leq \sqrt{2C}\}$, where $\|\cdot\|_F$ denotes the Frobenius norm.*

PROOF. Let us denote by M^* the optimal solution. Using the fact that $J(M^*; X) \leq J(0; X)$, we thus have

$$\frac{1}{2}\|M^*\|_F^2 = \frac{1}{2} tr((M^*)^\top M^*) \leq J(M^*; X) \leq J(0; X) = C$$

The second inequality is guaranteed by $tr(XLX^\top M) = \sum_{i,j} w_{ij}\|\mathbf{x}_i - \mathbf{x}_j\|_M^2 \geq 0$ and $\ell(M; (\mathbf{x}_{k_i^+}, \mathbf{x}_{k_i^-})) \geq 0$. □

Before presenting the theorem, we first introduce an important lemma that generalizes a result from [11].

Lemma 2. *Let $g_1, ..., g_T$ be a sequence of σ-strongly convex functions w.r.t the function $\frac{1}{2}\|\cdot\|_F^2$. Let $\mathcal{B}$ be a closed convex set and define $\Pi_{\mathcal{B}}(M) = \arg\min_{M' \in \mathcal{B}} \|M - M'\|_F$. Let $M_1, \ldots, M_{T+1}$ be a sequence of matrices such that $M_1 \in \mathcal{B}$ and for $t \geq 1$, $M_{t+1} = \Pi_B(M_t - \eta_t \nabla_t)$, where ∇_t is a subgradient of g_t at M_t and $\eta_t = 1/(\sigma t)$. Assume that for all t, $\|\nabla_t\| \leq G$. Then for all $M \in \mathcal{B}$ we have*

$$\frac{1}{T}\sum_{t=1}^{T} g_t(M_t) \leq \frac{1}{T}\sum_{t=1}^{T} g_t(M) + \frac{G^2(1 + ln(T))}{2\sigma T} \tag{13}$$

Based on Lemma 2, we are now ready to bound the average of the stochastic objective function $J(M_t; \mathcal{A}_t)$.

Theorem 1. *Assume that $\|\mathbf{x}_{q_i}\| \leq R_1 \; \forall i \in [Q]$, $\|x_j\|_2 \leq R_1 \; \forall j \in [N]$, and W is normalized such that $\sum_{i,j} W_{ij} = 1$. Let M^* be the optimal solution. Then, for $T \geq 3$ we have*

$$\frac{1}{T}\sum_{t=1}^{T} J(M_t; \mathcal{A}_t) \leq \frac{1}{T}\sum_{t=1}^{T} J(M^*; \mathcal{A}_t) + \frac{R^2 \ln(T)}{T} \tag{14}$$

where $R = \sqrt{2C} + (4\lambda + 8C)R_1^2$.

Proof. To simplify our notation we use the shorthand $J_t(M) = J(M; \mathcal{A}_t)$. The update of the algorithm can be rewritten as $M_{t+1} = \Pi_{\mathcal{B}_M}(M_t - \eta_t \nabla_t)$, where $\mathcal{B}_M$ is defined in Lemma 1 and $\nabla_t = \partial J(M_t; \mathcal{A}_t)/\partial M$. Thus, we only need to prove the conditions in Lemma 2 are satisfied.

Since g_t is the sum of a 1-strongly convex function ($\frac{1}{2}\|M\|_F^2$) and a convex function, it is also 1-strongly convex.

Next we would bound the norm of the sub-gradient:

$$\begin{aligned}\|\nabla_t\|_F \leq\ & \|M_t\|_F + \lambda\|XLX^\top\|_F \\ & + \frac{C}{N_a}\sum_{(\mathbf{x}_{q_i}, \mathbf{x}_{k_i^+}, \mathbf{x}_{k_i^-}) \in A_t^+} \|(\mathbf{x}_{k_i^-} - \mathbf{x}_{q_i})(\mathbf{x}_{k_i^-} - \mathbf{x}_{q_i})^\top\|_F \\ & + \frac{C}{N_a}\sum_{(\mathbf{x}_{q_i}, \mathbf{x}_{k_i^+}, \mathbf{x}_{k_i^-}) \in A_t^+} \|(\mathbf{x}_{k_i^+} - \mathbf{x}_{q_i})(\mathbf{x}_{k_i^+} - \mathbf{x}_{q_i})^\top\|_F\end{aligned}$$

Firstly, $\|M_t\|_F \leq \sqrt{2C}$ according to the design of the algorithm. And then, we would provide an upper bound on $\|XLX^\top\|_F$. Before proving the bound, we note that

$$\alpha^\top XLX^\top \alpha = \sum_{ij} w_{ij}(\alpha^\top \mathbf{x}_i - \alpha^\top \mathbf{x}_j)^2 \geq 0 \; \forall \alpha \in \mathbb{R}^{m \times 1}$$

