General Knowledge Embedded Image Representation Learning

Peng Cui, *Member, IEEE*, Shaowei Liu, and Wenwu Zhu, *Fellow, IEEE*

***Abstract*—Image representation learning is a fundamental problem in understanding semantics of images. However, traditional classification-based representation learning methods face the noisy and incomplete problem of the supervisory labels. In this paper, we propose a general knowledge base embedded image representation learning approach, which uses general knowledge graph, which is a multitype relational knowledge graph consisting of human commonsense beyond image space, as external semantic resource to capture the relations of concepts in image representation learning. A relational regularized regression CNN (R^3CNN) model is designed to jointly optimize the image representation learning problem and knowledge graph embedding problem. In this manner, the learnt representation can capture not only labeled tags but also related concepts of images, which involves more precise and complete semantics. Comprehensive experiments are conducted to investigate the effectiveness and transferability of our approach in tag prediction task, zero-shot tag inference task, and content-based image retrieval task. The experimental results demonstrate that the proposed approach performs significantly better than the existing representation learning methods. Finally, observation of the learnt relations show that our approach can somehow refine the knowledge base to describe images and label the images with structured tags.**

***Index Terms*—Image representation learning, knowledge base, multirelational graph embedding.**

I. Introduction

IMAGE representation is a fundamental problem in various image applications. In recent years, we have witnessed a fast advancement of image representation learning from hand-crafted features to learning image representation from scratch through deep neural network models. As deep models are trained in an end-to-end fashion, the learned image representations can perform much better than hand-crafted features in the target applications, as long as the training data is of sufficient quality and quantity. The side effect, however, is that the goodness of the learned representations heavily depends on the input training data, especially the supervisory information of the target application. If the goal is to learn image representations to bridge the semantic gap, the completeness, preciseness and richness of the semantic labels of images are decisive to the capability of the learned image representations to bridge the pixel data and semantics.

In existing image representation learning works, expert annotated semantic labels (e.g. ImageNet) are commonly used as the supervisory information. But the completeness and the quantity of these expert labels cannot be guaranteed due to expensive and limited human labors. Recently, a paucity of works start to exploit social tags (e.g. in Flickr) as the supervisory information. Although the problem of completeness is alleviated to some extent, the noisy and imprecise tags seriously affect the quality of the learned representations. In order to tradeoff the completeness and preciseness, [1], [2] consider to exploit the "isA" relations in image knowledge graph ImageNet to extend the labels of images. However, very small improvement was achieved over the methods without extending labels. One plausible reason is that the richness of ImageNet is quite limited and thus cannot bring much additional information. Meanwhile, most images in ImageNet are with high-quality and single salient object, which lead the learned image representations hard to apply in wild image applications. How to learn image representations from data guided by supervisory information with satisfactory completeness, preciseness and richness, is still an open problem.

In this paper, we explore the possibility of embedding general knowledge graph into the representation learning of wild images with multiple tags. Here the general knowledge graph is referred to the multi-type relational knowledge bases consisting of concept-level common-senses defined beyond image space. Let us take ConceptNet [3], a general commonsense knowledge base, as a representative. There are 3.4 million concepts in English in total, and 56 kinds of relations are included, such as "IsA", "AtLocation", "PartOf". By properly embedding these relational knowledge into image representations, we can get two notable benefits. First, the traditional supervisory information, like tags, can be extended, filtered and relationally structured, and thus can address the problem of completeness, preciseness and richness in supervisory information. Additionally, through end-to-end training in deep models, the learned image representations can well support relational reasoning if the supervisory information consists of relational tags. That is, we can predict precise tags for an image with relational structure. In contrast with the traditional flat-structured tags, the relational structured tags have larger potential in deeply understanding images and inferring user intents in different semantic levels or aspects.

Manuscript received November 15, 2016; revised April 28, 2017; accepted June 6, 2017. Date of publication July 11, 2017; date of current version December 14, 2017. This work was supported in part by the National Program on Key Basic Research Project under Grant 2015CB352300 and in part by the National Natural Science Foundation of China under Grant U1611461, Grant 61521002, Grant 61531006, and Grant 61210008. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Marco Bertini. *(Corresponding author: Peng Cui.)*

The authors are with the Department of Computer Science, Tsinghua University, Beijing 100084, China (e-mail: cuip@tsinghua.edu.cn; liushaowei@mails.tsinghua.edu.cn; wwzhu@tsinghua.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2017.2724843

