Heritage Image Annotation via Collective Knowledge

Junjie Zhang[a,b], Qi Wu[c], Jian Zhang[a,*],Chunhua Shen[c],
Jianfeng Lu[b], Qiang Wu[a]

[a] *University of Technology Sydney, Australia*

[b] *Nanjing University of Science and Technology, China*

[c] *The University of Adelaide, Australia*

Abstract

The automatic image annotation can provide semantic illustrations to understand image contents, and builds a foundation to develop algorithms that can search images within a large database. However, most current methods focus on solving the annotation problem by modeling the image visual content and tag semantic information, which overlooks the additional information, such as scene descriptions and locations. Moreover, the majority of current annotation datasets are visually consistent and only annotated by common visual objects and attributes, which makes the classic methods vulnerable to handle the more diverse image annotation. To address above issues, we propose to annotate images via collective knowledge, that is, we uncover relationships between the image and its neighbors by measuring similarities among metadata and conduct the metric learning to obtain the representations of image contents, we also generate semantic representations for images given collective semantic information from their neighbors. Two representations from different paradigms are embedded together to train an annotation model. We ground our model on the heritage image collection we collected from the library online open data. Annotations on the heritage image collection are not limited to common visual objects, and are highly relevant to historical events, and the diversity of the heritage image content is much larger than the current datasets, which makes it

*Corresponding author: jian.zhang@uts.edu.au

more suitable for this task. Comprehensive experimental results on the benchmark dataset indicate that the proposed model achieves the best performance compared to baselines and state-of-the-art methods.

Keywords: Annotation Diversity, Image Annotation, Representation Learning, Collective Knowledge, Heritage Image Collection

1. Introduction

The automatic image annotation is a fundamental research problem in computer vision and pattern recognition. Multiple applications require the image annotation to understand, search and guide massive image visual information. It is indispensable to train an automatic annotation model, which is in favor of understanding image contents and browsing them within a large-scale database.

Methods for automatically annotating images have been extensively studied for decades, and most mainstream methods have been focused on annotating common visual objects and attributes [1, 2, 3, 4]. Given a set of training images with ground-truth, the direct way to uncover the relationship between the image visual content and tags is to train classifiers independently for each tag. Hand-crafted image features like SIFT [5], LBP [6] and GIST [7] combined with classifiers such as random forest [8], SVM [9] and voting [10] are widely adopted to fulfill the task. Most recently, deep convolutional neural network (CNN) has shown the advanced ability on various computer vision and pattern recognition tasks, including the image classification [11, 12] and retrieval [13, 14]. CNN can obtain high-quality image representations based on the purely supervised learning, and training CNN with a logistic loss function [15] has established a solid baseline for the annotation problem, while multiple works have been focused on modifying the loss function to better fit the problem, such as the pair-wise ranking loss [16]. However, by treating each tag independently, this line of works only model the relationship between the image visual content and each tag but overlook semantic information carried by tags themselves.

To bridge the semantic gap, a multi-modal representation is often carried

(a)
blue, cloud, landscape, sunset, sky

(b)
procession, crowd, city street, band, festival, anniversary

Figure 1: Examples of the training image from Flickr dataset and the heritage image collection. As we can see, the tags of the heritage image collection are more diverse and semantical.

out to learn the image representation as well as the tag representation. Canonical correlation analysis (CCA) [17] and kernel canonical correlation analysis (KCCA) [18] based methods project both visual and semantic representations into a latent space to tackle the annotation and retrieval problem. There are also related works leverage the semantic information by capturing the dependencies between tag pairs. In general, images with multiple tags have strong correlations among attached tags. For example, the tag 'ocean' and 'boat' have the potential to appear in the same image, while 'ocean' and 'tiger' do not. Probabilistic graphical models (PGM) such as Markov random field (MRF) [19] and conditional random field (CRF) [20], as well as the widely used recurrent neural network (RNN) [21], have been proved their efficiencies on capturing high-order tag dependencies.

Aforementioned methods are mainly proposed for the image annotation of common objects and their attributes and have achieved satisfactory results. However, there are two major drawbacks when they handle the more diverse image annotation. First of all, state-of-the-art methods on the image annotation problem use CNN as the backbone model, which heavily relies on the fine-tuning the pre-trained image representation on the large-scale image dataset ImageNet

(a) (b) (c) (d)

Figure 2: The visual ambiguous of the heritage image collection. All images are annotated with the tag 'theatre.' However, the visual appearance of them is quite different. Fig. (a) is the exterior of a theatre, fig. (b) is the interior of a theatre, fig. (c) is the hall of a theatre, while fig. (d) is a group people taking a photo outside a theatre.

[22]. However, the tags of the real-world image can be more obscure and diverse. Take Fig. 1 as an example. The left image is sampled from the Flickr dataset [4]. These tags indicate the objects and their attributes, which are tightly related to the image visual content, while the right image is selected from the heritage image collection we collected online, it contains not only the easily recognized tags 'crowd' and 'city street', but also tags that can be inferred from context or related knowledge, such as 'anniversary' and 'festival'. Even for the same tag, the image visual appearance can be very different. For example in Fig. 2, all images are annotated with the tag 'theatre.' But the visual appearance of these images is quite diverse. Moreover, the domain difference between the training image set and ImageNet can be much larger than current datasets. Directly learning the image visual representation from the training set could degrade the performance. Secondly, since the tags of the real-world images can be very diverse, and classic annotation methods tend to treat tags as equally related to the image visual content or rule out ones that are less relevant, which makes their performances unsatisfactory.

To address above issues, we propose to uncover relationships between the image and its neighbors and utilize them to conduct the image representation learning and train an annotation model, namely annotation via collective knowledge. We ground our proposed model on the heritage image collection, given its visual and tag diversity. The whole framework is shown in Fig. 3. Specifically,

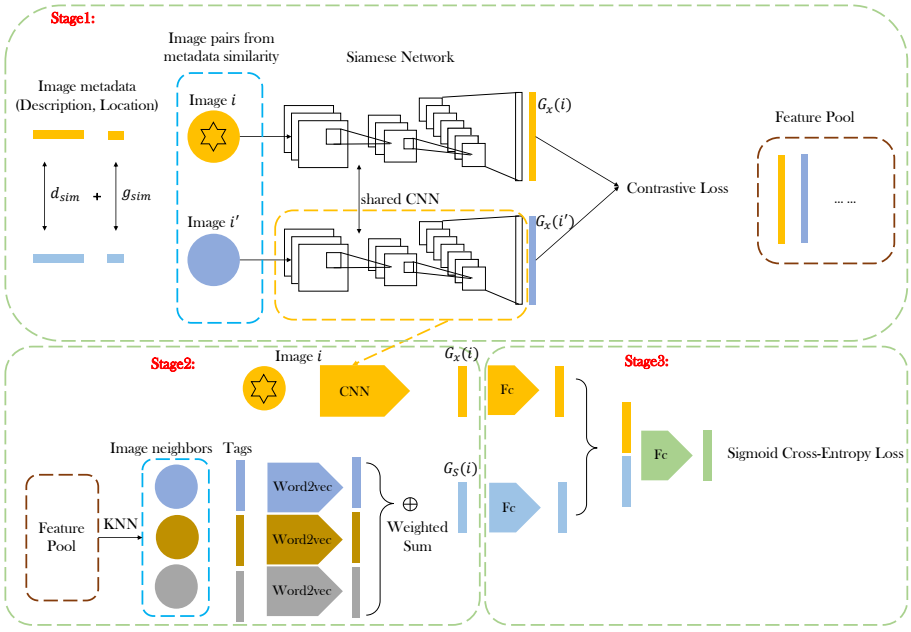


Figure 3: The framework of the proposed model. At the training stage, we first generate image pairs (i, i') for the visual representation learning based on the metadata similarity, in our case, the combination of the description similarity d_{sim} and location similarity g_{sim}. Image pairs are passed through the siamese network and trained with the contrastive loss. Then we retrieve image neighbors and summarise the semantic representation by adopting the weighted sum based on the image pair similarity and tag relevance. Both the visual representation $G_X(i)$ and the semantic representation $G_S(i)$ are fed into the fully-connected network to compute hidden states and further trained with the sigmoid cross-entropy loss.