Thus, $XLX^\top$ is positive semi-definite. We thus have

$$\begin{aligned}& \|XLX^\top\|_F \\ & = \sqrt{tr(XLX^\top XLX^\top)} \leq \sqrt{tr(XLX^\top)^2} = tr(XLX^\top) \\ & = \sum_{ij} w_{ij}\|x_i - x_j\|^2 \leq \sum_{ij} w_{ij}(\|x_i\| + \|x_j\|)^2 \leq \sum_{ij} w_{ij}4R_1^2 \\ & = 4R_1^2\end{aligned}$$

where the first inequality holds because $tr(AB) \leq tr(A)tr(B)$, when A and B are positive semi-definite matrices of the same order. Furthermore, we have

$$\begin{aligned}& \|(\mathbf{x} - \mathbf{x}_q)(\mathbf{x} - \mathbf{x}_q)^\top\|_F = (\mathbf{x} - \mathbf{x}_q)^\top(\mathbf{x} - \mathbf{x}_q) = \|\mathbf{x} - \mathbf{x}_q\|_2^2 \\ & \leq \quad (\|\mathbf{x}\|_2 + \|\mathbf{x}_q\|_2)^2 \leq 4R_1^2\end{aligned} \tag{15}$$

As a result,

$$\|\nabla_t\|_F \leq \sqrt{2C} + 4\lambda R_1^2 + 8CR_1^2 := R \tag{16}$$

In addition, it is easy to see that, when $T \geq 3$

$$\frac{1 + \ln(T)}{2T} \leq \frac{\ln(T)}{T} \tag{17}$$

Combining all of these results, the proof is done. □

Since Theorem 1 only provides a comparison for the functions $J(M; A_t)$, now the following theorem will provide a comparison between $J(M)$. For convenience, we denote $\mathcal{A}_i^j = (\mathcal{A}_i, \ldots, \mathcal{A}_j)$. Then we have the following theorem:

Theorem 2. *Assume that the conditions stated in Theorem 1 hold and for all t, $\mathcal{A}_t$ is chosen i.i.d from the set of all triplets. Let r be an integer picked uniformly at random from $[T]$. Then*

$$\mathbb{E}_{\mathcal{A}_1^T}\mathbb{E}_r[J(M_r)] \leq J(M^*) + \frac{R^2 \ln(T)}{T} \tag{18}$$

The proof of Theorem 2 can be found in the Appendix. Theorem 2 states that, in expectation, the SGD algorithm will converge quickly. The next theorem will provide a bound of the objective function in probability.

Theorem 3. *Assume that the conditions stated in Theorem 2 holds. Let $\delta \in (0, 1)$. Then, with probability of at least $1 - \delta$ over the choices of $\mathcal{A}_1, ..., \mathcal{A}_T$ and the index r, we have the following bound:*

$$J(M_r) \leq J(M^*) + \frac{R^2 \ln(T)}{\delta T} \tag{19}$$

Proof. Let $Z := J(M_r) - J(M^*) \geq 0$ be a random variable. Thus, from Markov inequality $P(Z \geq a) \leq \mathbb{E}[Z]/a$ and $P(Z \leq a) + P(Z \geq a) = 1$, we have $P(Z \leq a) = 1 - P(Z \geq a) \geq 1 - \frac{\mathbb{E}(Z)}{a}$. As a result, we have

$$P(Z \leq \frac{R^2 \ln(T)}{\delta T}) \geq 1 - \frac{\mathbb{E}(Z)}{\frac{R^2 \ln(T)}{\delta T}} \geq 1 - \frac{\frac{R^2 \ln(T)}{T}}{\frac{R^2 \ln(T)}{\delta T}} = 1 - \delta \tag{20}$$

In the above, we apply Theorem 2, i.e., $\mathbb{E}(Z) \leq \frac{R^2 \ln(T)}{T}$. □

We now use the above theorem to analyze the convergence of the last matrix M_{T+1}. We can treat $T + 1$ as a random index drawn from $\{1, \ldots, \hat{T}\}$, where $\hat{T} > T + 1$. Since M_{T+1} does not depend on $M_{T+2}, \ldots, M_{\hat{T}}$, we can terminate the algorithm after T iterations and return M_{T+1}. Using Theorem 3, we know that

$$J(M_{T+1}) - J(M^*) \leq \frac{R \ln(\hat{T})}{\delta \hat{T}} \leq \frac{R \ln(T)}{\delta T} \tag{21}$$

where the last inequality holds as $\frac{\ln(T)}{T}$ decreases in $[3, +\infty)$.

