Tagging Personal Photos with Transfer Deep Learning

Jianlong Fu [1]*, Tao Mei [2], Kuiyuan Yang [2], Hanqing Lu [1], and Yong Rui [2]
[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
No. 95, Zhongguancun East Road, Beijing 100190, China
[2]Microsoft Research, No. 5, Dan Ling Street, Haidian District, Beijing 10080, China
[1]{jlfu, luhq}@nlpr.ia.ac.cn, [2]{tmei, kuyang, yongrui}@microsoft.com

ABSTRACT

The advent of mobile devices and media cloud services has led to the unprecedented growing of personal photo collections. One of the fundamental problems in managing the increasing number of photos is automatic image tagging. Existing research has predominantly focused on tagging general Web images with a well-labelled image database, e.g., ImageNet. However, they can only achieve limited success on personal photos due to the domain gaps between personal photos and Web images. These gaps originate from the differences in semantic distribution and visual appearance. To deal with these challenges, in this paper, we present a novel transfer deep learning approach to tag personal photos. Specifically, to solve the semantic distribution gap, we have designed an ontology consisting of a hierarchical vocabulary tailored for personal photos. This ontology is mined from $10,000$ active users in Flickr with 20 million photos and 2.7 million unique tags. To deal with the visual appearance gap, we discover the intermediate image representations and ontology priors by deep learning with bottom-up and top-down transfers across two domains, where Web images are the source domain and personal photos are the target. Moreover, we present two modes (single and batch-modes) in tagging and find that the batch-mode is highly effective to tag photo collections. We conducted personal photo tagging on 7,000 real personal photos and personal photo search on the MIT-Adobe FiveK photo dataset. The proposed tagging approach is able to achieve a performance gain of 12.8% and 4.5% in terms of NDCG@5, against the state-of-the-art hand-crafted feature-based and deep learning-based methods, respectively.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithm, Performance, Experimentation

* This work was performed when Jianlong Fu was visiting Microsoft Research as a research intern.

WWW 2015, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3469-3/15/05.
http://dx.doi.org/10.1145/2736277.2741112.

Figure 1: Compared with Web images, the user-provided tags of personal photos are more subjective and the visual appearances are more complex. Here the personal photos are collected from real users, while Web images are from ImageNet.

Keywords

Personal photo, image tagging, deep learning, transfer learning, ontology.

1. INTRODUCTION

Recent years have witnessed the emergence of mobile devices (e.g., smart phones, digital cameras, tablets, etc.) and cloud storage services. This has led to an unprecedented growth in the number of personal photos. People are taking photos using their smart devices every day and everywhere. One of the fundamental challenges to managing this ever-increasing number of photos is providing appropriate tags for each photo. Therefore, image tagging has become an active research topic in the last few years, in order to label images with human-friendly *concepts*[1]. Before we dive into the details of various image tagging techniques, we first define *personal photos* and introduce their characteristics.

We define *personal photos* as photos that are usually captured by amateur users with personal digital devices (e.g., smart phones, digital cameras, etc.). Compared with general Web images, personal photos have several unique properties: 1) Personal photos lack accurate text descriptions in general as users are unlikely to label their photos. 2) The semantic distribution of personal photos is only a subset of a general vocabulary of Web images. For example, some celebrities such as "Barack Obama" and specific terms such as "mammal" and "placental," are not likely to appear in personal photos of common users. Moreover, the semantic distribution in personal photos is typically biased toward the concepts related to "landscape," "family," and so on. 3) The appearance of personal photos is more complex due to occlusion, lighting variation, clut-

[1]"concept" and "tag" are considered as interchangeable terms, and we don't differentiate them in this paper.

tered background, and considerable camera motion. The tags, if there are any, are very subjective. This has led to the challenge of understanding personal photos. If we refer to Fig. 1, we can see that the Web images collected from ImageNet[2] are more likely to reflect a concept. In personal photos, however, various objects such as people, tree, grass, and sunlight appear in Fig. 1(b), and a blur effect and lighting variation appear in Fig. 1(c). Additionally, the photo labelled with "airport" may be very subjective. 4) There is rich metadata (e.g., time, geo-location) that can be exploited for analyzing personal photos.

The extensive research on image tagging can be divided into model-based and model-free approaches. The model-based approaches heavily rely on pre-trained classifiers with machine learning algorithms [17] [19] [27] [30], while the model-free approach propagates tags through the tagging behavior of visual neighbors [18][29]. The two streams of approaches both assume that there is a well-labelled image database (source domain) that has the same or at least a similar data distribution as the target domain, so that the well-labelled database can ensure good generalization abilities for both classifier training and tag propagation. However, the well-labelled database is hard to obtain in the domain of personal photos. On one hand, although some photo sharing websites such as Flickr[3] can provide a huge number of personal photos and user-contributed tags, this data is not appropriate as supervised information, as half of the user-contributed tags are noises to the image content [4]. On the other, although ImageNet [6] can provide accurate supervised information, the two significant gaps, i.e., the semantic distribution and visual appearance gaps between the two domains pose grand challenges to personal photo tagging.

To address the above issues, we present a novel transfer deep learning approach with ontology priors to tag personal photos. First, we have designed an ontology specific for personal photos from 10,000 active users in Flickr. Although previous methods (e.g., [3][12]) defined about a dozen common types as concepts (e.g., "beach fun," "ball games," and "wedding") in the domain of personal photos, they are not enough to comprehensively describe the variety of content. Furthermore, the correlations between different concepts have not been fully exploited in previous research.

Second, we propose reducing the visual appearance gap by applying deep learning techniques. Existing image tagging methods often leverage hand-crafted features, e.g., Scale-Invariant Feature Transform (SIFT) [20], GIST [23], Histogram of Oriented Gradients (HOG) [5], and so on. Based on these features, visual representation algorithms (e.g., Bag-of-Features [24]) have been proposed to describe image content and assign keywords. However, these hand-crafted descriptors are designed for general tasks to capture fixed visual patterns by pre-defined feature types and are not suitable for detecting some middle-level features that are shared and meaningful across two specific domains. With the recent success in many research areas [1], deep learning techniques have attracted increasing attention. This method can automatically learn hierarchical deep networks from raw pixels and produce adaptive middle-level feature representations for a specific task, especially in computer vision. For example, deep convolutional neural networks achieved a winning top-5 test error rate of 15.3%, compared to the 26.2% achieved by the second-best approach which combines scores from many classifiers trained by a set of hand-crafted features [14]. Another breakthrough was achieved in [15], where the algorithm automatically learned the concepts of cat faces and human bodies from unlabelled data. Motivated by such promising performances, we have proposed to discover middle-level feature abstractions from raw image pixels (i.e., called "bottom-up transfer" in this paper) and high-level ontology priors (i.e., called "top-down transfer" in this paper) using deep learning techniques. Both the middle-level and high-level representations are shared and meaningful across the two domains and thus can facilitate algorithms to reduce the visual differences between the two domains. As a result, a type of deep network learned from the source domain can be transferred to the target with good generalization abilities.

To the best of our knowledge, this paper is one of the first attempts to design a domain-specific ontology for personal photos and solve the tagging problem by transfer deep learning. The main contributions of this paper can be summarized as follows:

- We design a domain-specific ontology for personal photos, which is the most comprehensive ontology in this domain.
- We propose a novel transfer deep learning approach with ontology priors to effectively discover intermediate image representations from deep networks and ensure good generalization abilities across the two domains (Web images as the source domain and personal photos as the target).
- We propose two modes, including a single-mode and a batch-mode, to highly efficiently tag personal photos by leveraging the discriminative visual descriptors and rich metadata in the personal photos.

The rest of this paper is organized as follows. Section 2 describes related work. Section 3 first presents the ontology collection scheme for personal photos, then Section 4 formulates the transfer deep learning approach. Section 5 further describes two modes to efficiently tag personal photos. Section 6 provides empirical evaluations, followed by the conclusion in Section 7.

2. RELATED WORK

In this section, we briefly review research related to our approach in two categories. Image tagging aims to automatically assign concepts to images and has been studied intensively in the past decade, while transfer deep learning has drawn a great deal of attention recently with the success of deep learning techniques.