we define neighborhood relationships between images based on their metadata, which are the descriptions of the background of images and locations. Images with similar contexts have the similar visual appearance, which can help us to eliminate the visual ambiguous. Therefore, we conduct the metric learning among image neighbors to learn effective visual representations. To handle the diversity of the tag set, we collect tag candidates from image neighbors and embed their semantic information with the visual representation to help infer all possible tags. The deep neural network is adopted in the representation learning and the annotation model. We collect a heritage image collection from the

library online open data as the benchmark dataset to verify the effectiveness of the proposed model. In summary, the main contributions of our model are as follows:

1) We propose to learn the image visual representation based on the metric learning. Different from previous works, we use metadata to measure similarities among image neighbors. The superior experimental results against hand-crafted features and CNN finetuned features indicate the significance of our proposed representation learning methods.

2) Different from previous works, our model annotates all tags via collective knowledge from image neighbors, including the visual representation learning and the semantic embedding, which is crucial for the real-world image annotation, since it focuses on not only the visually related tags but also the semantical tags which are related to the events.

3) We collect a heritage image collection as the benchmark dataset and ground our proposed model on it. Experimental results show that our model achieves the best performance against the state-of-the-art annotation methods.

The preliminary version of this work was published at ACPR 2017 [23]. The new material in this paper comprises a new perspective on the diverse image annotation, and we ground it on the heritage image collection, a new annotation model is proposed to tackle the problem. We collect a new benchmark dataset with more compared baselines, state-of-the-art methods, and ablation models are implemented and studied. More importantly, the proposed model achieves the best results against all compared methods.

2. Related Works

2.1. Mainstream Image Annotation Methods

As addressed before, the automatic image annotation, as a fundamental computer vision and pattern recognition problem, has been studied for decades. Multiple image datasets have been proposed for the research purpose, such as the PASCAL VOC [2], MS COCO [1], Flickr [4] and NUS-WIDE [3] etc. One

way to tackle the problem is to directly model the relationship between the image visual content and associated tags. Image visual representations can be chose from traditional hand-crafted features [5, 6, 7, 24] to the advanced CNN feature [11, 12], while the classifier can be a wide range of options including the random forest [8], SVM [9] and neural networks with suitable loss functions [15, 16]. These methods establish the baselines of the image annotation with satisfactory results, while some works focus on improving the baseline performances by leveraging image regions. In [25], Xue et al. use the multi-label multi-instance method to explore the image features both at the image level and the regional level. In [26], Wei et al. use BING [27] to generate object proposals and annotate tags independently inside each region, all regional annotations are fused together as the final annotation. Despite the fact that image visual representations are enhanced by introducing the regional information, most of these methods are targeting object-centered images.

To efficiently exploit the abundant semantic information carried by tags, several approaches [28, 29, 30, 31, 32] have been proposed to design a multi-modal representation of the image and its tags. The CCA [17] and KCCA [18] based methods and their variations are widely used in the image annotation and retrieval. In [33], Hwang et al. use the KCCA to leverage the importance of textual objects for the image annotation and retrieval. They reveal implied cues about object importance based on how people naturally annotate images with the text and then translate those cues into a dual-view semantic representation. In [30], a third view of the category or concept is added to the CCA to capture the high-level image semantics, which improves the retrieval performance. In [31], Murthy et al. propose to combine the CNN visual representation with the word embedding by using the CCA, while in [28], authors propose a label propagation framework based on the KCCA to tackle the annotation problem regardless whether the training set is annotated by experts. In [32], authors design a multi-modal curriculum learning (MMCL) strategy to tackle the semi-supervised image annotation problem. Different from previous works, our multi-modal representation is designed based on the visual

representation which is learned from the metric learning, and the semantic representation which is collected from image neighbors and the external knowledge base, in this way, we can infer all tags both visual and event-related.

There are also series of works focus on uncovering semantic information by modeling the tag pair correlation. Probabilistic graphical models are usually employed [34, 35, 36, 37, 38]. Different graph structures can model the visual representation-tag joint distribution from different perspectives, such as the Chow-Liu tree [34], directed acyclic graph [35], group sparsity [36], CRF [37] and ML-TLLT [38] etc. Most recently, RNN has been applied to capture the sequential dependencies, which is suitable for the image to language problem including the annotation and captioning, etc. In [21], Wang et al. have shown that RNN can efficiently capture high order label dependencies. They define each image ground-truth as an ordered sequence of tags and use CNN-RNN as one unified framework to annotate images in an end-to-end fashion. However, when it comes to the more complicated annotation such as the heritage image collection, the tag pair correlation can be too diverse to be captured as an ordered sequence.

There are also related works that utilize metadata to assist the annotation process. In [39], Johnson et al. propose to generate image neighbors by exploiting image metadata, and build a framework to merge the visual information between image and its neighbors. In [40], Jin et al. use WordNet as the knowledge base to analyze the hierarchical relationship in the tag set. In [41], personal annotation preference is considered in the form of tag statistics computed from images a user has uploaded in the past. These past images are used in [42] to learn a user-specific embedding space. Photo timestamps are exploited for time-sensitive image retrieval [43], where the connection between image occurrence and various temporal factors is modeled. In [44], time-constrained tag co-occurrence statistics are considered to refine the output of visual classifiers for the tag assignment. Different from these prior works, the metadata in our model is used to identify image neighbors and guide the image visual representation learning process by conducting the metric learning at the training

stage.

2.2. Tag Relevance Analysis

Given the diversity of the tag set in the image training set, tag relevance can guide the annotation process in an effective way. Therefore, we review related works on the tag relevance analysis. In early works [45, 46, 47], the tag relevance is estimated based on the semantic similarities between tag pairs. In [48], Sun et al. propose two distance metrics to quantify the tag relevance of the image visual content, which is measuring the tag visual relevance at a global level. In [49], Li et al. use the low-level visual feature similarity to find each image's neighbors and employ the KNN method to vote the tag relevance to the image content, which is measuring the image-specific tag relevance. In [50], the tag relevance is initially leveraged by the kernel density estimation; then the random walk is employed to refine the relevance based on the visual and semantic information. In [15, 51], nearest neighbor voting is used to estimate the tag visual relevance and annotate images. Inspired by these works, we also measure the tag relevance to guide the combination of neighborhood contributions for collective knowledge.