3.6 Tagging Images with Optimized Metrics

Finally, we briefly describe the process of automated image tagging by applying the optimized metric M learned by applying distance metric learning techniques.

In particular, given a novel unlabeled image $\mathbf{x}_q$ for tagging, the first step is to conduct similarity search to retrieve a subset of similar images with tags from social image database. In our approach, we retrieve a set of k-nearest neighbors of the query image, i.e.,

$$\mathcal{N}_k(\mathbf{x}_q) = \{i \in [1, \ldots, n] | \mathbf{x}_i \in \text{kNN} - \text{List}(\mathbf{x}_q)\}, \tag{22}$$

where n is the total number of images in the social image repository, and the $\mathrm{kNN-List}$ is found by measuring the distances with the optimized metric M, i.e., $\|\mathbf{x}_q - \mathbf{x}_i\|_M^2$.

With the set of similar social images $\mathcal{N}_k(\mathbf{x}_q)$, the next step is to perform a tag ranking by adapting the idea of majority voting. Specifically, we define a set of candidate tags $\mathcal{T}_w$ as:

$$\mathcal{T}_w = \bigcup_{i \in \mathcal{N}_k(\mathbf{x}_q)} \mathcal{T}_i \tag{23}$$

where $\mathcal{T}_i$ represents the set of tags associated with social image $\mathbf{s}_i$. Further, we calculate the frequency of each candidate tag $w \in \mathcal{T}_w$, denoted as $f(w)$, which indicates the number of times the tag is associated with the k social images. Finally, we conduct the automated image tagging by following the intuition: to assign the query image with a tag of *high* frequency and *small* average distance. Specifically, we tag the novel image $\mathbf{x}_q$ by incrementally adding a tag using the following approach:

$$w^* = \arg\max_{w \in \mathcal{T}_w \wedge w \notin \mathcal{T}_q} \frac{f(w)}{avg\_d_M(\mathbf{x}_q, w) + \kappa} \tag{24}$$

where $avg\_d_M(\mathbf{x}_q, w)$ represents the average distance (with optimized metric M) between the query image and those candidate social images that have tag w, and κ is a smoothing parameter fixed to 1 in our experiments.

4. EXPERIMENTS

In this section we discuss our experiments for evaluating the performance of our unified distance metric learning approach for automated photo tagging.

4.1 Experimental Testbed

We conducted our experiments on a real-world social images testbed, which consists of 200,000 images crawled from Flickr website. These social images contain rich information, including user-generated tags and other metadata.

To simplify the experiments, we employed tags and visual features to represent a social image. For text information, we sorted all tags in the dataset by their frequencies, the top 100,000 of which were used to construct a dictionary, and the others were abandoned. To improve the quality of annotation, we manually removed some clearly noisy tags from the dictionary by applying a list of stopwords. We adopted each image's associated tags in this dictionary as its text features. For visual features, we extracted four kinds of effective and compact visual features, including grid color moment, local binary pattern, Gabor wavelet texture, and edge direction histogram. In total, a 297-dimensional feature vector was used to represent each image. The set of features had been used in some previous CBIR studies [30, 12, 26].

We randomly split the 200,000 images data set into 3 sets: *training* set, *test* set and *database* set.

- The *training* set is used as input training data for distance metric learning. We randomly sampled 15,000 images with their associated metadata from the whole dataset. These social images were used to generate side information for DML.
- The *test* set is adopted to test the tagging performance. In particular, we randomly chose 2,000 images as query images and treated their associated tags in the dictionary as the annotation ground truth directly.
- The *database* set consists of the rest 183,000 images. It is used as social image repository for the retrieval-based tagging process.

4.2 Compared Methods

To evaluate the performance of the proposed UDML method, we compared it extensively with two major categories of metric learning techniques. One is to learn metrics with explicit class labels, such as NCA[10], LMNN[25]. The other is to learn metrics from pairwise constraints, such as RCA[1], DCA[13],OASIS[4]. Specifically, the compared schemes include:

- **Euclidean**: the baseline method.
- **DCA**[13]: Discriminative Component Analysis, which leans a linear projection using only equivalent constraints.
- **RCA**[1]: Relevance Component Analysis that learns a linear projection using only equivalent constraints.
- **ITML**[5]: Information Theoretic Metric Learning which trains the metric with the goal that minimizes the differential relative entropy between two multivariate Gaussians under constraints on the distance function.
- **RDML**[19]: Regularized Distance Metric Learning that adopts the correlation between users' relevance feedback and low-level image features.
- **LMNN**[25]: Large Margin Nearest Neighbor whose goal is that k-nearest neighbors always belong to the same class while examples from different classes are separated by a large margin.
- **NCA**[10]: Neighbourhood Components Analysis which maximizes a stochastic variant of the leave-one-out kNN score.
- **OASIS**[4]: Online Algorithm for Scalable Image Similarity learning, which is an online dual approach based on the passive-aggressive algorithm and is to learn a bilinear similarity measure over sparse representations.
- **LRML**[12]: Laplacian Regularized Metric Learning whose goal is to leverage the unlabeled data information and to ensure metric learning smoothness through a regularization learning framework.
- **pRCA**[26]: probabilistic Relevant Component Analysis, which learns an optimal metric from probabilistic side information.
- **UDML**: the proposed Unified Distance Metric Learning method.

4.3 Experimental Setup

As no explicit side information is available in the experiments, in order to apply DML techniques, we applied the proposed side information generation approach described in Section 3.2 to derive side information from the training set of social images. In particular, we randomly chose one social image as query from the dataset, and generated a set of 100 triplets for each query. We ran the random sampling process 1000 times, and totally generated $100,000$ triplets as side information for our experiments. The same set of side information was used/converted to other appropriate formats (e.g. chunklets) for other DML methods. Regarding parameter settings, we simply fixed tradeoff parameters $\lambda = 1$, $C = 10000$, the size of active set $N_a = 100$, and the total number of iterations $T = 1000$ for the proposed UDML algorithm.

To evaluate the performance of DML approaches for automated image tagging, we applied the retrieval-based tagging procedure as described in Section 3.6. Specifically, a query image was chosen from the test set, and then used to search similar images from the database set by applying the optimized distance metrics. In particular, a set of top k (we set $k = 30$ in default) images were retrieved, and then top t tags ranked by equation (24) were suggested to tag the query image. The annotation performance was then evaluated based on the relevance of the top ranked tags ranging from top 1 to top 10 tags. The standard average precision (AP) and average recall (AR) were employed as the performance metrics.

4.4 Experimental Results

Figures 4 and 5 show the average precision and recall results achieved by different DML methods, where the horizontal axis denotes the number of the top t tags annotated. Figure 5 shows a comparison of the precision-recall curves by different DML methods. For all these comparisons, we fixed the number of similar images $k = 30$ in the annotation procedure. From these experimental results, we can draw several observations as follows.

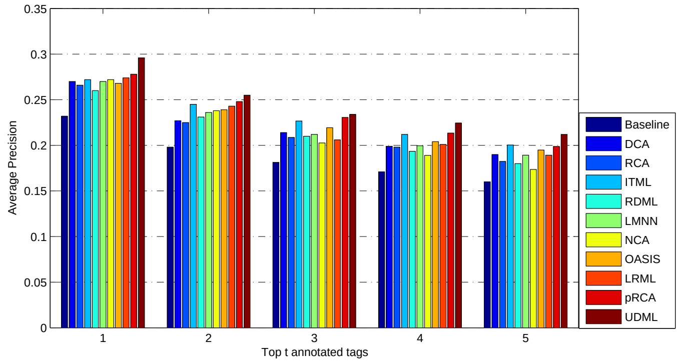


Figure 4: Average precision at top t annotated tags

First of all, we found that all the DML based approaches performed significantly better than the baseline tagging approach that simply adopts Euclidean distance. This shows that the approach of applying DML to optimize the metrics is beneficial and important for the retrieval-based image tagging task.

Second, among all the compared methods, we observed that the proposed UDML method considerably surpassed all the other approaches for most cases. For instance, in terms of the average precision performance, UDML achieved about 29.6%, while the baseline approach only had 23.2% and the results of other DML methods ranged from 26.0% to 27.8%.

Lastly, despite the above encouraging improvements, we noticed that the average precision values of all the compared methods are still quite low. The possible reasons include (1) the difficulty of retrieving the similar social images without a very large-scale database; (2) the associated tags of some social image are quite noisy that could degrade the tagging performance; and (3) the optimized distance measure may be still not perfect to return the most similar social images relevant to the query image, which shows that there might be still a large room to study more effective distance metric learning techniques in the future.