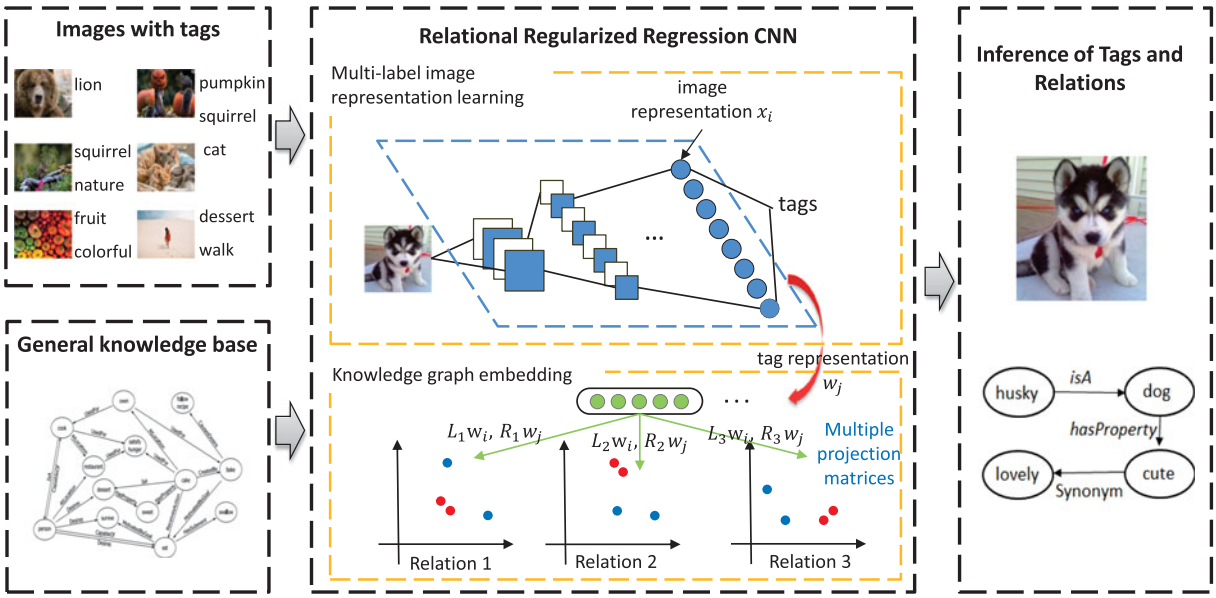

Fig. 1. Framework of general knowledge embedded image representation learning.

However, we face the following challenges in embedding multi-type relational knowledge into image representations:

1) *The semantic spaces of images and the concepts in knowledge graph are not consistent:* We cannot simply learn the image representations and concept embeddings independently and then bridge them together because it is very difficult to make the semantic spaces of images consistent to the concepts in knowledge graph. First, there are a lot of relations in knowledge graph that cannot be used to describe images, thus they are not helpful in knowledge embedding. Second, if the existed image representation has lost some semantic information, bringing it to knowledge embedding will confuse the model. Thus, we need to jointly optimize image representation learning problem and knowledge graph embedding problem in a framework.

2) *The multiple relations are not equally effective in image representation learning:* Different from traditional semantic hierarchy, there are multiple types of relations in a knowledge graph. However, they are not equally effective to describe images. For example, the relations "CausedBy" and "MotivatedByGoal" can hardly be applied to image tagging. Thus, we cannot simply conduct image annotation and then extend related tags. To solve this problem, we need to consider the relations and their corresponding visual contents at the same time.

3) *Both symmetric and asymmetric relations need to be incorporated:* Knowledge graph is a directed graph, which includes asymmetric relations, such as "IsA", "MadeOf", and symmetric relations, such as "RelatedTo" and "Synonym". We need to consider the types of relations and incorporate all of them.

In order to address the above challenges, we propose a new framework to embed the multiple relational multi-type relational knowledge into image representations, as shown in Fig. 1. In our approach, we design a Relational Regularized Regression Convolutional Neural Network (R^3CNN), which jointly optimize the multi-label image representation learning problem and knowledge graph embedding problem. In image representation learning task, a regression CNN is adopted to learn image representation from image-tag data. In knowledge graph embedding task, we map a concept in knowledge graph into multiple hidden spaces for different types of relations. The mapping functions of left concept and right concept in a relation are different, which makes the distance of two concepts asymmetric. Therefore, both of visual contents of images and relations in knowledge graph are involved in the model through joint optimization. The learnt representation can capture the information of not only labeled tags, but also relations of concepts in knowledge graph. In addition, for zero-shot tags unseen in training data, we can easily infer their representations based on the existed tag embeddings in knowledge graph, and then predict its relevance to the images.