2.3. Heritage Image Research

There are some researches focus on the heritage image annotation problem. In our previous work [52], we establish a baseline by employing the basic CNN model for the heritage image collection annotation. In [53], Zhao et al. propose a CNN based framework (Sherlocknet) to tag the British Library one million images dataset into twelve categories and discover trends in art styles over historical time.

3. Image Annotation via Collective Knowledge

3.1. Model Overview

The key characteristic of our model is that we use collective knowledge to solve the image annotation problem, which is grounded on the heritage image

collection. The collective knowledge is reflected from two perspectives. One is the image visual representation. We uncover relationships between the image and its neighbors by measuring similarities among their metadata and conduct the metric learning to obtain the effective visual representation. The other is that we generate the semantic representation for the image by summarising contributions from its neighbors' tags. The visual and semantic representations are embedded together and passed through a neural network to finalize the annotation model. The entire model is shown in Fig. 3.

In the following part, the metadata similarity measurement and the metric learning for the image visual representation are first introduced in Sec. 3.2, and the tag relevance analysis and the semantic representation are described in Sec. 3.3. The final annotation model and training details are summarised in Sec. 3.4 and Sec. 3.5 respectively.

3.2. Visual Representation Learning

Considering the visual diversity of the training set, and tags are related to the image visual content in varying degrees, directly finetuning the visual representation of the deep neural network towards the training ground-truth can compromise the representational capacity. However, metadata as the reflection of the image content can be used to eliminate the image visual ambiguous and locate image neighbors. In our case, the heritage image collection is attached with the abundant information noted by photographers and librarians when cataloging them. These metadata are presented as the descriptions of historical events when the images were taken, and the locations of where these events happened, which reflect the visual content in a semantical way. See Fig. 4 for an example. Inspired by [39], we define image neighbors for the heritage collection based on metadata and conduct the metric learning to obtain the effective visual representation.

An ideal annotation model should be flexible to handle the different forms of metadata, to this end, we measure the similarity between metadata nonparametrically to define image neighbors. Let I be the heritage image collection,

(a)
Description: St Patrick's Day procession
Location: Hyde Park, College Street, Sydney

(b)
Description: Royal Easter Show
Location: Sydney Showground

(c)
Description: Women's winter fashion
Location: Snow's Department Store, Pitt Street, Sydney

Figure 4: Examples of the description and the location of the heritage image collection. As we can see, these metadata can help understand the image at a semantical level.

T be the set of tag candidates, and $D = \{(i,t)|i \in I, t \subseteq T\}$ indicates images associated with a set of tags. Metadata of the heritage image collection can be various, and based on the benchmark dataset we use, we consider the descriptions of historical events and the locations.

The description of a historical event is presented as a free-form phrase. Given the image i, we tokenize the description d_i into words, which results in a vocabulary V of word candidates. Therefore, for each image $i \in I$, a subset $v_i \subseteq V$ represents the description. We use the Jaccard similarity to compare two descriptions, that is, given two images i and i':

$$d_{sim} = |v_i \cap v_{i'}|/|v_i \cup v_{i'}| \tag{1}$$

Since each image i only possesses one geographical location g_i, we simply use the indicator function to represent the location similarity:

$$g_{sim} = \mathbf{1}(g_i, g_{i'}) : \begin{cases} 1 & \text{if} \quad g_i = g_{i'} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

After we have the similarities of the different type of metadata, we now give the final form of the metadata similarity, where λ is used to balance the two metrics:

$$meta_{sim} = d_{sim} + \lambda g_{sim} \tag{3}$$

Based on the metadata similarity, we apply the nearest neighbor approach to generate image pairs for the metric learning. For each image i, we select top

n_p similar images to form positive pairs, while negative pairs are composed of top n_n dissimilar images. We pass image pairs through the siamese network to conduct the metric learning for the visual representation. The siamese network consists of two identical neural networks; each receives one of input image pairs, the last layers of networks are fed to a contrastive loss function [54], which captures the similarity between two images. That is, given the image pair i and i', with the label y indicates the similarity, 1 for the positive pair, while 0 for the negative one. The networks are trained by minimizing the loss L:

$$D_W = \sqrt{\{G_X(i) - G_X(i')\}^2} \tag{4}$$

$$L = y\frac{1}{2}D_W^2 + (1-y)\frac{1}{2}\{max(0, m - D_W)\}^2 \tag{5}$$

where G_X is the output of the neural network that stands for the visual representation of the input image, m is the margin. D_W measures the Euclidean distance between two image representations $G_X(i)$ and $G_X(i')$. The contrastive loss L constrains the image pairs with the similar metadata to have the closer visual representations, while pushing the dissimilar pairs beyond the margin m. Since the metadata is the reflection of the image content, by training based on the metadata similarity, we can eliminate the image visual ambiguous and locate image neighbors, and by training with image pairs, we avoid finetuning the visual representation directly based on the diverse ground-truth. The experimental results, which are presented in Sec. 4, prove the effectiveness of the metric learning for the visual representation.

3.3. Tag Relevance Analysis & Semantic Summarisation

Since we consider the annotation for the individual image via collective knowledge from its neighbors, it is important to evaluate each tag's relevance to the image content. We choose evaluation metrics from two aspects: the semantic and visual information, namely the semantic field [45] and the neighbor voting, which is orthogonal to each other.

Given the image i, semantic field evaluates the tag $\tau \in t$ based on average semantic similarities between τ and other tags that are associated with the

image i. We consider the semantic similarity from two aspects: the context of the current image collection and the general knowledge base. Given two tags τ_j and τ_k, the context similarity refers to the semantic similarity that is obtained based on the context information of the current image collection. We adopt the normalized Google distance [55], where context statistics, including the tag frequence and tag pair co-occurrence, are sampled from the image collection instead of the webpage. For the general knowledge base, we use the Euclidean distance between word2vec vectors that are pre-trained from the Google News corpus [56]. We combine two metrics to balance the general and domain-specific knowledge for the given collection, that is:

$$s_{ctx}(\tau_j, \tau_k) = \frac{max(logf(\tau_j), logf(\tau_k)) - logf(\tau_j, \tau_k)}{logN - min(logf(\tau_j), logf(\tau_k))} \tag{6}$$

$$s(\tau_j, \tau_k) = s_{ctx}(\tau_j, \tau_k) + \alpha s_{w2v}(\tau_j, \tau_k) \tag{7}$$

$$\xi_{\tau_{j,k}} = exp(-s(\tau_j, \tau_k)^2/\sigma) \tag{8}$$

where $f(\tau)$ indicates the frequence of tag τ in the image collection, $f(\tau_j,\tau_k)$ is the co-occurrence of tag τ_j and τ_k, α is the weight, σ is the medium value of ξ, and N is the total number of images in the collection. Then the semantic filed similarity score of tag τ_j for image i can be computed by:

$$r_{ss}(\tau_j|i) = \frac{1}{|t|-1} \sum_{k=1,k\neq j}^{|t|-1} \xi_{\tau_{j,k}} \tag{9}$$

$$r(\tau_j|i) = \text{CombSUM}(r_{ss}(\tau_j|i), r(\tau_j|i)) \tag{10}$$

The neighbor voting measures the tag relevance of τ_j with respect to the image i by examining the visually similar images. Based on the visual representation we obtained from the last step, the metric retrieves multiple nearest neighbors, and the number of images annotated with tag τ_j is used as the similarity score $r_{nn}(\tau_j|i)$. Finally, two relevance scores are normalized and merged by the CombSUM as $r(\tau_j|i)$. By performing the tag relevance analysis, we convert the ground-truth of the training set from the hard assignment 0-1, where 0-1 denotes the absence-present of a tag, to a confidence vector, where each dimension indicates the tag relevance (0 occupies the absent tag). Considering

the diversity of the tag set, this step can benefit the summarization of collective knowledge by leveraging the tag relevance.