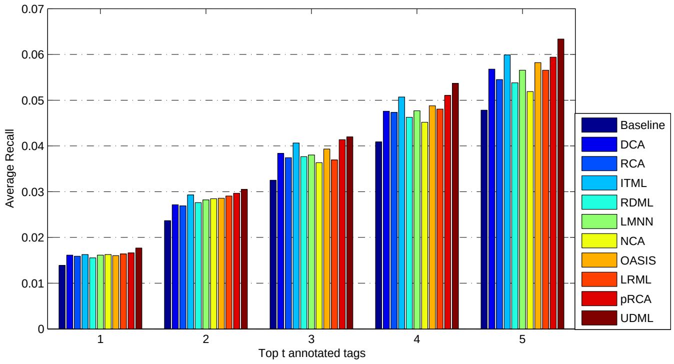


Figure 5: Average recall at top t annotated tags

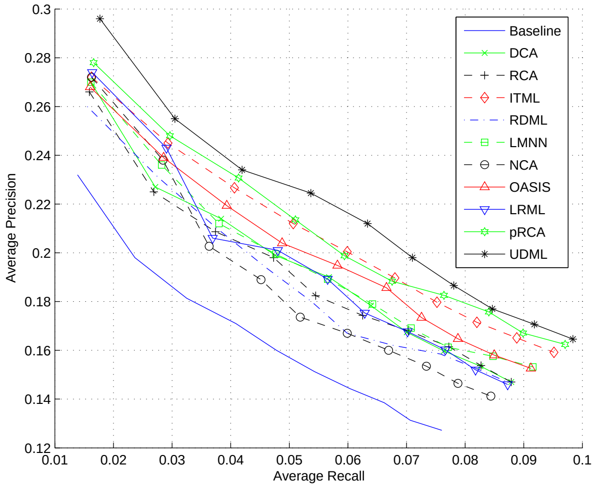


Figure 6: The precision-recall curves

4.5 Evaluation of Varied k Values

Figure 7 shows the performance of UDML at top t tags by varying k, the number of top retrieved similar images from 10 to 60. From the results, we observed that k affects the annotation performance. In particular, when k is about 40 to 50, the proposed method achieved the best average precision. This is reasonable because if k is too small, some relevant images may not be retrieved, while if k is too large, lots of irrelevant images could be retrieved, leading to engage many noisy tags in the list of candidate tags. Both of the above situations could degrade the annotation performance.

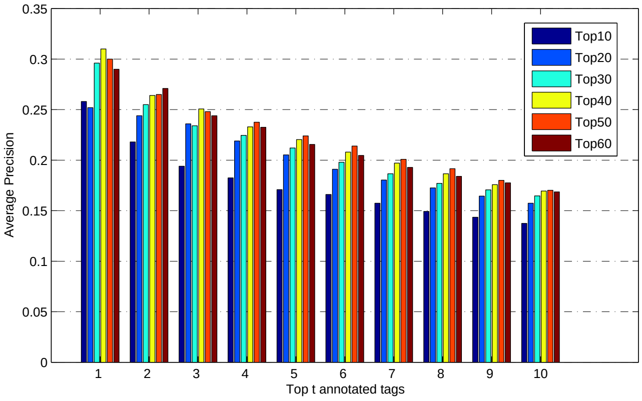


Figure 7: Comparisons of average precision under different top k similar images used

4.6 Comparison of Qualitative Performance

Our last experiment is to examine the qualitative tagging performance achieved by different DML methods for automated image tagging tasks. To achieve this purpose, we randomly chose several images from the test set, and applied a number of different DML methods to annotate them using the proposed retrieval-based annotation approach. Figure 8 shows the top 10 annotated tags under different metrics. The relevant tags are marked in "blue" font. The results show that UDML often achieves better quality among all the 11 approaches.

5. RELATED WORK

Our work is related to several groups of research, including social image search and mining with applications to automated image/photo annotation and object recognition [17, 21, 28], and distance metric learning (DML) studies [27, 1, 19, 5], etc. Due to limited space, we briefly review some most representative and relevant studies below.

5.1 Web/Social Image Mining

Our work is related to web/social image search and mining as well as automated image annotation. Image annotation has been actively studied over the past decade in multimedia community. Conventional approaches often train some classification models, e.g. SVM [7], from a collection of human-labeled training data for a set of predefined semantic concept/object categories [2, 3, 6, 23].

Recently, there is a surge of emerging interests in exploring web photo repositories for image annotation. A promising approach is the retrieval-based (or termed "search-based") paradigm [17, 24, 21, 22]. Russell et al. [17] built a large collection of web images with ground truth labels for helping object recognition research. Wang et al. [24] proposed a fast search-based approach for image annotation by some efficient hashing technique. Torralba et al. [21] proposed efficient image search and scene matching techniques for exploring a large-scale web image repository. These work usually concerned more on fast indexing and search techniques, while we focus on learning more effective distance metrics. Finally, our work mainly follow the recent study of exploring social images for automated photo tagging [26], but we propose a new and empirically more effective method.