It is worthy to highlight the contributions of the proposed approach as follows:

1) In order to address the problem of completeness, preciseness and richness of supervisory information in image representation learning, we propose a new framework to embed knowledge graph into image representations.
2) We propose a novel R^3CNN model to jointly optimize the multi-label image representation learning problem and knowledge graph embedding problem in a deep model, where both symmetric and asymmetric relations are incorporated.
3) Extensive experiments are conducted to demonstrate the effectiveness of the proposed method in tag prediction, zero-shot tag inference, and content-based image retrieval tasks. The results demonstrate that our method significantly outperforms state-of-the-art image representation learning methods.

The rest of the paper is organized as follows: Section II gives a brief comparison of related works. Section III introduces some statics and observations of data and defines our problem. In Section IV, we present the proposed R^3CNN model and introduce the applications of our approach. Then, the experimental results are reported to demonstrate the effectiveness of our approach in Section V. Finally, Section VI summarizes the paper.

II. Related Work

A. Deep-Model-Based Image Representation Learning

To bridge the gap between low-level features and high-level semantics, deep model based image representation learning methods show their superiority to traditional hand-crafted features. In these methods, Convolutional Neural Network (CNN) [4], [5] is widely used and proved to be very effective. These works usually use image classification dataset with single label(such as ImageNet). To date, this problem is well solved and its top-5 error is reduced to less than 5% [6]. However, CNN based representation learning methods rely on the supervisory information in the dataset very much. The completeness and richness of supervisory information determines the effectiveness of the learned representations. The concept of the salient object is not enough to cover the all the semantic information in an image due to the lack of motion, attributes, background, etc. In recent years, some researchers explore to learn image representation based on the corresponding textual information (e.g. tags and sentences) because such textual information can cover more semantics in an image. Some of these works treat each tag as an independent classifier, thus convert this problem into a multi-label classification problem [7]–[10]. Also, [11], [12],[13] use multimodal methods to bridge the visual modality and text modality, and [14] proposes a novel deep relative attributes (DRA) algorithm to learn visual features. In order to alleviate the intention gap problem in image applications, [15] proposes an Asymmetric Multi-task CNN model to embed social signals into image representations. However, all the above solutions face the the noisy and incomplete problems of the wild textual data. Intuitively, only using image-text data cannot solve this problem because such wild data cannot provide enough clues to filter the noisy tags and complete the missing tags.

B. Image Understanding With Extra Semantic Resources

Image annotation with noisy tags is a well-studied problem. Traditional methods usually use nearest-neighbor-based approaches [16], [17], [18], where the semantic meaning of the words is ignored. In recent years, researchers explore to consider the relation of labels with extra semantic resources to capture the relation labels. Visual-semantic embedding models [1], [2] to embed visual ontology ImageNet into image classification task. It considers the ontology of the 1000 classes in ImageNet to make the results more semantically reasonable and can be used for zero-shot prediction (predict unseen categories). However, ImageNet ontology cannot include all the semantic relations of the concepts, such as property and location. To import general world knowledge into image understanding, Xie *et al.* uses the general knowledge base ConceptNet as semantic resources to learn tags relations for image tagging [19]. However, in this work, the tag relations is evaluated by simple similarity in a common space, which cannot reflect the property of multi-relation and single-direction in knowledge base. Im *et al.* uses DBPedia [20] to extend the annotation results based on the existed tags to improve recall [21]. [22] constructs a semantic-visual knowledge base to encode the rich event-centric concepts to enhance video event recognition. [23] proposes a cross-domain learning method to classify web multimedia objects by transferring the correlation knowledge among different information sources. However, for vision tasks, there are a lot of useless relations in knowledge base. Besides, there are a lot of ambiguous tags. Thus, tag extension should consider the visual content, which is better to be conducted at representation learning stage rather than tagging stage. More recently, knowledge graph embedding and their combination with visual information arouse some research interest [24], [25]. [26] borrows the idea of relation representation in knowledge graph and propose a novel visual translation network for visual relation detection. This paper differs in that we exploit the semantic relationships (e.g. UsedFor, IsA) among labels for representation learning and tagging, while these works exploit the visual relationships (e.g. ride-on, walk-to) among object labels for image content understanding.