2.1 Image Tagging

A large body of work on image tagging proceeds along two dimensions, i.e., model-based and model-free approaches [7]. Model-based approaches heavily rely on pre-trained classifiers with machine learning algorithms. Tag ranking estimates initial relevance scores for tags by using probability density functions and performs a random walk process to refine the tagging results [19]. To address the problem of large-scale annotation of Web images, Visual synset applies multi-class one-vs-all linear Support Vector Machine models, which are learned from the automatically generated visual synsets of a large collection of Web images [27]. In [13], Ji *et al.* exploit both low-level visual features and high-level semantic context into a unified conditional random fields (CRF) model for solving the image tagging problem, which achieves significant improvement and more robustness results on two benchmarks. Although the model-based approach can achieve good performance, it may suffer from a limited vocabulary and less scalability on large-scale datasets.

In contrast to the model-based approach, the model-free approach [11][18][29] was developed to learn visual concepts from Web images. The intuition is that we can measure the relevance between an image and tags by searching and voting from its visual duplicates in

[2]www.image-net.org
[3]www.flickr.com

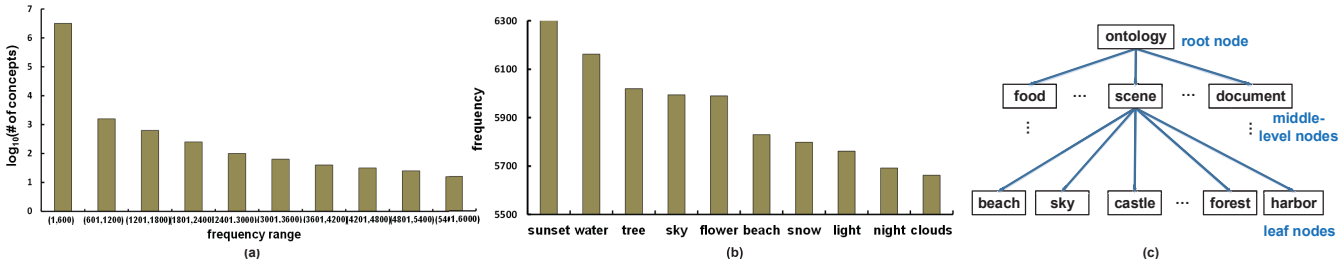

Figure 2: The statistics from 10K active users in Flickr and an illustration of the defined ontology. (a) Distribution of concepts on frequency ranges. (b) The top 10 concepts with frequencies. (c) An illustration of the curated ontology in the domain of personal photos. The concepts in the leaf nodes are the focus of the subsequent tagging procedure.

a well-labelled image database. Note that the model-free approach does not impose any model training. In [18], Li *et al.* leverage such techniques to measure social tag relevance by neighbor voting. In [29], Wang *et al.* adopt web-scale images to reformulate the image tagging problem as a search for semantically and visually similar images. In [11], Guillaumin *et al.* propose a novel nearest neighbor voting scheme that predicts tags by taking a weighted combination of the tag absence or presence among neighbors. However, the size of the well-labelled image database is limited in practice, and thus irrelevant tags may be propagated by mistake due to the well-known "semantic gap."

However, these works, in both dimensions, only leverage hand-crafted features. For example, in [18] and [29], global features (e.g., color correlogram, texture moment) are widely used as global similarity metrics. In [11] and [27], local shape descriptors and face signatures are employed, respectively. Although the approaches based on these hand-crafted features have achieved good results, it is still unclear how they should be selected to achieve better results for a desired task. Additionally, prior research for personal photos only focuses on event recognition for about 10 common types that have relatively small datasets [3][12]. A vocabulary of such size is less descriptive and comprehensive for a variety of personal photos. The goal of this paper is to show that a deep network can automatically learn image representations on a large domain-specific ontology for personal photos, which removes the need for engineered feature representations and can transfer to new tasks.

2.2 Transfer Deep Learning

Deep learning began emerging as a new area of machine learning research in 2006 [1]. The techniques developed from deep learning research exploit many layers of non-linear information processing for supervised or unsupervised feature extraction and for pattern analysis and classification, and they have shown many promising and exciting results [14].

Transfer deep learning targets the transfer of knowledge from a source domain to a target domain using deep learning algorithms. In [22], transfer learning problems are divided into two categories. One class tries to improve the classification performance of categories that have deficit training examples by discovering other similar categories and transferring knowledge among them (e.g., [25]). Other works aim to solve the different data distribution problems, so as to effectively transfer the knowledge learned from the source domain to the target (e.g., [22]).

Compared to the above methods, our task is more challenging, because there are few labelled images for all categories in the domain of personal photos. The approaches of [22] and [25], which all need labelled training images in the target domain, are unsuitable for our task. More similar to our work, Bengio *et al.* learn to extract a meaningful representation for each review text for different products using a deep learning approach in an unsupervised fashion [9]. In contrast to [9], which is applied to text applications, we need to handle the high-dimensional problem of images, which results in more difficulties. Furthermore, unlike [9], which only uses a bottom-up transfer, we propose both a bottom-up transfer and a top-down transfer in a unified framework to better reduce the domain gap.

3. ONTOLOGY FOR PERSONAL PHOTOS

We propose to design an ontology for personal photos, since the vocabulary of general Web images is often too large and not specific to personal photos. To obtain the ontology, we explored the semantic distributions in the domain of personal photos by mining frequent tags from active users in Flickr. Although Flickr cannot provide us with accurate tags to image content, we can mine a set of semantic words frequently used in personal photos by common users and collect them into this ontology. We collected more than $30,000$ users and selected about $10,000$ active users who had uploaded more than 500 photos in the most recent six months and with a registration time of more than two years. There are about 20 million photos and 2.7 million unique tags in total. For each user, we crawled all the photos and considered each user-contributed tag as a concept. After eliminating the stop words, we aggregated the concepts among all the users, ranked these concepts in the decreasing order by frequency and selected the top concepts whose frequencies were larger than $3,000$. Finally, we obtained 272 concepts, which forms the vocabulary in the subsequent image tagging. Fig. 2(a) and Fig. 2(b) show the distribution of concepts in terms of frequency and the top 10 concepts mined from the 10,000 active users, respectively.

To reflect the correlations between different concepts, we construct a three-layer ontology which consists of a hierarchical vocabulary. The top is a root node, followed by 20 human-curated middle-level nodes which represent various topics in personal photos. These topics are defined as: entertainment, social activity, daily routine, home, public places, people, plant, animal, transportation, clothing, regular items, furniture, kitchen ware, electronics, food, beverage, instrument, public facilities, scene, and document. The 272 concepts are considered as leaf nodes in the ontology and eventually grouped into the 20 middle-level nodes according to the word similarity in WordNet[4]. For example, as shown in Fig. 2(c), "beach," "sky," "castle," etc. are grouped into the topic of "scene." This ontology embeds the relationships of concepts, which can be considered as prior probabilities to improve tagging performance.

As we aim to conduct transfer learning from ImageNet to personal photos, the concepts in the source and target domains should be matched. However, there is a "label bias" between the two do-

[4]http://wordnet.princeton.edu/

mains [26]. To solve this problem, we calculated the word similarity between the 272 concepts and the ImageNet-22K labels by using WordNet. As a result, each concept in the domain of personal photos can be mapped to the closest label in the ImageNet. The images corresponding to these labels in the ImageNet form the training data in the source domain.

4. TRANSFER DEEP LEARNING WITH ONTOLOGY PRIORS

In this section, we propose a six-layer deep neural network for simultaneously harnessing the labelled images in the source domain and the unlabelled images in the target to reduce the visual appearance gap across the two domains. Fig. 3 shows an overview of the proposed approach, which consists of three components, i.e., (a) the training set, (b) the network of the transfer deep learning with ontology priors and (c) the ontology. In the testing stage, the personal photos that have been input in (d) can be annotated by the transferred network in (b) and the resultant tags are shown in (e).

First, the stacked convolutional autoencoders (CAES) are pretrained on both the source and target domains in an unsupervised manner, from which the shared deep feature representations can be discovered from raw pixels. A fine-tuning process is then implemented using the supervision in the source domain to give the network stronger discriminability. Note that although the fine-tuning is guided by the supervision in the source domain, it starts with the network parameters discovered across the two domains. Therefore, the fine-tuned network can still produce the shared feature representations across domains. Once the shared deep feature representations are fine-tuned, the top layer, i.e., a fully connected layer with ontology priors (FCO), is further trained. Since the shared deep feature representations and the ontology take effect across the two domains, the resultant parameters can be transferred to the target domain in the testing stage to obtain middle-level feature representations (a bottom-up transfer) and high-level confidence scores (a top-down transfer).