Since the tags are relevant to the image visual content in various degree, some of them are easier predicted based on the combination of semantical information than the individual image. We generate a semantic representation of the image to help infer all possible tags by summarising tag information from its neighbors. This is intuitive that when librarians manually annotate the heritage image collection, they not only observe the individual image but also look into the archived collection to find connections between images, and consider whether to transfer the tags.

Given image i, we retrieve m nearest neighbors based on the visual representation to form a candidate set Z_i, each image $i_j \in Z_i$ is associated with a relevance vector r_{i_j}, where $|r_{i_j}| = |T|$, we reserve the KD-tree structure for training set after we perform the nearest neighbor approach, which will be used during the tag prediction for new images. The summarised tag information of i is indicated as:

$$\varphi_{i,i_j} = exp(-\|G_X(i) - G_X(i_j)\|^2/\sigma) \tag{11}$$

$$G_S(i) = \sum_{j=1}^{m} \varphi_{i,i_j}(r_{i_j} \cdot H) \tag{12}$$

where σ is the medium value of φ. φ_{i,i_j} is the similarity of the image pair i and i_j, H is the matrix of the pre-trained word2vec, and each row of H stands for the word2vec of a tag. $G_S(i)$ gathers all the tag information from image neighbors based on the image similarity and the tag relevance of each neighbor. The reason that we adopt the weighted sum is to better leverage the importance of the annotation carried by the image neighbor with respect to the query image. Therefore, we can summarise the semantic information in a more accurate way.

3.4. Annotation via Collective Knowledge

For each image $i \in I$, we now have a visual representation $G_X(i)$ by the metric learning and a semantic representation $G_S(i)$ by summarising from its neighbors. We compute h_1 and h_2 dimensional hidden states for each image's visual

and semantic representations respectively, by a fully connected layer followed by a ReLU nonlinearity transform ψ, which are parameterized by (w_X, b_X), and (w_S, b_S). At this point, we concatenate hidden states of visual and semantic representation, and feed into a third fully connected layer parametered by (w_P, b_P) to obtain the probabilities of tags, that is:

$$\vartheta_X = \psi(w_X \cdot G_X(i) + b_X) \tag{13}$$

$$\vartheta_S = \psi(w_S \cdot G_S(i) + b_S) \tag{14}$$

$$P(t|i) = w_P \cdot [\vartheta_X; \vartheta_S] + b_P \tag{15}$$

where ϑ_X and ϑ_X indicate the hidden states of visual and semantic representations respectively. The neural network can be trained by minimizing the sigmoid cross-entropy loss towards the image ground-truth.

3.5. Training and Prediction

We use the VGGNet-16 as our backbone network for the metric learning, the output of the fc7 layer with ReLU transform is used as the visual representation. The training process is three-stage: first, we generate positive and negative image pairs for the visual representation learning based on similarities between image metadata. Then, we measure the tag relevance of the image content by exploring the semantic and visual information from image neighbors, and summarise the semantic representation for each image based on collective knowledge. And finally we fuse the visual and semantic representations together and feed it into fully connected layers. The whole training process is shown in Alg. 1.

As for the tag prediction during the test, when a new image k arrives, we first extract its visual representation by $G_X(k)$, since we build the KD-tree during training, we can easily query its neighbours from the training set, and summarise the semantic representation $G_S(k)$. Then we feed two representations into the annotation model to compute hidden states and predict tags.

Algorithm 1 Training stages of the annotation model.

Input: D, $\forall i \in I$, $t \subseteq T$, $v_i \subseteq V$

1: Measure the metadata similarity $meta_{sim}$ of image pairs based on eq. 3;
2: Generate positive and negative training samples for the metric learning by minimizing eq. 5, and obtain the visual representation $G_X(i)$;
3: Evaluate each tag's relevance of image i based on eq. 10;
4: Summarise the semantic representation $G_S(i)$ based on eq. 12;
5: Compute the hidden states of representations and concatenate them to pass through fully connected layers for the tag generation based on eq. 15;
6: Training the network with the sigmoid cross-entropy loss.

4. Experiments

In this section, we present the experimental details and results. Our model is evaluated on the heritage image collection we collected online. By comparing with baseline models and state-of-the-art methods, we show that our model achieves the best performance, and comprehensive ablation studies indicate the significance of each component in our model.

4.1. Data Preprocessing & Evaluation Metrics

We collect the raw heritage image collection from the library online open data. The dataset contains 37,931 valid images with textual information, including tag, description, and location, etc., which are annotated by librarians. To conduct the effective annotation, we lemmatize all the textual words to their dictionary forms, then we exclude tags that blew the occurrence threshold (0.2% of $|I|$) to avoid the insufficient sampling. Finally, images without tags are removed. The preprocessing results in 31,815 images with a size of the tag set $|T| = 257$; each image is attached with metadata: a description and a geographic location. We use 21,210 images for the training and 10,605 images for the test, the metadata is only used with training images. [1]

[1] The raw dataset is available at https://bitbucket.org/Junjie_Avalon/heritage_image_dataset

For overall and per-tag evaluation metrics, we use the average precision (AP). As an effective annotation model, the relevant tag should be ranked higher than irrelevant ones with respect to the image, and the same applies to the image with respect to a tag query. Therefore, we compute both the mean image average precision (imAP), which is averaging APs over all images and the mean average precision (mAP), which is averaging APs over all tags. Moreover, to conduct the quantitative evaluation, we predict up to three highest ranked tags above the threshold for each image to compare against the ground-truth. Overall and per-tag precision/recall/F1 score noted as $(O_P, O_R, O_{F1}, C_P, C_R, C_{F1})$ are reported.

4.2. Implementation Details

As mentioned in Sec. 3.5, the output of last two fully connected layer in the VGGNet-16 is used for the visual representation. For the neighbor voting in the tag relevance analysis and the semantic summarization, we retrieve $m = 50$ neighbors. The pre-trained word2vec for each tag is queried from the Google News corpus, which is a 300-d vector. And we set the hidden state dimension $h_1 = 512$ and $h_2 = 256$ for the visual and semantic representation respectively. As for the hyperparameters including λ and α, by referring to the previous works [39, 57], we adopt the grid-search to tune them. Our experimental result $\lambda = 0.85$ indicates that, in the given heritage image collection, we rely on the description similarity d_{sim} more, which is reasonable considering multiple historical events can happen in the same location. Similarly, α is used to balance the domain-specific and general knowledge for the given collection. The experimental result $\alpha = 0.76$ illustrates that we value domian-specific knowledge s_{ctx} more when measuring the semantic similarity, and the general knowledge base s_{w2v} is jointly considered to ensure the similarity metric's generalization. And we set $n_p = 1$ and $n_n = 2$ for the metric learning[2]. We train all models in-

[2]We varied the values of n_p and n_n during the experiments and no significant differences were observed.

cluding the siamese network and fully-connected layers for 30 epochs, with the SGD optimization and the learning rate decreases from 0.001 to 1/10 every ten epochs. During the test, a new image is extracted with the visual and semantic representation and passed through the annotation model to get the prediction.