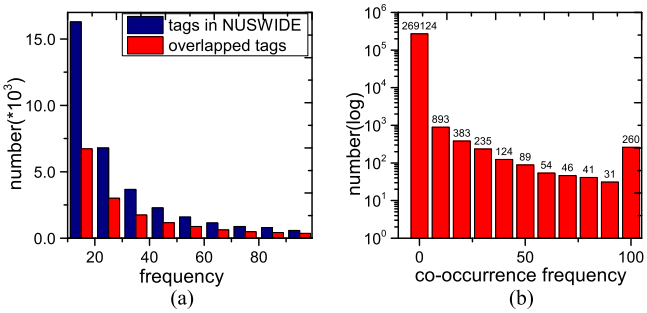

Fig. 2. Statistics of overlapped tags and relations in NUSWIDE and ConceptNet.

III. Preliminary Study

In this section, we preliminarily demonstrate the feasibility of exploiting general knowledge graph in image representation learning by data statistics, and then give the formal problem definition.

A. Statistics and Observation

In this work, we use two data resources, a visual resource NUSWIDE and a semantic resource ConceptNet. NUSWIDE includes 269,648 images crawled in Flickr[1] with 425,001 tags labeled by users. ConceptNet is a knowledge graph consisting of general common sense knowledge in multi-language. There are 3.4 million concepts, 56 kinds of relations and 11 millon relations in total. We conduct statistics on the above resources to answer the following two questions: 1) Whether these two resources can be bridged? 2) Can the relations in ConceptNet help to understand the tag relations in describing images?

To answer the first question, we give statistics on how many shared tags[2] occur in the two resources. In the 425,001 tags in NUSWIDE, there are 92,595 tags also occur in ConceptNet. Fig. 2(a) illustrates the distribution of number of tags versus the occurrence frequency in NUSWIDE among all tags in NUSWIDE and the shared tags of NUSWIDE and ConceptNet.

[1][Online]. Available: http://www.flickr.com

[2]A word is called "tag" in NUSWIDE and "concept" in ConceptNet. In this paper, we denote a shared entity as tag or label without confusion.

We can observe that the higher occurrence frequency of tags, the higher possibility of sharing between NUSWIDE and ConceptNet. For the tags that occur more than 50 times in NUSWIDE, more than 50% of them also occur in ConceptNet. Thus, ConceptNet can encompass most of the common tags in NUSWIDE, and the shared tags can play the role of a bridge between these two resources.

For the second question, we use the most typical tag relation in NUSWIDE, i.e., co-occurrence relation to observe whether it is enough to describe tag relations without ConceptNet. Due to the fact that ConceptNet has multiple types of relations, we use the relation types "IsA", "AtLocation", and "InstanceOf" as representation, which are very useful for image tagging in our observation. For any two overlapped tags, if they have at least one of the above relations in ConcepNet, we conduct count their co-occur frequency in NUSWIDE. The statistical histogram is illustrated in Fig. 2(b). We can observe that most of the relational pairs do not occur in ConceptNet. It indicates that co-occurrence relation cannot include the above three relations in ConceptNet. Besides, There are 33,401 tag pairs occur more than 100 times in NUSWIDE. While only 3,412 of them occur in ConceptNet. It indicates that the relations in ConceptNet and co-occurrence in NUSWIDE are complementary.

Based on the above statistics and observations, we can find that bridging NUSWIDE and ConceptNet together and incorporating the relations of these resources can help us better understand image semantics.

B. Notation and Problem Definition

Our target is to jointly learn image representation x_i for image i and tag representation w_j for tag j, which can satisfy the requirements of two tasks: multi-label image representation learning and knowledge graph embedding.

Definition 1. (Image Representation Learning): In image representation task, out goal is to learn x_i, w_i to make $y_{ij} = f(x_i, w_j)$ for a given function $f(\cdot)$, where y_{ij} is the ground truth that denote whether tag j belongs to image i.

Definition 2. (General Knowledge Graph): A general knowledge graph is a multi-type relational graph whose vertices are concepts (one or several words), and edges are the types of relations between two tags. A relation is represented as a triplet $< i, j, p >$, which means concept i and concept j has relation type p. To be noted that if the relation type p is symmetric such as "RelatedTo", both of $< i, j, p >$ and $< j, i, p >$ will occur in our knowledge graph. Thus, we can simply regard the knowledge graph as a directed graph can consider the asymmetric relations.