Let $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N] \in \mathbb{R}^{d \times N}$ be a set of N training data with d dimensions and $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N] \in \mathbb{R}^{K \times N}$ be the corresponding label matrix. Here the labels in the target domain are unknown, while in the source domain each label $\mathbf{y}_i$ is a K dimensional output for leaf nodes. The value of 1 is for the correct concept in the defined ontology and 0 otherwise. Let $\mathbf{W}$ denote the set of parameters of CAES (i.e., weights and biases), and $\mathbf{B}$ denote parameters of the top FCO layer, $\mathbf{B} \in \mathbb{R}^{K \times D}$. Here D represents the dimension of the transformed feature after CAES. Given $\mathbf{X}$, the parameter learning is determined by a conditional distribution over $\mathbf{Y}$, which can be formulated as:

$$\max_{\mathbf{W},\mathbf{B}} \{P(\mathbf{Y}|\mathbf{X})\} = \max_{\mathbf{W},\mathbf{B}} \left\{ \sum_{\mathbf{W},\mathbf{B}} P(\mathbf{Y}|\mathbf{X},\mathbf{W},\mathbf{B}) P(\mathbf{W}) P(\mathbf{B}) \right\}, \tag{1}$$

where $\mathbf{W}$ and $\mathbf{B}$ need to be optimized in the subsequent transfer deep learning procedures.

4.1 Deep Learning with Bottom-Up Transfer

To ensure good generalization abilities in transfer learning, a shared middle-level feature abstraction is first learned in an unsupervised pre-training and a supervised fine-tuning from both the source and target domains, in which $\mathbf{W}$ is optimized.

The autoencoder (AE) is one of the methods to build deep networks that is often used for learning an effective encoding of the original data without using supervised labels [1]. An autoencoder consists of an encoder function $f_{\mathbf{W}}(\mathbf{x}_i)$ and a decoder function $g_{\mathbf{W}'}(\mathbf{x}_i)$, where $\mathbf{x}_i$ is an input, $\mathbf{W}$ and $\mathbf{W}'$ are the parameters of the encoder and decoder, respectively. The fully connected AE is a basic form of an autoencoder. However, the fully connected AE ignores the high dimensionality and spatial structure of an image. Inspired by convolutional neural networks (CNN) [14], the convolutional autoencoder (CAE) has been proposed [21]. In contrast to the full connected AE, weights of the CAE are shared among all locations in the input. Therefore, CAE scales better to the realistic-sized high-dimensional images. For the input $\mathbf{x}_i$, the hidden representation of the j^{th} feature map is given by:

$$\mathbf{h}_j = f_{\mathbf{W}_j}(\mathbf{x}_i) = \sigma(\mathbf{x}_i * \mathbf{W}_j), \tag{2}$$

where σ is an activation function and $*$ denotes the two-dimensional convolution. The reconstruction of $\mathbf{x}_i$, i.e., $\mathbf{r}_i$, is obtained by:

$$\mathbf{r}_i = g_{\mathbf{W}'}(f_{\mathbf{W}}(\mathbf{x}_i)) = \sigma\left(\sum_{j \in \mathbf{H}} \mathbf{h}_j * \mathbf{W}'_j \right), \tag{3}$$

where $\mathbf{H}$ denotes the set of all the hidden feature maps and $\mathbf{W}'$ is usually forced to be the transpose of $\mathbf{W}$. A cost function is defined to minimize the reconstruction error over all the training data using mean squared error (MSE):

$$cost(\mathbf{W}) = \frac{1}{2N} \sum_{i=1}^{N} (\mathbf{x}_i - \mathbf{r}_i)^2. \tag{4}$$

The cost function can be solved using the back propagation algorithm [14] as standard networks.

To build deep networks, we cascade several convolutional autoencoders to form a stacked CAES. The input of each layer is the encoding of the layer below. The unsupervised training can be implemented in a greedy layer-wise fashion. Although unsupervised pre-training can guide the learning and support better generalizations from the training set, the discriminability can be further enhanced through a supervised fine-tuning. Therefore, based on the learned unsupervised architecture, a fine-tuning procedure is conducted using the supervision in the source domain. For the sake of simplicity, we denote $\mathbf{W}$ as the overall parameter of the five layers after fine-tuning. Note that the fine-tuned network not only retains the shared architecture across the two domains, but is more discriminative than the unsupervised network. The architecture of the five-layer stacked convolutional autoencoders (CAE1 to CAE5) is shown in Fig. 3(b).

Once $\mathbf{W}$ has been learned, we can obtain a transformed feature representation for $\mathbf{X}$, and thus Eqn. (1) can be further represented by:

$$\max_{\mathbf{B}} \{P(\mathbf{Y}|\mathbf{X})\} = \max_{\mathbf{B}} \left\{ \sum_{\mathbf{B}} P(\mathbf{Y}|f_{\mathbf{W}}(\mathbf{X}),\mathbf{B}) P(\mathbf{B}) \right\}. \tag{5}$$

4.2 Deep Learning with Top-Down Transfer

Following the CAES, an FCO layer with ontology priors is learned on the shared feature abstractions in the source domain and transferred to the target. An ontology is a high-level semantic structure reflecting whether concepts are close to each other or not. Such priors with respect to this relationship can be more discriminative to different concepts, especially those with great differences. For example, "beach" and "sky" belong to the same middle-level node "scene," while "dog" belongs to "animal." Therefore, the priors of "beach" and "sky" are similar, but very different from that of "dog." In this paper, we have defined an ontology curated for the domain of personal photos. For the concepts in this defined ontology, the relationship among different concepts is inherited across the two

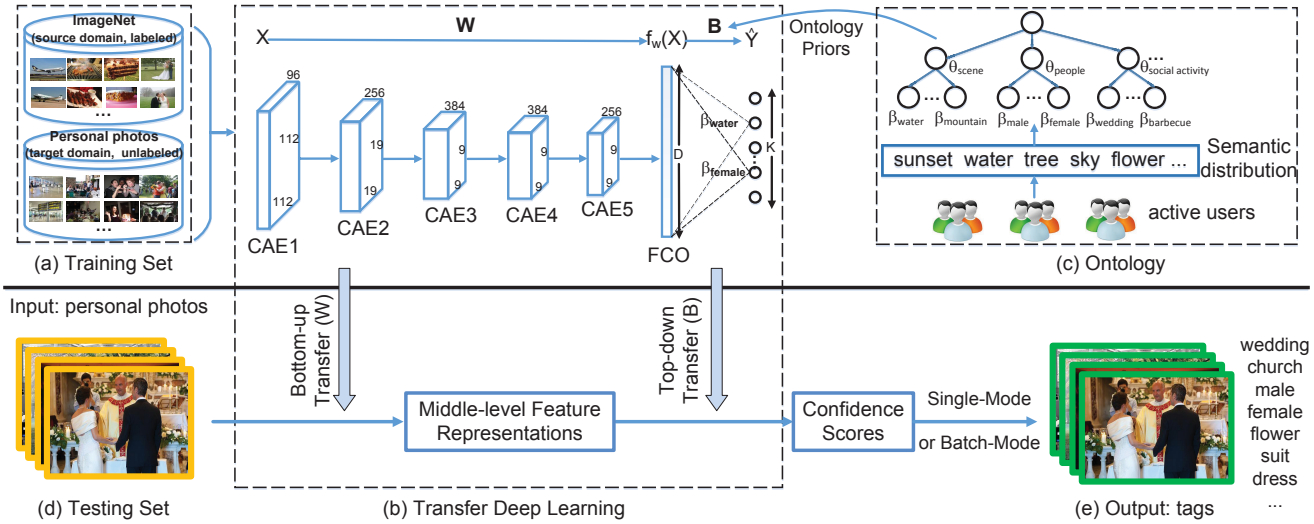

Figure 3: The network of the proposed transfer deep learning. (a) The training set contains the labelled source images and the unlabelled target images. (b) The network of the transfer deep learning with ontology priors. It is first trained on both ImageNet (the source domain) and personal photos (the target domain) by pre-training and fine-tuning for discovering shared middle-level feature abstractions across domains. Once the shared feature abstractions are learned, the top layer with ontology priors is further trained. In the testing stage, the resultant parameters $\mathbf{W}$ and $\mathbf{B}$ can be transferred to the target domain to obtain the middle-level feature representations (a bottom-up transfer) and high-level confidence scores (a top-down transfer). (c) An illustration of the ontology collecting scheme. (d) The input, in the testing stage, is highly flexible which can either be a single photo or a photo collection. (e) The tagging result.

domains. For example, a middle-level node "animal" is composed of the same leaf nodes (e.g., "cow," "bird," etc.) in both domains. Therefore, based on the shared feature abstractions and the inherited relationship, the parameters of the FCO layer can be learned from the source domain and transferred to the target without much of a gap. The ontology priors can enhance the correlations among close concepts and weaken those among dissimilar ones, and thus boost the prediction accuracy. Meanwhile an ontology built upon the shared feature abstractions across domains can help reduce the domain gap, which is the main difference between our approach and some single-domain learning methods, e.g., [25].