4.3. Baselines and Compared Methods

To evaluate the effectiveness of our proposed model, we implement four baselines for comparisons, here we give the descriptions of these baselines and compared state-of-the-art annotation methods:

RandomGuess. All the annotation methods should achieve better results against RandomGuess [57], which randomly selects a subset of tags. The random prediction is run 50 times, and average evaluation scores are reported.

Multi-CNN. This is a standard CNN model without involving any additional information [52]. The VGGNet-16 is used as the backbone model and trained on the image ground-truth with the sigmoid cross-entropy loss. The model is trained for 30 epochs with the SGD optimization and the learning rate decreases from 0.001 to 1/10 every ten epochs.

KNN. This is a simple and widely used annotation baseline model [58]. KNN annotates the image by retrieving the m nearest neighbors, and tags are assigned based on the occurrence among neighbors. We set $m = 50$, and use the 4096-d visual representation from Multi-CNN last two fully-connected layer.

CNN+LSTM. This is a equivalent model proposed in [21]. CNN+LSTM annotates images by leveraging the image visual-tag relationship and tag pair dependencies simultaneously at the image level. Tags associated with the image are ranked as an ordered sequence based on the occurrence rate. The image is first sent to the CNN for the visual representation extraction, then the visual representation and tags are recurrently encoded and fed to the LSTM to infer the next tag. The embedded dimensions for the visual representation and tag are 512 and 256 respectively.

Compared Methods. We compare our proposed model with several state-of-the-art annotation methods, including CCA [59], TagProp [15], HCP [26] TagFeature [60], and TagExample [51]. For fair comparisions, the image representation used in these model are 4096-d vector from Multi-CNN last two fully-connected layer. Moreover, since these methods are not originally designed to take the metadata into consideration, to further show the effectiveness of our proposed model, we implement five comparison experiments that utilize the metadata, including KNN*, CCA*, TagProp*, TagFeature* and TagExample*, where * stands for the image features in these models are obtained after the proposed visual representation learning.

4.4. Results on Image Annotation

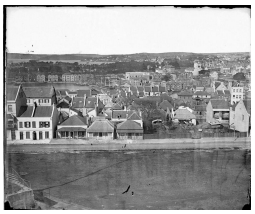
Tab. 1 shows that our proposed model achieves the best performance on all evaluation metrics against all baselines and state-of-the-art methods. As expected, all methods outperform the baseline RandomGuess in a large margin, which proves the learning from the image ground-truth is necessary for the effective annotation. Multi-CNN baseline directly models the relationship between the image visual content and associated tags, since our model uses metadata for the visual representation learning and summarise the semantic representation from image neighbors, the performance of our model surpasses this baseline. KNN baseline only uses the image visual representation to retrieve relevant neighbors to vote the annotation; therefore, with the more accurate visual representation obtained from the metric learning, our model also outperforms this baseline. CNN+LSTM is the state-of-the-art model for the image annotation, which models the image-tag and the tag sequence correlation by utilizing a unified framework of CNN and LSTM. However, it is hard for LSTM to directly model the diverse tag correlation as an ordered sequence.

As for the compared methods, CCA [59] learns a latent space to embed the image and tag representation together to enhance the representation, and the similar idea is also used in TagFeature [60], where the predictions from tag classifiers are concatenated with the image representation to retrain the annotation

Table 1: Results of the Image Annotation.

Method	imAP	mAP	O_P	O_R	O_{F1}	C_P	C_R	C_{F1}
RandomGuess [57]	0.008	0.009	0.008	0.011	0.009	0.008	0.000	0.000
Multi-CNN [52]	0.411	0.346	0.541	0.315	0.398	0.515	0.200	0.288
KNN [58]	0.330	0.282	0.399	0.360	0.378	0.425	0.143	0.214
CNN+LSTM [21]	0.410	0.342	0.539	0.316	0.399	0.517	0.314	0.391
CCA [59]	0.424	0.367	0.498	0.384	0.434	0.509	0.286	0.366
TagProp [15]	0.430	0.366	0.555	0.329	0.413	0.520	0.257	0.344
HCP [26]	0.449	0.390	0.536	0.394	0.454	0.519	0.308	0.387
TagFeature [60]	0.417	0.357	0.520	0.297	0.378	0.509	0.229	0.315
TagExample [51]	0.425	0.368	0.525	0.353	0.422	0.517	0.314	0.391
KNN*	0.391	0.331	0.475	0.350	0.403	0.464	0.286	0.354
CCA*	0.443	0.386	0.500	0.410	0.450	0.513	0.310	0.387
TagProp*	0.475	0.408	0.541	0.403	0.462	0.529	0.257	0.346
TagFeature*	0.425	0.359	0.499	0.378	0.430	0.506	0.305	0.381
TagExample*	0.441	0.376	0.511	0.389	0.442	0.513	0.342	0.410
Our Model	**0.511**	**0.445**	**0.607**	**0.416**	**0.494**	**0.575**	**0.371**	**0.451**

model. We also learn the hidden states of both image and semantic representation for the annotation, however, since we obtain the semantic representation from image neighbors and we have better image visual representation based on the metric learning, our model outperforms both methods. TagProp [15] is a trained nearest neighbor approach, which directly maximizes the probability of the tag distribution in training set by the integration of the metric learning. Different from this method, our metric learning for the visual representation is performed under the guidance of image metadata instead of modeling the tag distribution, considering the diversity of tags. The superior performance against TagProp verifies this assumption. Instead of operating at the image level, HCP [26] proposes to apply the bottom-up proposal method to generate image regions, captures image regions with associated tags and fuse them together. However, considering the various range of the image visual content in the training set, the bottom-up proposal methods like BING [27] fail to capture valid image regions in most cases. TagExample [51] explores both positive and negative training samples by analyzing the tag relevance with respect to the image content. Compared to this method, we use the metadata similarity for the image representation learning to avoid any visual ambiguous caused by the



(a)
Ground-Truth: historic building; house; panoramic view; suburb; city view
KNN: historic building; city street; city view
Multi-CNN: historic building; panoramic view
Our Model: historic building; house; city street; panoramic view; city view, suburb

(b)
Ground-Truth: festival; procession; official event; bridge; float procession; band
KNN: festival; anniversary; procession; bridge
Multi-CNN: festival; procession; official event; bridge;
Our Model: festival; procession; official event; bridge; float procession

(c)
Ground-Truth: historic building; children; teacher; bank roof
KNN: historic building;
Multi-CNN: historic building;
Our Model: historic building; children; teacher

(d)
Ground-Truth: crowd; evening clothes; theater foyer
KNN: crowd; theater; association; audience
Multi-CNN: crowd
Our Model: crowd; evening clothes; theater foyer