5.2 Distance Metric Learning

In literature, DML has been actively studied in two major domains. One is to learn metrics with explicit class labels, which are often studied for classification tasks [14, 8, 9, 25, 29]. The other is to learn metrics from pairwise constraints that are mainly used for clustering and retrieval [1, 13, 27]. Moreover, from machine learning perspective, most existing DML studies belong to inductive learning methods, although there are some recent studies that have attempted to explore transductive learning for DML [12].

Our study is quite different from existing DML approaches in data mining and machine learning. Unlike most existing DML methods that assume explicit side information is provided in the form of either class labels or pairwise constraints, in our DML problem, no explicit side information is directly given for the learning task. Hence, in our study, we actually learn metrics from implicit side information, which is hidden in the rich contents of social image data in our application. Finally, we unify both inductive and transductive learning principles in a systematic framework.

6. CONCLUSIONS

This paper investigated a machine learning approach for mining social images towards automated image tagging applications. In particular, we proposed a novel unified distance metric learning (UDML) method, which learns metrics from implicit side information hidden in massive social images on the web. Unlike regular metric learning studies, the proposed UDML method fully exploits both textual and visual contents for learning an effective metric in a unified and systematic learning framework. To handle a real large scale web mining problem, we proposed an efficient stochastic gradient descent algorithm and showed its convergence property by providing theoretical proofs. Experimental results on a real social image testbed show that our UDML method is effective and promising for mining social images for solving automated image tagging applications. In future work, we plan to enlarge the social image database, and investigate more sophisticated tag ranking techniques for improving the annotation performance.

Acknowledgement

The work was supported by Singapore National Research Foundation Interactive Digital Media R&D Program under research grant NRF2008IDM-IDM004-006.

Appendix: Proof of Theorem 2

Proof. Taking expectation of the inequality of Theorem 1 leads to the following:

$$\mathbb{E}_{A_1^T}[\frac{1}{T}\sum_{t=1}^{T}J(M_t;A_t)] \leq \mathbb{E}_{A_1^T}[\frac{1}{T}\sum_{t=1}^{T}J(M^*;A_t)] + \frac{R^2\ln(T)}{T}$$

Since M^* does not depend on the choice of triplets, we have

$$\begin{aligned}\mathbb{E}_{A_1^T}[\frac{1}{T}\sum_{t=1}^{T}J(M^*;A_t)] &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{A_1^T}J(M^*;A_t) \\ &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{A_t}J(M^*;A_t) = J(M^*)\end{aligned}$$

$$\begin{aligned}\mathbb{E}_{A_1^T}[\frac{1}{T}\sum_{t=1}^{T} J(M_t;A_t)] &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{A_1^T} J(M_t;A_t)\\ &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{A_1^t} J(M_t;A_t)\end{aligned}$$

Recall the law of total expectation implies that for any two random variables X, Y, $\mathbb{E}_X[J(X)] = \mathbb{E}_Y\mathbb{E}_X[J(X)|Y]$, thus

$$\begin{aligned}\mathbb{E}_{A_1^t}[J(M_t;A_t)] &= \mathbb{E}_{A_1^{t-1}}[\mathbb{E}_{A_1^t}[J(M_t;A_t)|A_1^{t-1}]]\\ &= \mathbb{E}_{A_1^{t-1}}[J(M_t)] = \mathbb{E}_{A_1^T}[J(M_t)]\end{aligned}$$

Putting the above together, we can obtain

$$\mathbb{E}_{A_1^T}[\frac{1}{T}\sum_{t=1}^{T} J(M_t;A_t)] = \mathbb{E}_{A_1^T}[\frac{1}{T}\sum_{t=1}^{T} J(M_t)]$$

Furthermore, since $\mathbb{E}_r[J(M_r)] = \frac{1}{T}\sum_{t=1}^{T} J(M_t)$, combining all the above leads to complete our proof. □

7. REFERENCES

[1] A. Bar-hillel and D. Weinshall. Learning a Mahalanobis Metric from Equivalence Constraints. *Journal of Machine Learning Research*, 6:937–965, 2005.

[2] G. Carneiro, A. B. Chan, P. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Tran. PAMI*, pages 394–410, 2006.

[3] G. Carneiro and N. Vasconcelos. Formulating semantic image annotation as a supervised learning problem. In *IEEE CVPR*, pages 163–168, 2005.