Definition 3. (Knowledge Graph embedding): Let $d_{ij}^p = g_p(w_i, w_j)$ denote the distance of concept i and j in the space p. Knowledge graph embedding task is to find the optimal concept embedding w_i for tag i, which makes the distance d_{ij}^p consistent to the relations knowledge graph, i.e., d_{ij}^p is small when $< i, j, p >$ exists in the knowledge graph, and vise versa.

In image representation learning and knowledge graph embedding, tag representation w_i is a common variable, which bridges these two tasks together for joint optimization.

IV. Relational Regularized Regression CNN

In this section, we introduce the proposed model R^{3}CNN, where the image representation learning and knowledge graph embedding are jointly optimized.

A. Multilabel Image Representation Learning

Most existing CNN models are designed for single-label classification task, which cannot be directly applied to multi-label image representation learning problem. Let us take AlexNet [4] as an example. In AlexNet, the activation function for fc7 (the last fully connection) layer is softmax. Softmax is a function considering multiple tags as a joint probabilistic distribution, which makes different tags mutually exclusive. As wild images are often labeled with noisy and incomplete tags, we can not know the real distribution of multiple labels. Using the observed noisy tags instead with softmax function will lead the learned distribution to have a large discrepancy with the real distribution. Therefore, the softmax function cannot work well in multi-label image representation problem. To solve this problem, we propose to use sigmoid function as in [27] to replace softmax function, and regard each tag to be independent. In this way, the influence of noisy and incomplete tags will be reduced in the objective function by reducing the weight of negative samples. In addition, the original loss function in AlexNet is based on cross entropy, which is used to measure the distance of two probabilistic distribution. In our case, as the output is not a joint distribution of multi-labels but independent probability for each label, we use L2 loss as our loss function. In this model, the structure of the first 10 layers is the same with AlexNet, including 5 convolution layers, 3 pooling layers, and 2 fully connection layers. Then, we use sigmoid function as the activation function of fc7 layer and adopt Euclidean loss ($L2$ loss) instead of cross entropy . Here the loss function is

$$\mathcal{L}_1 = \sum_{i=1}^{N} \left(\sum_{y_{ij}=1} \frac{1}{2}(y_{ij} - t_{ij})^2 + \alpha \sum_{y_{ij}=0} \frac{1}{2}(y_{ij} - t_{ij})^2 \right) \tag{1}$$

where N is the number of images, t_{ij} is the ground truth of tags where $y_{ij} = 1$ denotes images i contains tag j, and α is the parameter which balances the positive samples and negative samples. y is the output of the fc7 layer, which is calculated by

$$y_i = \sigma(W x_i^{(7)} + b) \tag{2}$$

where $x_i^{(7)}$ is the input of fc7 layer, W and b are the parameters, and σ is the sigmoid activation function $\sigma(x_i) = \frac{1}{1+e^{-x_i}}$. By replacing $[x_i^{(7)}, 1]$ by new x_i and $[W, b]$ by new W, (2) can be written as

$$y_i = \sigma(W x_i). \tag{3}$$

In the rest of this paper, we use this formulation instead of (2).

In (3), each tag can be regarded as an independent classifier. Thus, for a single tag j, (2) can be written as

$$y_{ij} = \sigma(w_j \cdot x_i) \tag{4}$$

where w_j is the j-th row of the weight matrix W. Then, x_i is the representation of image i and w_j is the representation of tag j.

Thus, the final output y_{ij} is determined by the inner product of the representations of image i and tag j.

B. General Knowledge Graph Embedding

Graph embedding is to represent nodes in a hidden vector space and maintain the graph edges by distance metric in vector space. As general knowledge graphs consists of multiple relation types, and these relation types have heterogeneous characteristics, we cannot use one vector space, as usual, to reflect these multiple relation types. Also, there are both symmetric and asymmetric relations in knowledge graphs. Traditional similarity metric is inadequate to maintain direct relations.