(e)
Ground-Truth: streetscape; coronation; decoration; city street
KNN: streetscape; commercial establishment; flag
Multi-CNN: crowd; city street; procession; decoration
Our Model: city street; streetscape; decoration

(f)
Ground-Truth: group people; uniform; football team
KNN: group people; football team; children
Multi-CNN: group people; football team
Our Model: group people; uniform; football team

Figure 5: Some example annotation results on the heritage image collection.

tag diversity and apply tag relevance analysis later to help summarise the semantic representation from neighbors. The advanced performance indicates the effectiveness of our proposed model. These compared methods are not originally designed to take the metadata into consideration, which is one of the advantages

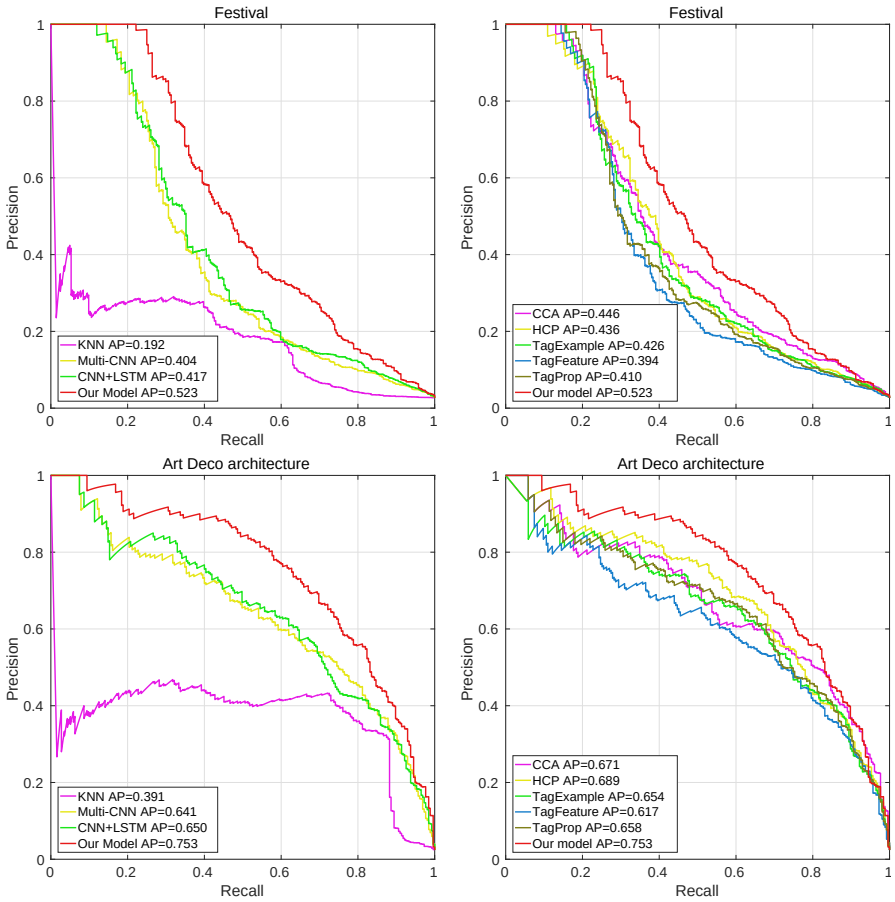


Figure 6: (a) The first row is the PR-curves of the tag 'festival' compared with baselines and state-of-the-art methods. (b) The second row is the PR-curves of the tag 'Art Deco architecture' compared with baselines and state-of-the-art methods. The average precision is also given in the figure. Better view in color.

of our proposed model when we tackle the diverse image annotation problem on the heritage image collection. To include the metadata in these models, we also report the results of the *-models in Tab. 1. As we can see, the *-models outperform their original models in most cases, which validate the effectiveness of the learned image visual representation. Moreover, our model achieves best performance against *-models again shows the significance of the annotation via collective knowledge. Annotation examples of the proposed model are shown in

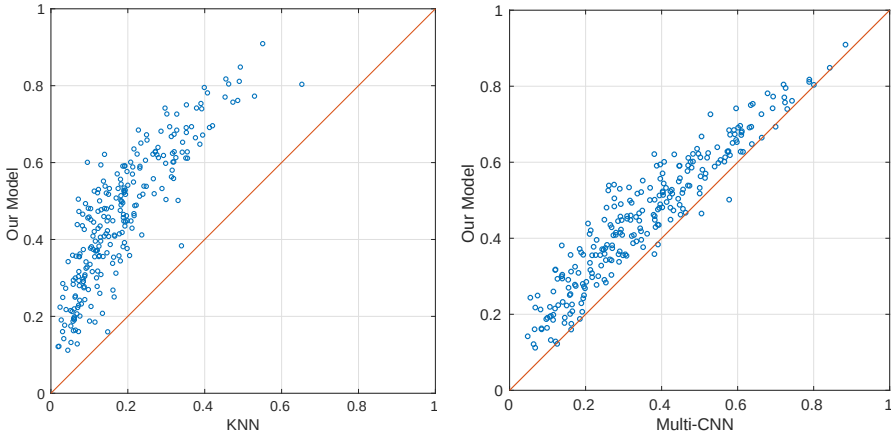


Figure 7: The comparisions of the average precision (AP) values between the proposed model and baselines on all tags. The left one is the AP values of our model against KNN baseline, while the right one is our model against Multi-CNN baseline. Better view in color.

Fig. 5.

In Fig. 6, we show the precision-recall curves of our model against baselines and compared methods on the tag 'festival' and 'Art Deco architecture.' As we can see, our model achieves the best performance compared against baselines and state-of-the-art methods. The tag 'festival' and 'Art Deco architecture' are both semantical and related to the image visual content. By combining the visual and semantic representation, we show large improvements compared to baselines. We also compare the average precisions (AP) between the proposed model and KNN/Multi-CNN baselines. The results are shown in Fig. 7, where the x-ray stands for the baseline AP value, and y-ray is the corresponding proposed model's AP value. As we can see, the majority of the values are above the $y = x$, which proves the effectiveness of our model on the whole tag set.

4.5. Ablation Study

We conduct the comprehensive ablation study to further investigate the individual contribution of each component in the proposed model. In this section, we first analyze the effectiveness of the metric learning on the visual representation. Then we investigate the contribution of the semantic summarisation for

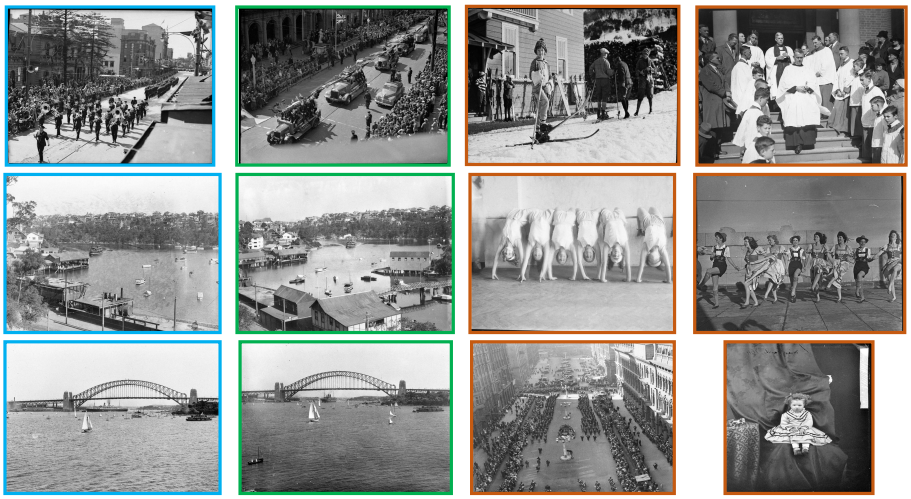


Figure 8: The examples of training pairs for the metric learning based on the metadata similarity. The first column is the query image with the blue box, the second column is its positive pair with the green box, and the third and fourth columns are negative ones with brown boxes.

the annotation.