[4] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large Scale Online Learning of Image Similarity Through Ranking. *Journal of Machine Learning Research*, 11:1109–1135, 2010.

[5] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216, 2007.

[6] P. Duygulu, K. Barnard, J. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, pages 97–112, 2002.

[7] J. Fan, Y. Gao, and H. Luo. Multi-level annotation of natural scenes using dominant image components and semantic concepts. In *ACM Multimedia*, pages 540–547, 2004.

[8] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Elsevier, 1990.

[9] A. Globerson and S. Roweis. Metric learning by collapsing classes. In *NIPS'05*, 2005.

[10] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. *In Advances in Neural Information Processing Systems*, 17, 2005.

[11] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Mach. Learn.*, 69(2-3):169–192, 2007.

[12] S. C. Hoi. Semi-supervised distance metric learning for Collaborative Image Retrieval. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, June 2008.

[13] S. C. Hoi, W. Liu, M. R. Lyu, and W.-Y. Ma. Learning distance metrics with contextual constraints for image retrieval. In *CVPR*, New York, US, June 17–22 2006.

[14] G. H. J. Goldberger, S. Roweis and R. Salakhutdinov. Neighbourhood components analysis. In *NIPS17*, 2005.

[15] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *SIGIR'03*, pages 119–126, Toronto, Canada, 2003.

[16] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004.

[17] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *Int. J. Comput. Vision*, 77(1-3):157–173, 2008.

[18] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *ICML*, pages 807–814, Corvalis, Oregon, USA, 2007.

[19] L. Si, R. Jin, S. C. H. Hoi, and M. R. Lyu. Collaborative image retrieval via regularized metric learning. *ACM Multimedia Systems Journal*, 12(1):34–44, 2006.

[20] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. PAMI*, 22(12):1349–1380, 2000.

[21] A. Torralba, Y. Weiss, and R. Fergus. Small codes and large databases of images for object recognition. In *CVPR*, 2008.

[22] C. Wang, L. Zhang, and H.-J. Zhang. Learning to reduce the semantic gap in web image retrieval and annotation. In *SIGIR'08*, pages 355–362, Singapore, 2008.

[23] M. Wang, X. Zhou, and T.-S. Chua. Automatic image annotation via local multi-label classification. In *ACM CIVR*, pages 17–26, 2008.

[24] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma. Annosearch: Image auto-annotation by search. In *CVPR'06*, pages 1483–1490, 2006.

[25] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2006.

[26] L. Wu, S. Hoi, R. Jin, J. Zhu, and N. Yu. Distance metric learning from uncertain side information with application to automated photo tagging. In *ACM Multimedia*, pages 135–144. 2009.

[27] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. In *NIPS2002*, 2002.

[28] R. Yan, A. Natsev, and M. Campbell. A learning-based hybrid tagging and browsing approach for efficient manual image annotation. In *IEEE CVPR'08*, 2008.

[29] L. Yang, R. Jin, R. Sukthankar, and Y. Liu. An efficient algorithm for local distance metric learning. In *AAAI*, 2006.

[30] J. Zhu, S. C. Hoi, M. R. Lyu, and S. Yan. Near-duplicate keyframe retrieval by nonrigid image matching. *Proceeding of the 16th ACM international conference on Multimedia - MM '08*, page 41, 2008.