As $\mathbf{W}$ has been learned, we fix $\mathbf{W}$ and introduce the ontology priors into Eqn. (5) to maximize:

$$P(\mathbf{Y}|\mathbf{X}) = \sum_{\mathbf{B},\mathbf{\Theta}} P(\mathbf{Y}|f_{\mathbf{W}}(\mathbf{X}),\mathbf{B})P(\mathbf{B}|\mathbf{\Theta})P(\mathbf{\Theta}), \quad (6)$$

where $\mathbf{B} = [\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K]^T \in \mathbb{R}^{K\times D}$ and $\mathbf{\Theta} = [\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_M]^T \in \mathbb{R}^{M\times D}$ are the priors of the leaf nodes and the middle-level nodes in the defined ontology, respectively. The M and K are the number of middle-level nodes and leaf nodes, respectively. The prior over a leaf node is constrained by its immediate middle-level node (i.e., parent node) in the form of a conditional probability. We define a function $parent(\cdot)$ as a mapping from leaf nodes to their middle-level nodes, i.e., if k and m are indexes of a leaf node and a middle-level node separately, then $parent(k) = m$.

The typical choice for priors $\mathbf{B}$ and $\mathbf{\Theta}$ is Gaussian distribution. We thus define the following forms for $\mathbf{B}$ and $\mathbf{\Theta}$:

$$\boldsymbol{\beta}_k \sim \mathcal{N}(\boldsymbol{\theta}_{parent(k)}, \frac{1}{\lambda_1}\mathbf{I}_D), \boldsymbol{\theta}_{parent(k)} \sim \mathcal{N}(0, \frac{1}{\lambda_2}\mathbf{I}_D), \quad (7)$$

where $\boldsymbol{\beta}_k \in \mathbb{R}^D$ denotes the prior for the k^{th} leaf node, whose mean is determined by its parent $\boldsymbol{\theta}_{parent(k)}$ and $\mathbf{I}_D$ is a diagonal covariance. Let $\boldsymbol{\theta}_m$ be a prior of the m^{th} middle-level node in the ontology. $\boldsymbol{\theta}_m$ consists of a set of $\boldsymbol{\beta}_k$ where $parent(k) = m$. λ_1 and λ_2 are the scale factors of the two covariance matrix. We define $C_m = |\{k|parent(k) = m\}|$, where $|\cdot|$ denotes the cardinality of a set. As $\boldsymbol{\beta}_k$ and $\boldsymbol{\theta}_m$ are Gaussian distributions, given $\boldsymbol{\beta}_k$, $\boldsymbol{\theta}_m$ can be represented as in [25], which is:

$$\boldsymbol{\theta}_m = \frac{1}{C_m + \lambda_2/\lambda_1} \sum_{parent(k)=m} \boldsymbol{\beta}_k, \quad (8)$$

where $\boldsymbol{\theta}_m \in \mathbb{R}^D$. In general, we resort to MAP estimation to determine the value of the FCO layer's parameters $\mathbf{B}$ and $\mathbf{\Theta}$, which is to maximize:

$$\log P(\mathbf{Y}|f_{\mathbf{W}}(\mathbf{X}),\mathbf{B}) + \log P(\mathbf{B}|\mathbf{\Theta}) + \log P(\mathbf{\Theta}). \quad (9)$$

By selecting the mean squared error (MSE) as loss, the loss function can be expressed as:

$$\min_{\mathbf{B},\mathbf{\Theta}} \left\{ ||\mathbf{B}\ f_{\mathbf{W}}(\mathbf{X}) - \mathbf{Y}||^2 + \frac{\lambda_1}{2}\sum_{k=1}^{K} ||\boldsymbol{\beta}_k - \boldsymbol{\theta}_{parent(k)}||^2 + \frac{\lambda_2}{2}||\mathbf{\Theta}||^2 \right\}. \quad (10)$$

4.3 Optimization for Top-Down Transfer

To efficiently solve the above loss function, we propose to transform the $\mathbf{\Theta} \in \mathbb{R}^{M\times D}$ matrix into the same dimension as $\mathbf{B}$. Let $\mathbf{\Theta} = [\boldsymbol{\theta}_{parent(1)}, \boldsymbol{\theta}_{parent(2)}, ..., \boldsymbol{\theta}_{parent(K)}]^T \in \mathbb{R}^{K\times D}$, then Eqn. (10) can be simplified into the following form:

$$\min_{\mathbf{B},\mathbf{\Theta}} \left\{ ||\mathbf{B}\ f_{\mathbf{W}}(\mathbf{X}) - \mathbf{Y}||^2 + \frac{\lambda_1}{2}||\mathbf{B} - \mathbf{\Theta}||^2 + \frac{\lambda_2}{2}||\mathbf{\Theta}||^2 \right\}. \quad (11)$$

By fixing $\mathbf{\Theta}$, we set the derivative of $\mathbf{B}$ of the above loss function to zero, then $\mathbf{B}$ can be updated according to the following rules:

$$\mathbf{B} = \left(2\mathbf{Y}\ f_{\mathbf{W}}(\mathbf{X})^T + \lambda_1\mathbf{\Theta}\right)\left(2f_{\mathbf{W}}(\mathbf{X})f_{\mathbf{W}}(\mathbf{X})^T + \lambda_1\mathbf{I}\right)^{-1}, \quad (12)$$

where $\mathbf{I}$ is an identity matrix. Once we obtain an updated $\mathbf{B}$, we can recalculate $\mathbf{\Theta}$ using Eqn. (8) and transform it again. Therefore, Eqn. (11) can be optimized by iteratively updating $\mathbf{B}$ and $\mathbf{\Theta}$ until the difference between two successive iterations is below a threshold, e.g., 10^{-4}.

5. PERSONAL PHOTO TAGGING

Once the deep network is trained, we describe two tagging modes to highly efficiently tag personal photos, i.e., a single-mode for tagging a single photo and a batch-mode for tagging a photo collection. The single-mode only takes visual content into account, while the batch-mode further combines visual content with time constraints in a photo collection since the time information has been demonstrated as an essential constraint in the domain of personal photos [8].

5.1 Tagging with Single-Mode

Let $\mathbf{x} \in \mathbb{R}^d$ be the raw pixels of a single photo. We feed $\mathbf{x}$ into the learned stacked convolutional autoencoders and obtain a transformed feature representation $f_{\mathbf{W}}(\mathbf{x}) \in \mathbb{R}^D$. Then the tagging problem can be formulated as the following objective function:

$$\min_{\mathbf{y}} ||\mathbf{B}\ f_{\mathbf{W}}(\mathbf{x}) - \mathbf{y}||^2, \tag{13}$$

where $\mathbf{y} \in \mathbb{R}^K$ denotes the label vector indicating a confidence score for each concept. We can directly obtain a closed formed solution of $\mathbf{y}$, which is

$$\mathbf{y} = \mathbf{B}\ f_{\mathbf{W}}(\mathbf{x}). \tag{14}$$

Typically, we can utilize $\mathbf{y}$ in two ways. One way is to sort concepts according to their scores in $\mathbf{y}$ in decreasing order and select the top k concepts as the tagging results. An alternative way is to set a threshold and concepts whose scores are above the threshold can be selected as the tagging results.