Table 2: Results of the Ablation Study.

Method	imAP	mAP	O_P	O_R	O_{F1}	C_P	C_R	C_{F1}
Multi-CNN	0.411	0.346	0.541	0.315	0.398	0.515	0.200	0.288
MeL+Fc	0.455	0.403	0.508	0.410	0.454	0.514	0.286	0.367
Our Model	**0.511**	**0.445**	**0.607**	**0.416**	**0.494**	**0.575**	**0.371**	**0.451**

4.5.1. Visual Representation

To conduct quantitative analysis, we extract the image visual representation after the metric learning and fed into the fully-connected network to train an annotation model. The network is trained with the sigmoid cross-entropy loss, same as Multi-CNN baseline. We note this model as MeL+Fc and show the results in Tab. 2. As we can see, MeL+Fc shows large improvement compared to baseline Multi-CNN, which proves the metric learning based on the metadata similarity is necessary to obtain the accurate visual representation.

(a)

Ground-Truth:
crowd; festival; city street; procession; band; anniversary

Suggestions from neighbors:
crowd; procession; band; festival; spectator; city street; association; anniversary;

(b)

Ground-Truth:
official event; harbor; bridge

Suggestions from neighbors:
harbor; bridge; official event; bridge construction; ship

(c)

Ground-Truth:
evening clothes; table setting; dance

Suggestions from neighbors:
evening clothes; table setting; dance; group people; convention;

(d)

Ground-Truth:
theatrical costume; actor; theatrical production

Suggestions from neighbors:
portrait; theatrical costume; actor; theatrical production

Figure 9: Some examples of tag suggestions retrieved from image neighbors. We summarise these suggestions as the semantic representation.

In Fig. 8 we show some examples of training pairs for the metric learning based on the metadata similarity. As we can see, it is reasonable to generate positive and negative samples for the visual representation learning based on the metadata similarity.

4.5.2. Semantic Summarization

As we mentioned in Sec. 3.3, after we obtain the visual representation, we retrieve nearest neighbors and utilize them to generate the semantic representation. In Tab. 2, we observe the improvement from the comparison between MeL+Fc and our final model, which proves that summarise semantic information from neighbors can help boost the annotation performance. We show some examples of tag suggestions retrieved from image neighbors in Fig. 9. As we

can see, these suggestions can assist the annotation model to infer all tags.

5. Conclusions

Images are the visual reflections of the real world. Building an automatic annotation model is a crucial step to understand these images as well as efficiently retrieve them. In this paper, we ground the diverse image annotation on the heritage collection and conduct the image representation learning based on collective knowledge. The proposed image representation is consist of both visual and semantic information. That is, we allocate the image neighbors by measuring the metadata similarity and obtain the visual representation of the image by performing the metric learning within the neighborhood. Moreover, we generate the semantic representation by summarizing the neighborhood annotations based on the tag relevance. Comprehensive experiments are conducted on the heritage image dataset, and advanced results against compared models indicate the significance of the proposed method. Since the textual based metadata is mainly used in our work, the precision of the provided metadata can influence the annotation accuracy, we will investigate other types of metadata and the combination of late-fusion methods to improve the robustness as our future work.

References

[1] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: Proc. Eur. Conf. Comp. Vis., 2014, pp. 740–755.

[2] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vision 88 (2) (2010) 303–338.

[3] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, NUS-WIDE: a real-world web image database from national university of singapore, in: Proceedings of the 8th ACM International Conference on Image and Video

Retrieval, CIVR 2009, Santorini Island, Greece, July 8-10, 2009, 2009, p. 48.

[4] M. J. Huiskes, M. S. Lew, The MIR flickr retrieval evaluation, in: Proceedings of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval, MIR 2008, Vancouver, British Columbia, Canada, October 30-31, 2008, 2008, pp. 39–43.

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

[6] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.

[7] A. Oliva, A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope, Int. J. Comput. Vision 42 (3) (2001) 145–175.

[8] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[9] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 27.

[10] N. S. Altman, An introduction to kernel and nearest-neighbor nonparametric regression, The American Statistician 46 (3) (1992) 175–185.

[11] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR abs/1409.1556.

[12] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2016, pp. 770–778.

[13] A. Gordo, J. Almazán, J. Revaud, D. Larlus, Deep image retrieval: Learning global representations for image search, in: Proc. Eur. Conf. Comp. Vis., Springer, 2016, pp. 241–257.

[14] F. Zhao, Y. Huang, L. Wang, T. Tan, Deep semantic ranking based hashing for multi-label image retrieval, in: Proc. IEEE Conf. Comp. Vis. Patt. Recogn., IEEE, 2015, pp. 1556–1564.

[15] M. Guillaumin, T. Mensink, J. Verbeek, C. Schmid, Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation, in: Proc. IEEE Int. Conf. Comp. Vis., 2009, pp. 309–316.

[16] Y. Gong, Y. Jia, T. Leung, A. Toshev, S. Ioffe, Deep convolutional ranking for multilabel image annotation, CoRR abs/1312.4894.

[17] W. K. Härdle, L. Simar, Canonical correlation analysis, in: Applied Multivariate Statistical Analysis, Springer, 2015, pp. 443–454.

[18] N. Rasiwasia, D. Mahajan, V. Mahadevan, G. Aggarwal, Cluster canonical correlation analysis, in: Artificial Intelligence and Statistics, 2014, pp. 823–831.

[19] S. Z. Li, Markov random field modeling in image analysis, Springer Science & Business Media, 2009.

[20] J. Lafferty, A. McCallum, F. Pereira, et al., Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proc. Int. Conf. Mach. Learn., Vol. 1, 2001, pp. 282–289.

[21] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, W. Xu, Cnn-rnn: A unified framework for multi-label image classification, in: Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2016, pp. 2285–2294.

[22] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proc. Advances in Neural Inf. Process. Syst., 2012, pp. 1097–1105.

[23] J. Zhang, J. Zhang, Q. Wu, Q. Wu, J. Xu, J. Lu, R. Phua, K. Curr, Z. Tang, Historical image annotation by exploring the tag relevance, in: 2017 4th

IAPR Asian Conference on Pattern Recognition (ACPR), IEEE, 2017, pp. 640–645.

[24] Z. Lu, L. Wang, Learning descriptive visual representation for image classification and annotation, Pattern Recogn. 48 (2) (2015) 498–508.