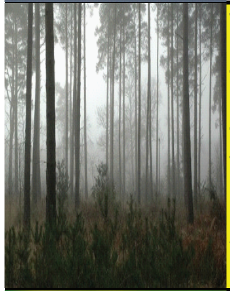
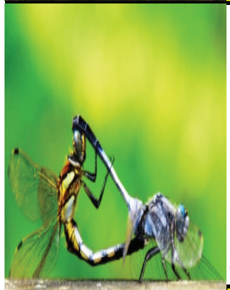
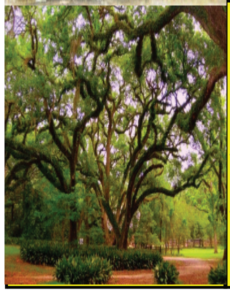
	Baseline	DCA	RCA	ITML	RDML	LMNN	NCA	OASIS	LRML	pRCA	UDML
	nature	**nature**	**nature**	**nature**	**nature**	**nature**	**nature**	**nature**	**nature**	**nature**	**nature**
	blue	**bird**	**wildlife**	**reflection**	blue	**bird**	blue	**wildlife**	**bird**	**reflection**	**bird**
	reflection	bravo	**bird**	**bird**	**bird**	**reflection**	**bird**	**bird**	**wildlife**	**water**	**egret**
	bird	**reflection**	**reflection**	bravo	**reflection**	**wildlife**	beach	**reflection**	**reflection**	**animal**	**wildlife**
	animal	**birds**	specanimal	**wildlife**	**water**	canon	specanimal	**water**	**birds**	**bird**	**animal**
	water	specnature	animalkingdc	**water**	**wildlife**	**animal**	**reflection**	bravo	**animal**	**egret**	**reflection**
	canon	**animal**	**animal**	**birds**	mountain	specanimal	**wildlife**	specanimal	**egret**	bravo	bravo
	egret	**egret**	**water**	**animal**	specanimal	**egret**	**animal**	specnature	specanimal	animalkingdc	**birds**
	specanimal	specanimal	**birds**	**egret**	**egret**	blue	plane	**animal**	specnature	mountain	**animals**
	wildlife	**water**	**egret**	animalkingdc	specnature	animalkingdc	bravo	**egret**	japan	specanimal	**heron**
	**forest**	**trees**	**fog**	**forest**	**forest**	**trees**	**forest**	**forest**	**fog**	**forest**	**forest**
	nature	**forest**	**trees**	**fog**	**nature**	**forest**	**fog**	**trees**	**forest**	**nature**	**fog**
	trees	**fog**	**forest**	**trees**	**trees**	**tree**	**tree**	**fog**	**trees**	**trees**	**nature**
	light	**light**	bravo	**tree**	**fog**	**nature**	**nature**	**tree**	**mist**	explore	**tree**
	tree	explore	**mist**	**nature**	**tree**	**fog**	**trees**	**nature**	**tree**	**tree**	**trees**
	fog	morning	**tree**	**landscape**	**light**	black	**mist**	germany	**nature**	**fog**	**wood**
	explore	**tree**	**light**	morning	soe	canon	**light**	morning	alberi	**landscape**	**mist**
	bravo	**nature**	**landscape**	germany	**woods**	**light**	canon	**landscape**	explore	bravo	germany
	woods	soe	**nature**	**mist**	art	winter	morning	**wood**	misty	**mist**	**landscape**
	landscape	**landscape**	morning	wood	vision	white	**wood**	winter	**park**	**light**	morning
	**nature**	**macro**	**macro**	**nature**	**nature**	**nature**	**macro**	**green**	**nature**	**macro**	**green**
	macro	**nature**	**green**	**macro**	**macro**	**macro**	**nature**	**macro**	**green**	**green**	**macro**
	green	**insect**	**nature**	**green**	**green**	bird	**green**	**nature**	**macro**	**nature**	**nature**
	insect	butterfly	bug	**insect**	canon	**insect**	canon	canon	canon	**insect**	bug
	canon	bokeh	**insect**	soe	butterfly	butterfly	**insect**	bug	bird	leaf	**insect**
	bravo	canon	leaf	dof	bird	best	leaf	eos	light	bravo	canon
	bug	soe	rain	**dragonfly**	**insect**	bravo	bravo	**insect**	landscape	canon	dof
	bird	**green**	water	photo	bravo	bug	butterfly	excapturema	leaf	plant	soe
	specanimal	bug	dof	explore	specanimal	bokeh	eos	flower	water	photos	bokeh
	flower	flower	canon	butterfly	flower	taiwan	bird	**dragonfly**	dof	butterfly	**dragonfly**
	**garden**	**nature**	autumn	**trees**	**garden**	**nature**	**garden**	**trees**	**nature**	**green**	**trees**
	old	explore	**forest**	**nature**	old	**trees**	**park**	**nature**	**trees**	**trees**	**nature**
	vancouver	canon	fall	**forest**	fall	**green**	**green**	spring	fall	river	old
	park	**trees**	**trees**	autumn	**park**	old	japanese	old	autumn	autumn	**tree**
	spring	river	**nature**	**green**	vancouver	**garden**	winner	**forest**	**forest**	**garden**	**park**
	fall	usa	perfect	river	**tree**	explore	bravo	autumn	soe	soe	bridge
	explore	travel	love	bravo	favoritegarde	perfect	**tree**	soe	explore	**nature**	**green**
	favoritegarde	art	**tree**	landscape	winner	landscape	canada	**park**	**tree**	canon	**leaves**
	canon	wood	**leaves**	fall	spring	canon	landscape	bravo	**leaves**	**leaves**	colors
	car	fall	waterfall	spring	**green**	**forest**	**nature**	japan	bravo	bravo	bravo

Figure 8: Examples showing the tagging results by 11 different methods. For each row, the first image is a test image and each following block shows top 10 tags annotated by one method. The correct tags are highlighted by blue color.