5.2 Tagging with Batch-Mode

One of the most distinct characteristics of personal photos is the metadata stored in the digital photo files. This metadata includes the timestamp when the photo was taken and the geo-location of the photo. The metadata can be very useful in bridging the semantic gap in multimedia understanding. The single-mode only considers the visual content of a single photo, because an absolute timestamp is not very useful for understanding a photo via algorithms. However, a series of timestamps of a photo collection can be exploited for discovering the relationship among photos and boosting the tagging performance. For example, if the timestamps of two photos have a short interval between them, we can infer that the two photos were taken at the same event and the tagging results of the two photos should be highly correlated. Since geo-location information is not always available, we only leverage the time information in this paper.

Suppose there is a photo collection $\mathbf{X} \in \mathbb{R}^{d \times N}$ containing N photos and a label matrix $\mathbf{Y} \in \mathbb{R}^{K \times N}$. To reflect the time constraints, we construct an affinity matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$ by:

$$S_{i,j} = \begin{cases} \exp\{-\frac{||t_i - t_j||^2}{\gamma^2}\}, & |t_i - t_j| < T, \\ 0, & \text{otherwise}, \end{cases} \tag{15}$$

where t_i denotes the timestamp of photo i, γ is a free parameter to control the decay rate and T is a threshold. In a photo collection, if the difference of timestamps between two photos is smaller than T, the two photos are likely to share the tagging results. Considering the time constraints and visual clues simultaneously, the objective function of the batch-mode is formulated as:

$$\min_{\mathbf{Y}} \left\{ Tr[\mathbf{Y}^T \mathbf{L} \mathbf{Y}] + ||\mathbf{B}\ f_{\mathbf{W}}(\mathbf{X}) - \mathbf{Y}||^2 \right\}, \tag{16}$$

where $\mathbf{L} = \mathbf{A}^{-1/2}(\mathbf{A} - \mathbf{S})\mathbf{A}^{-1/2}$ and $\mathbf{A}$ is a degree matrix defined as the diagonal matrix with the degrees $a_1, ..., a_N$ on the diagonal and $a_i = \sum_{j=1}^{N} S_{i,j}$. By setting the derivative of $\mathbf{Y}$ to zero, the above optimization has a closed formed solution:

$$\mathbf{Y} = 2\mathbf{B}\ f_{\mathbf{W}}(\mathbf{X})(\mathbf{L} + \mathbf{L}^T + 2\mathbf{I})^{-1}, \tag{17}$$

where $\mathbf{I}$ is an identity matrix and the matrix $\mathbf{Y}$ indicates the tagging results of the whole collection, where each column is a set of confidence scores of a single photo.

6. EXPERIMENTS

In this section, we evaluate the proposed approach on two datasets. Personal photo tagging with single-mode and batch-mode is evaluated on a real personal photo dataset collected from 25 volunteers. Besides, an application of personal photo search is conducted on the public MIT-Adobe FiveK photo dataset [2], as it covers a broad range of topics in personal photos.

6.1 Dataset

Training: For each of the 272 concepts, we randomly selected about 650 images and obtained 180,000 images in total from ImageNet as the training data in the source domain. Although the user tags are subjective and noisy, to learn the intermediate feature representation across the two domains, we selected about $180,000$ photos tagged with the 272 concepts (650 photos for each) from the $10,000$ active users in Flickr as the training data in the target domain. The training set contained about 0.36 million images in total.

Testing: In photo tagging, the testing photos were collected from 25 volunteers. The 25 volunteers, including 17 males and 8 females, were from different educational backgrounds, including computer science, mathematics, physics, business, management science, art and design. All the volunteers were familiar with photography and liked taking photos. Among the volunteers, 19 of them were students ranging from 20 to 28 years old, while the rest were employees ranging from 30 to 45 years old. Each volunteer was asked to contribute at least 500 photos of his/her own and all volunteers contributed 35,217 testing photos in total.

As there is no well-labelled datasets in the domain of personal photos, to conduct the evaluation and comparison with other approaches, we organized a ground-truth dataset with manual labeling. Since the labeling procedure was very time-consuming, the 25 volunteers were asked to randomly annotate one fifth of their own personal photos. The 272 concepts were annotated for each photo on three levels: 2–Highly Relevant; 1–Relevant; 0–Non Relevant. Before labeling a photo, each volunteer was strictly requested to browse the 272 concepts, from middle-level nodes to leaf nodes in the ontology. Finally, we obtained 7,000 annotated personal photos in total, which were used in the following evaluations. The distribution of the 7,000 photos on the different topics (represented by middle-level nodes) is shown in Fig. 4. The top five topics of the personal photos were related to "scene," "public places," "plant," "people," and "home," which demonstrates the semantic bias in the domain of personal photos. Fig. 5 further shows the statistics of the number of concepts in photos. Each column expresses the number of photos for a given number of concepts. Among the 7,000 photos, there were 10 photos having 15 relevant or highly relevant

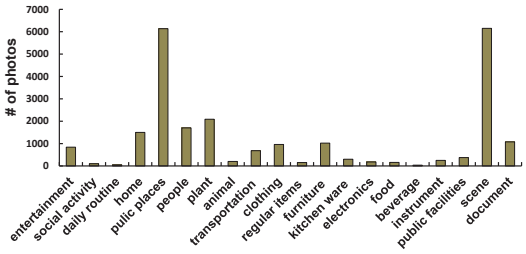

Figure 4: The photo distribution on middle-level nodes in our ontology. Each photo can contain multiple concepts.

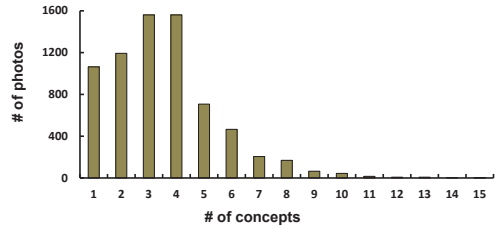

Figure 5: The concept distribution for personal photos.

concepts and about 70% of the photos present more than three relevant or highly relevant concepts which indicates the complexity in the visual appearances of personal photos.

6.2 Experiment Settings

Compared approaches: The following approaches were compared for the performance evaluation:

1. Tag ranking [19]: A typical approach uses hand-crafted features (225-d color moment and 128-d wavelet feature). It estimates initial relevance with concepts by a Gaussian kernel function and refines them by random walk. The kernel function is learned on ImageNet with 1,000 images for each concept, and the tag graph is built as in [19].
2. **D**yadic **T**ransfer **L**earning (DTL) [28]: A nonnegative matrix tri-factorization based transfer learning framework for image tagging, where the same hand-crafted features are extracted as in Tag ranking.
3. Transfer Learning with **G**eodesic **F**low **K**ernel (GFK) [10]: An unsupervised approach to learn domain-invariant features by leveraging the subspaces that lie on the geodesic flow. Hand-crafted features are extracted as in Tag ranking and the nearest neighbor classifier is adopted as in [10]. We selected GFK because of its best performance over other hand-crafted feature-based transfer learning methods.
4. **D**eep **l**earning with no transfer (DL) [14]: A deep learning approach with five convolutional layers and three fully connected layers. The networks are trained in the source domain (i.e., ImageNet), with about 650 images for each of the 272 concepts and about 180,000 training images in total.
5. **D**eep **l**earning with **Flickr** training data (DL(Flickr)): We trained the same network as DL except for using the $180,000$ Flickr training data in the target domain.
6. **D**eep **l**earning with **t**op-down **t**ransfer (DL+TT): The same architecture and training set as DL except for the ontology priors embedded in the top, fully connected layer.
7. **D**eep **l**earning with **b**ottom-up **t**ransfer (DL+BT): A deep learning approach with five-layer CAES and one fully connected layer. The networks are trained on both domains.
8. **D**eep **l**earning with **f**ull **t**ransfer (DL+FT) (i.e., bottom-up and top-down transfer): The same architecture and training set as DL+BT except for the ontology priors embedded in the top, fully connected layer.

Note that DL+TT, DL+BT, and DL+FT are proposed in this paper.