[25] X. Xue, W. Zhang, J. Zhang, B. Wu, J. Fan, Y. Lu, Correlative multi-label multi-instance image annotation, in: Proc. IEEE Int. Conf. Comp. Vis., IEEE, 2011, pp. 651–658.

[26] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, S. Yan, Hcp: A flexible cnn framework for multi-label image classification, IEEE Trans. Pattern Anal. Mach. Intell. 38 (9) (2016) 1901–1907.

[27] M.-M. Cheng, Z. Zhang, W.-Y. Lin, P. Torr, Bing: Binarized normed gradients for objectness estimation at 300fps, in: Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2014, pp. 3286–3293.

[28] T. Uricchio, L. Ballan, L. Seidenari, A. D. Bimbo, Automatic image annotation via label transfer in the semantic space, Pattern Recogn. 71 (2017) 144–157.

[29] X. Ke, M. Zhou, Y. Niu, W. Guo, Data equilibrium based automatic image annotation by fusing deep model and semantic propagation, Pattern Recogn. 71 (2017) 60–77.

[30] Y. Gong, Q. Ke, M. Isard, S. Lazebnik, A multi-view embedding space for modeling internet images, tags, and their semantics, Int. J. Comput. Vision 106 (2) (2014) 210–233.

[31] V. N. Murthy, S. Maji, R. Manmatha, Automatic image annotation using deep learning representations, in: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Shanghai, China, June 23-26, 2015, 2015, pp. 603–606.

[32] C. Gong, D. Tao, S. J. Maybank, W. Liu, G. Kang, J. Yang, Multi-modal curriculum learning for semi-supervised image classification, IEEE Trans. Image Process. 25 (7) (2016) 3249–3260.

[33] S. J. Hwang, K. Grauman, Learning the relative importance of objects from tagged images for retrieval and cross-modal search, Int. J. Comput. Vision 100 (2) (2012) 134–153.

[34] J. K. Bradley, C. Guestrin, Learning tree conditional random fields, in: Proc. Int. Conf. Mach. Learn., 2010, pp. 127–134.

[35] M.-L. Zhang, K. Zhang, Multi-label learning by exploiting label dependency, in: Proc. ACM Int. Conf. Knowledge discovery & data mining, 2010, pp. 999–1008.

[36] Y. Yang, Z. Huang, Y. Yang, J. Liu, H. T. Shen, J. Luo, Local image tagging via graph regularized joint group sparsity, Pattern Recogn. 46 (5) (2013) 1358–1368.

[37] M. Tan, Q. Shi, A. van den Hengel, C. Shen, J. Gao, F. Hu, Z. Zhang, Learning graph structure for multi-label image classification via clique generation, in: Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2015, pp. 4100–4109.

[38] C. Gong, D. Tao, J. Yang, W. Liu, Teaching-to-learn and learning-to-teach for multi-label propagation, in: Proc. Conf. AAAI, 2016, pp. 1610–1616.

[39] J. Johnson, L. Ballan, L. Fei-Fei, Love thy neighbors: Image annotation by exploiting image metadata, in: Proc. IEEE Int. Conf. Comp. Vis., 2015, pp. 4624–4632.

[40] Y. Jin, L. Khan, L. Wang, M. Awad, Image annotations by combining multiple evidence & wordnet, in: Proc. ACM Int. Conf. Multimedia., 2005, pp. 706–715.

[41] N. Sawant, R. Datta, J. Li, J. Z. Wang, Quest for relevant tags using local interaction networks and visual content, in: Proc. ACM Int. Conf. Multimedia Retrieval, ACM, 2010, pp. 231–240.

[42] J. Liu, Z. Li, J. Tang, Y. Jiang, H. Lu, Personalized geo-specific tag recommendation for photos on social websites, IEEE Trans. Multimedia 16 (3) (2014) 588–600.

[43] G. Kim, E. P. Xing, Time-sensitive web image ranking and retrieval via dynamic multi-task regression, in: Proceedings of the sixth ACM international conference on Web search and data mining, ACM, 2013, pp. 163–172.

[44] P. McParlane, S. Whiting, J. Jose, Improving automatic image tagging using temporal tag co-occurrence, in: International Conference on Multimedia Modeling, Springer, 2013, pp. 251–262.

[45] S. Zhu, C.-W. Ngo, Y.-G. Jiang, Sampling and ontologically pooling web images for visual concept learning, IEEE Trans. Multimedia 14 (4) (2012) 1068–1078.

[46] H. Xu, J. Wang, X.-S. Hua, S. Li, Tag refinement by regularized lda, in: Proc. ACM Int. Conf. Multimedia., ACM, 2009, pp. 573–576.

[47] B. Sigurbjörnsson, R. Van Zwol, Flickr tag recommendation based on collective knowledge, in: Proc. Int. Conf. World Wide Web., ACM, 2008, pp. 327–336.

[48] A. Sun, S. S. Bhowmick, Quantifying tag representativeness of visual content of social images, in: Proc. ACM Int. Conf. Multimedia., 2010, pp. 471–480.

[49] X. Li, C. G. Snoek, M. Worring, Learning social tag relevance by neighbor voting, IEEE Trans. Multimedia 11 (7) (2009) 1310–1322.

[50] D. Liu, X.-S. Hua, L. Yang, M. Wang, H.-J. Zhang, Tag ranking, in: Proc. Int. Conf. World Wide Web., ACM, 2009, pp. 351–360.

[51] X. Li, C. G. Snoek, Classifying tag relevance with relevant positive and negative examples, in: Proc. ACM Int. Conf. Multimedia., ACM, 2013, pp. 485–488.

[52] J. Zhang, J. Zhang, J. Lu, C. Shen, K. Curr, R. Phua, R. Neville, E. Edmonds, Slnsw-uts: A historical image dataset for image multi-labeling and retrieval, in: Proc. Int. Conf. Digital Image Computing: Techniques and Applications, 2016, pp. 1–6.

[53] L. Zhao, K. Wang, B. Do, Sherlocknet: Exploring 400 years of western book illustrations with convolutional neural networks, Technical Report (2016) 1–9.

[54] S. Chopra, R. Hadsell, Y. LeCun, Learning a similarity metric discriminatively, with application to face verification, in: Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2005, pp. 539–546.

[55] R. L. Cilibrasi, P. M. Vitanyi, The google similarity distance, IEEE Trans. Knowl. Data Eng. 19 (3).

[56] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proc. Advances in Neural Inf. Process. Syst., 2013, pp. 3111–3119.

[57] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. Snoek, A. D. Bimbo, Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval, ACM Computing Surveys 49 (1) (2016) 14.

[58] A. Makadia, V. Pavlovic, S. Kumar, A new baseline for image annotation, in: Proc. Eur. Conf. Comp. Vis., Springer, 2008, pp. 316–329.

[59] V. N. Murthy, S. Maji, R. Manmatha, Automatic image annotation using deep learning representations, in: Proc. ACM Int. Conf. Multimedia Retrieval, ACM, 2015, pp. 603–606.

[60] L. Chen, D. Xu, I. W. Tsang, J. Luo, Tag-based image retrieval improved by augmented features and group-based refinement, IEEE Trans. Multimedia 14 (4) (2012) 1057–1067.