Network architecture: The architecture of our network is summarized in Fig. 3(b) and contains five convolutional layers and one fully connected layer with detailed specifications, CAE1 including 96 filters of size $7 \times 7 \times 3$ with a stride of 2 pixels, CAE2 including 256 filters of size $5 \times 5 \times 96$ with a stride of 2 pixels, CAE3 including 384 filters of size $3 \times 3 \times 256$ with a stride of 1 pixel, CAE4 including 384 filters of size $3 \times 3 \times 384$ with a stride of 1 pixel and CAE5 including 256 filters of size $3 \times 3 \times 384$ with a stride of 1 pixel. The input photos were color images of size $224 \times 224 \times 3$. To achieve higher computational efficiency and robustness, max-pooling layers were used following CAE1, CAE2 and CAE5 with the same window size of 3×3 and strides of 3, 2, 2, respectively. In the convolutional layers, rectifier linear unit $max(0, x)$ was adopted as the activation function, while in the top layer a linear activation function was used. The parameters λ_1 and λ_2 related to the prior distributions in the top layer were learned on a validation set in the source domain and set as $\lambda_1 = 30$ and $\lambda_2 = 10$.

Evaluation metrics: we adopted Normalized Discounted Cumulative Gain (NDCG) as metrics to evaluate photo tagging and Precision@K to evaluate photo search. The NDCG measures multi-level relevance and assumes the relevant tags are more useful when appearing higher in a ranked list. This metric at the position of p in the ranked list is defined by:

$$NDCG@p = Z_p \sum_{i=1}^{p} \frac{2^{r^i} - 1}{log(1+i)}, \tag{18}$$

where 2^{r^i} is the relevance level of the i^{th} tag and Z_p is a normalization constant such that $NDCG@p = 1$ for the perfect ranking. Additionally, the top-N error rates were adopted in tagging as they are more intuitional for common users. For example, the top-5 error rate is the ratio of testing photos whose top-5 tags are all irrelevant.

6.3 Evaluation of Tagging with Single-Mode

The problem of tagging with single-mode is to assign one or more relevant concepts to a given personal photo based on its visual content. As presented in Section 6.1, the 7,000 photos of real users with ground truth were evaluated. Fig. 6 shows the NDCG of different approaches for tagging personal photos. Obviously, we can see that DL(Flickr) is even far below than the method DL trained on ImageNet without transfer learning, which indicates the large percentage of noises in the user-provided tags in Flickr. Fig. 7 shows the error rates of different approaches over the 7,000 personal photos and an ideal performance of the DL approach (denoted as "DL+withinDomian") which is trained and tested on ImageNet. Overall, the tagging performances across domains were inferior to that within the same domain. The results verify our observation that there are significant domain gaps between Web images and personal photos.

It can also be observed from Fig. 6 and Fig. 7 that the existing tagging approach (Tag ranking) using hand-crafted features was the worst to bridge the domain gaps. It indicates that pre-defined feature types have difficulty discovering the shared feature representations across the two domains. By adopting cross-domain learning ideas, DTL [28] and GFK [10] were superior to the Tag ranking, but were inferior to the deep learning-based approach (DL). This shows stronger learning and generalization abilities of deep learning than the hand-crafted features. When further integrating transfer learning to deep learning, DL+TT, DL+BT and DL+FT achieve better performance than the DL approach. DL + FT achieved the best re-

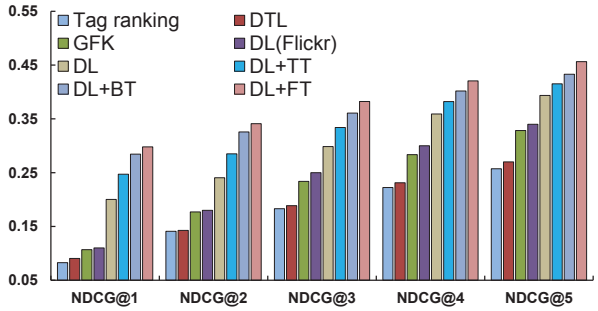

Figure 6: The NDCG of different approaches for tagging personal photos.

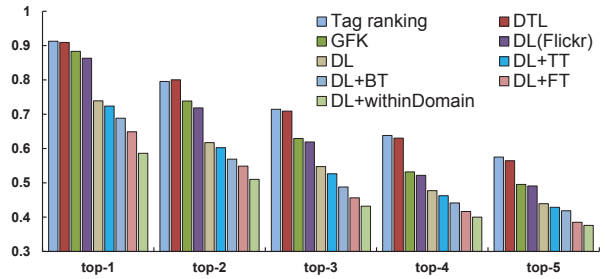

Figure 7: The top-N error rates of different approaches for tagging personal photos and an ideal performance obtained by training and testing on ImageNet (denoted as DL+withinDomain).

sult. The NDCG@5 of the full transfer network increased 12.8% and 4.5% against the GFK and the DL approach, respectively. The top-5 error rate of the full transfer network was 40.0%, which is very close to the top-5 error rate 37.6% within the domain. The superior performance derives from the fact that the full model can discover the shared feature abstractions from low-level raw pixels and the shared ontology from semantic structures across the two domains. In Fig. 7, we can observe that DL+FT (guided by the ontology) relatively decreased about 8.0% (in terms of top-5 "error" rate), compared to DL+BT. By extending the ontology beyond the three hierarchies with deeper networks, the top-down transfer can get further improvement. For the full model (DL+FT), we also varied the relative percentage of the labelled training set (source domain) as other typical transfer learning methods. The result is drawn in Fig. 13.

Fig. 10 shows the tagging results of different approaches ranked by confidence scores. We can observe that the typical approaches with hand-crafted features work well for "simple" photos with a clean background and few objects (e.g., the first photo in the left column), because of the limited descriptive power of these features. However, personal photos are always captured in the real world with uncontrolled environments, which are more likely to present complex visual patterns. This challenge can deteriorate the performance of the hand-crafted feature-based approaches. On the other hand, the deep learning-based approaches show stronger generalization abilities. However, it is still challenging for algorithms to predict the correct tags, since the photos are annotated by owners so that the ground truth is sometimes very subjective. For example, in Fig. 10, the content of the first photo in the right column is very similar to an "exhibition hall" or an "office," but the owner of the photo considers the two tags are wrong, because he took the photo in a restaurant. We will resort to the batch-mode to solve this problem.

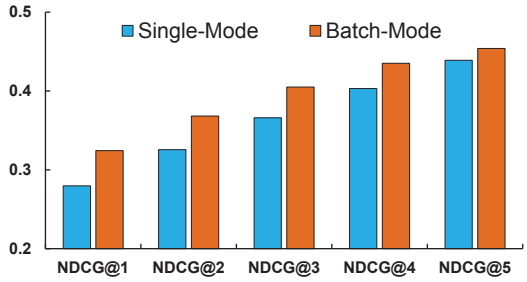

Figure 8: The NDCG of single-mode and batch-mode for tagging personal photos.

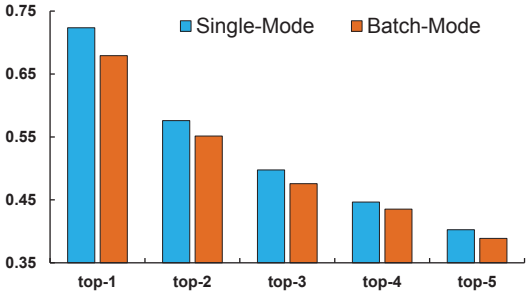

Figure 9: The top-N error rates of single-mode and batch-mode for tagging personal photos.

6.4 Evaluation of Tagging with Batch-Mode

The problem of tagging with batch-mode is how to assign concepts to a given photo collection, where the visual content and rich metadata embedded in personal photos is exploited simultaneously. As the geo-locations are not always available for photos, we only leverage the time information in this paper. Since the personal photos have been organized into several folders when submitted by the 25 volunteers, we considered the photos in each folder as a photo collection. We randomly selected three photo collections with various sizes (from 10 to 600 photos) from each user. In total, about 4,000 photos from 75 photo collections were evaluated.

Empirically, we set $\gamma = 2$ in Eqn. (15) to make $S_{i,j}$ in a proper scale, and we counted t_i by hours. Fig. 8 and Fig. 9 show the NDCG and error rates by using the DL+FT network with single-mode and batch-mode, respectively. In the single-mode, each photo in a photo collection was annotated individually by algorithms. While in the batch-mode, we ran the photo collection at one time. Compared to the single-mode, the batch-mode could consistently achieve a better performance, with gains of 4.4%, 4.3%, 3.9%, 3.2% and 1.5% from the NDCG@1 to NDCG@5. The improvement partly benefited from the idea to formulate the photo collection as a group, so that a photo can contribute its tags to others and vice versa depending on their time consistency. In our experiment, we set T at one hour. The shorter and longer time intervals led to inferior results over a validation set. Our analysis shows that, the shorter interval limited the tag propagation within a photo collection as the event in a photo collection usually lasts three to four hours. However, the longer interval inevitability led to noise.

Fig. 11 shows the tagging results for the same photo collections with single-mode and batch-mode, respectively. We found that tagging with batch-mode was more affective and could partially solve the problem of subjective annotation. For example, in the case of photo(c) of user #1, collection #3, the tagging algorithm with single-mode merely predicted "library," and "conference room" based on visual content. However, the batch-mode simul-

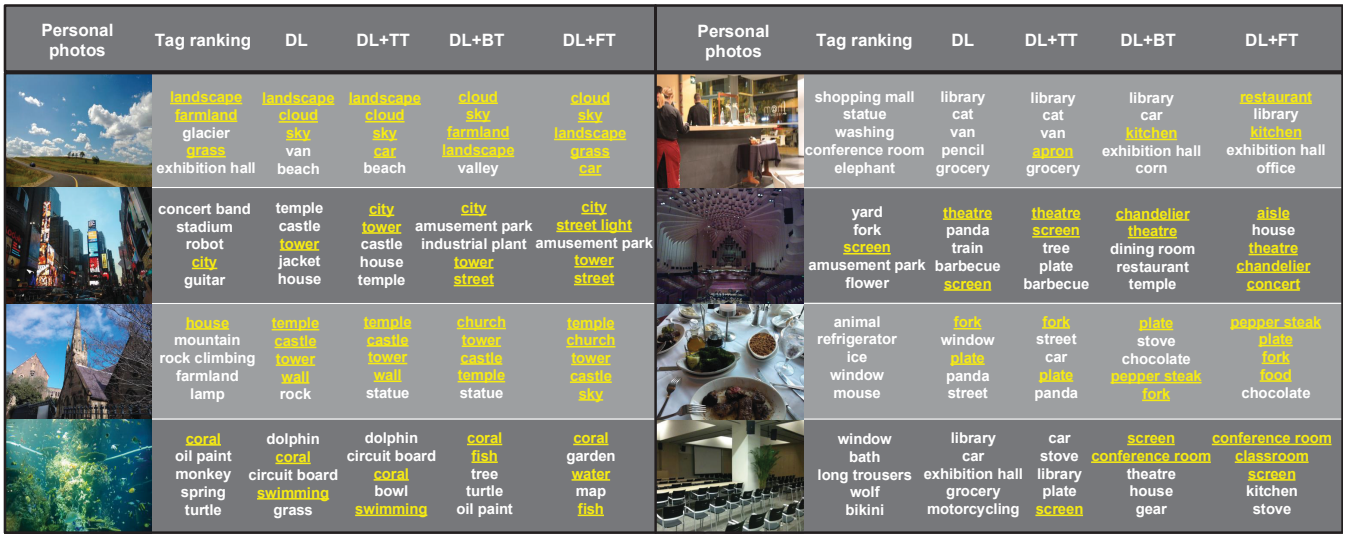

Figure 10: Example tagging results for eight personal photos by different approaches. Tag ranking and DL are produced by [19] and [14], respectively. The DL+TT, DL+BT and DL+FT are proposed in this paper. The top five tags are listed for each approach and the correct tags are highlighted in yellow and underlined.

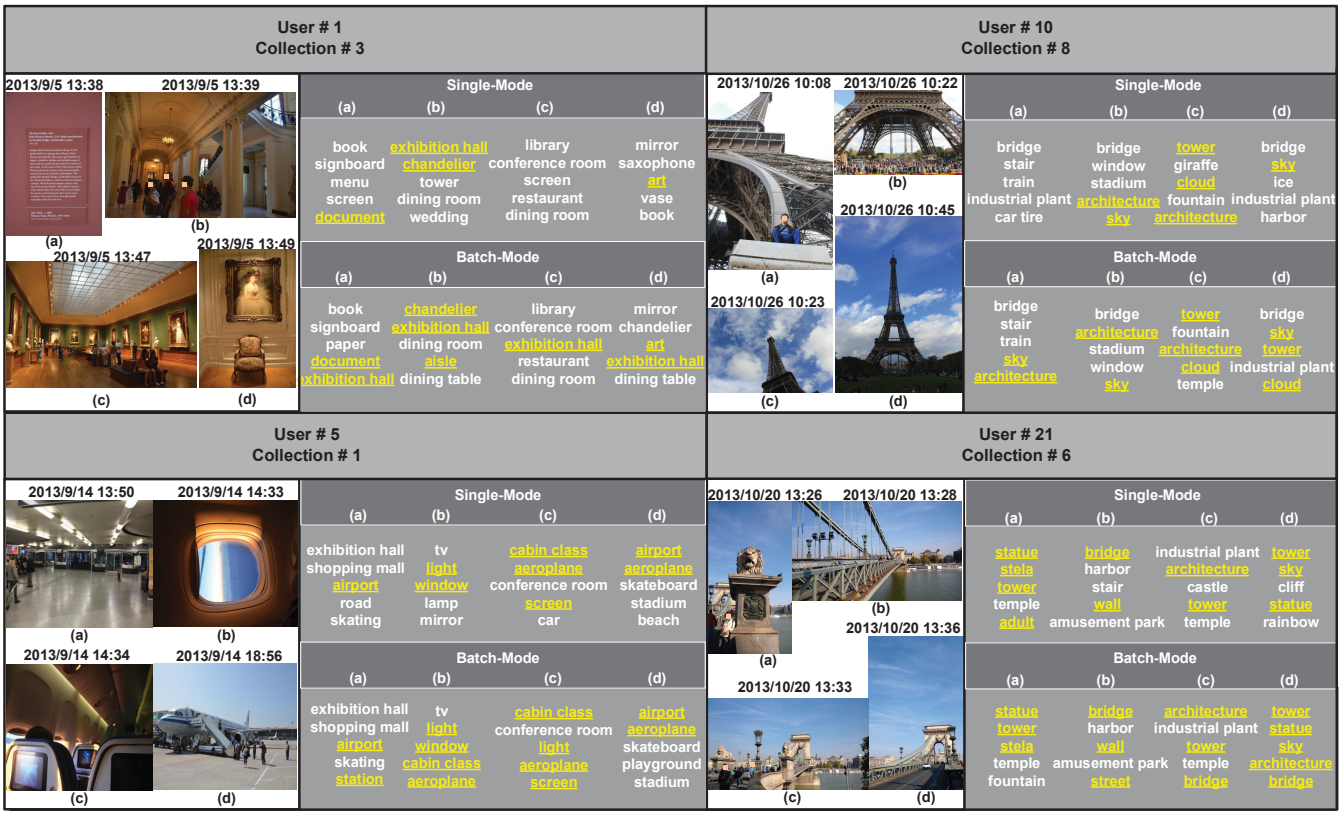

Figure 11: The tagging results for four example photos in each of four collections by single-mode and batch-mode, respectively. We adopted the network of DL+FT to produce the results for both modes. In single-mode, each photo was annotated individually. In batch-mode, we ran a photo collection at one time. The top five tags are listed for each photo and the correct tags are highlighted in yellow and underlined.

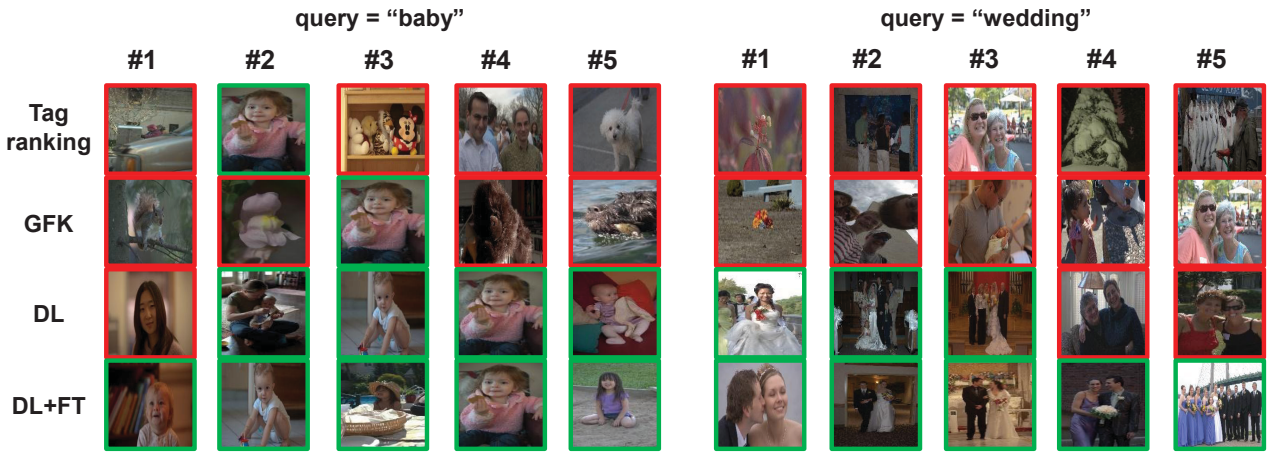

Figure 12: Photo search results on two exemplary queries for four typical compared approaches in the MIT-Adobe FiveK photo dataset. The green and red rectangles mark relevant and irrelevant photos, respectively. [best viewed in color]

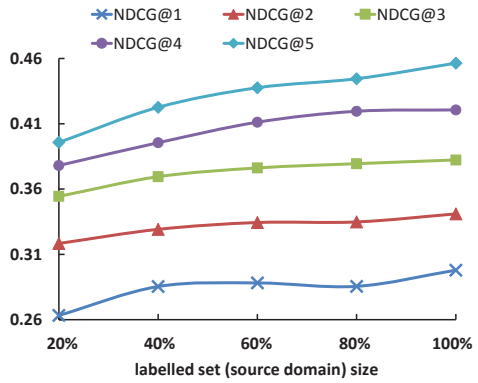

Figure 13: NDCG@1 to NDCG@5 with different percentages of the labelled training set.

Table 1: The average precision@K of photo search in the MIT-Adobe FiveK dataset.

	Tag ranking	GFK	DL(Flickr)	DL	DL+TT	DL+BT	DL + FT
P@1	0.34	0.41	0.42	0.55	0.61	0.71	0.77
P@5	0.28	0.32	0.31	0.46	0.52	0.62	0.74
P@10	0.22	0.26	0.25	0.40	0.45	0.56	0.72

taneously considered visual content and correlations from time-adjacent photos. Therefore, “exhibition hall” was propagated from time-adjacent neighbors. The batch-mode with time constraints complements the visual content analysis and enhances the robustness of our tagging algorithm.

6.5 Evaluation of Personal Photo Search

One of the most useful features of photo tagging is to help users recall and reconstruct the situation what he (or she) experienced. It can be conducted by photo search and ranking [16] with the user-typed queries. MIT-Adobe FiveK photo dataset was used in this evaluation. As there is no specific time information in this public dataset, we annotated photos by the proposed single-mode to generated the top-1 tag for each photo. Each tag was associated with a probability score produced by the deep learning networks. We used each of the 272 concepts in the personal photo ontology as queries. A photo is returned to users if its top-1 tag can be matched to the query. For each query, photos are ranked in the decreasing order according the probability scores. We retrieved ten photos for each query and manually judged the relevance of each photo. We calculated the average precision@K on all queries for different compared methods. The result is in Tab. 1. We can see that our proposed approach achieves much better performance compared to the other baselines. DL+FT achieves the best results, which demonstrates that DL+FT can build more accurate links between visual features and tags through the bottom-up and top-down transfer. Fig. 12 further illustrates some exemplary photo search results. For each query, the top-5 photos are returned. We can clearly see that the retrieved photos by the proposed approach can provide users with better results.

6.6 Complexity Analysis

We trained the six-layer network through the training set of 0.36 million images, which took three to four days on a NVIDIA GTX 580 GPU. For tagging with single-mode, the algorithm took less than 10 milliseconds on a PC with Intel Core Quad CPU with 2.83GHz and 4GB RAM. For tagging with batch-mode, it took three seconds for a photo collection of 200 photos (800*600 pixels). The high efficiency ensures an immediate response, and thus the transfer deep learning approach with two modes can be adopted as a prototype model for real-time mobile applications, such as photo tagging and event summarization on mobile devices.

7. CONCLUSIONS

In this paper, we have studied the problem of tagging personal photos. To effectively leverage supervised Web resource and reduce the domain gap between general Web images and personal photos, we have proposed a transfer deep learning approach to discover the shared representations across the two domains. Such representations can guide knowledge transfer from the source to the target domain. We conducted personal photo tagging on 7,000 real personal photos and personal photo search on the MIT-Adobe FiveK photo dataset. Experiments demonstrated the superiority of the transfer deep learning approach over the state-of-the-art hand-crafted feature-based methods and deep learning-based methods.

8. ACKNOWLEDGMENTS

This work was supported by 863 Program (2014AA015104), and National Natural Science Foundation of China (61273034, and 61332016).

9. REFERENCES

[1] Y. Bengio. Learning deep architectures for AI. *Found. Trends Mach. Learn.*, 2(1):1–127, Jan. 2009.

[2] V. Bychkovsky, S. Paris, E. Chan, and F. Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *CVPR*, pages 97–104, 2011.

[3] L. Cao, J. Luo, H. A. Kautz, and T. S. Huang. Annotating collections of photos using hierarchical event and scene models. In *CVPR*, pages 1–8, 2008.

[4] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval*, pages 1–9, 2009.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, pages 248–255, 2009.

[7] J. Fu, J. Wang, Y. Rui, X.-J. Wang, T. Mei, and H. Lu. Image tag refinment with view-dependent concept representations. In *IEEE Transactions on Circuits and Systems for Video Technology*. IEEE, 2014.

[8] A. C. Gallagher, C. Neustaedter, L. Cao, J. Luo, and T. Chen. Image annotation using personal calendars as context. In *ACM Multimedia*, pages 681–684, 2008.

[9] X. Glorot, A. Bordes, and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, pages 513–520, 2011.

[10] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, pages 2066–2073, 2012.

[11] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, pages 309–316, 2009.

[12] N. Imran, J. Liu, J. Luo, and M. Shah. Event recognition from photo collections via pagerank. In *ACM Multimedia*, pages 621–624, 2009.

[13] C. Ji, X. Zhou, L. Lin, and W. Yang. Labeling images by integrating sparse multiple distance learning and semantic context modeling. In *ECCV*, pages 688–701, 2012.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012.

[15] Q. V. Le, M. Ranzato, R. Monga, M. Devin, G. Corrado, K. Chen, J. Dean, and A. Y. Ng. Building high-level features using large scale unsupervised learning. In *ICML*, 2012.

[16] C. Li, Q. Liu, J. Liu, and H. Lu. Learning ordinal discriminative features for age estimation. In *CVPR*, pages 2570–2577, 2012.

[17] T. Li, T. Mei, I.-S. Kweon, and X.-S. Hua. Contextual bag-of-words for visual categorization. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(4):381–392, Apr. 2011.

[18] X. Li, C. G. Snoek, and M. Worring. Learning tag relevance by neighbor voting for social image retrieval. In *Proc. ACM International Conference on Multimedia Information Retrieval*, pages 180–187, 2008.

[19] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag ranking. In *WWW*, pages 351–360, 2009.

[20] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004.

[21] J. Masci, U. Meier, D. C. Ciresan, and J. Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *ICANN*, pages 52–59, 2011.

[22] O. Maxime, B. Leno, L. Ivan, and S. Josef. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, pages 1717–1724, 2014.

[23] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Visual Perception, Progress in Brain Research*, 155, 2006.

[24] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, 2003.

[25] N. Srivastava and R. Salakhutdinov. Discriminative transfer learning with tree-based priors. In *NIPS*, pages 2094–2102, 2013.

[26] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR*, pages 1521–1528, 2011.

[27] D. Tsai, Y. Jing, Y. Liu, H. A. Rowley, S. Ioffe, and J. M. Rehg. Large-scale image annotation using visual synset. In *ICCV*, pages 611–618, 2011.

[28] H. Wang, F. Nie, H. Huang, and C. Ding. Dyadic transfer learning for cross-domain image classification. In *ICCV*, pages 551–556, 2011.

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

[30] P. Wu, S. C.-H. Hoi, P. Zhao, and Y. He. Mining social images with distance metric learning for automated image tagging. In *WSDM*, pages 197–206, 2011.