In this paper, we use multiple hidden spaces to maintain multiple types of relations. To address the problem of asymmetric relations, we generate two mappings from original concept representations to the hidden space p: L_p for the left concept, and R_p for the right concept. Then, we calculate the distance of concept w_i and w_j in hidden space p by the Euclidean distance

$$d_{ij}^p = (L_p w_i - R_p w_j)^T (L_p w_i - R_p w_j) \tag{5}$$

where L_p and R_p are $k \times m$ projection matrices (m is the dimension of concept representation and k is the dimension of hidden space). Then, the target of knowledge base embedding learning is to learn mappings L_p and R_p for each relation p, which make d_{ij}^p small if concepts i and j have relation p in knowledge graph, and vise versa. In order to reduce down the dimensionality of the embedding space, we impose orthogonal constraints on L_p and R_p as in [12] which demonstrates that imposing orthogonal constraint on the mapping functions can result in orthogonal embedding representations. In this way, the redundancy of different embedding dimensions can be largely reduced, leading to fewer required embedding dimensions. Therefore, the loss function is defined as

$$\begin{aligned}\mathcal{L}_2 = & -\frac{1}{|E|}\sum_{e_{ij}\in E_p}(d_{ij}^p)^2 + \frac{\gamma}{N^2-|E|}\sum_{e_{ij}\notin E_p}(d_{ij}^p)^2 \\ & + \lambda\sum_p(||L_p^T L_p - I||_F^2 + ||R_p^T R_p - I||_F^2)\end{aligned} \tag{6}$$

where $|E|$ is the number of edges in knowledge graph, N is the number of vertices (tags) in the knowledge graph, $e_{ij} \in E_p$ means the edge from i to j is the p^{th} type of relation.

C. Joint Optimization

In order to jointly optimize image representation learning and knowledge graph embedding, we combine the loss function in Regression CNN (1) and the loss for knowledge base (6) to formulate the final loss function of the proposed R^{3}CNN

$$\mathcal{L} = \mathcal{L}_1 + \beta\mathcal{L}_2 \tag{7}$$

where β is the trade-off parameter to balance image tagging and the constraints of knowledge base embedding. By using the tag representation w_i as bridge for joint optimization, the learnt image representation can capture more precise and complete semantics, and the tag representation can capture not only semantic relations but also visual similarities.

In the loss function 7, there are two groups of variables: 1) The parameters in Regularized Regression CNN (weights and biases of each layer) 2) The projection matrices L_p and R_p. The parameters in Regression CNN are usually optimized using back propagation with stochastic gradient descent (SGD). L_p and R_p can be optimized by gradient descent. Therefore, we can iteratively optimize these two groups of parameters. First, in Regression CNN, the weights W of output layer is special because it occurs in both two terms in the loss function. Its gradient is calculated as

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial w_i} = & \sum_{j=1}^{N} y_{ji}(1-\sigma(w_i\cdot x_j^{(7)})) + \alpha(1-y_{ji})\sigma(w_i\cdot x_j^{(7)}) \\ & - \frac{2\beta}{|E|}\sum_{e_{ij}\in E_p} L_p^T(L_p w_i - R_p w_j) \\ & + \frac{2\beta\gamma}{N^2-|E|}\sum_{e_{ij}\notin E_p} L_p^T(L_p w_i - R_p w_j).\end{aligned} \tag{8}$$

The other parameters in Regression CNN can be easily calculated by back propagation. After the parameters in Regression CNN are converged, we optimize L_p and R_p using gradient descent

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial L_p} = & -\frac{2\beta}{|E|}\sum_p\sum_{e_{ij}\in E_p}(L_p w_i - R_p w_j)w_i^T \\ & + \frac{2\beta\gamma}{N^2-|E|}\sum_p\sum_{e_{ij}\notin E_p}(L_p w_i - R_p w_j)w_i^T \\ & + 4\beta\lambda\sum_p L_p(L_p^T L_p - I)\end{aligned} \tag{9}$$

similarly

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial R_p} = & -\frac{2\beta}{|E|}\sum_p\sum_{e_{ij}\in E_p}(R_p w_j - L_p w_i)w_j^T \\ & + \frac{2\beta\gamma}{N^2-|E|}\sum_p\sum_{e_{ij}\notin E_p}(R_p w_j - L_p w_i)w_j^T \\ & + 4\beta\lambda\sum_p R_p(R_p^T R_p - I).\end{aligned} \tag{10}$$

We alternately optimize all these parameters until they are converged. To be noted that when updating W, the all of the parameters in CNN are also updated through back propagation.

D. Algorithm and Complexity

We summarize the algorithm general knowledge embedded image representation learning as described in Algorithm 1.

The complexity of optimizing L_p and R_p in an iteration in total is $\mathcal{O}(N^2pkm)$, where N is the number of tags, p is the number of relations types, k is the dimension of hidden space and m is the dimension of tag representation. The complexity of optimizing W for each epoch (a whole iteration in CNN) is $\mathcal{O}(n(pm^2+N^2))$, where n is the number of training samples. The complexity of optimizing the whole Regression CNN is difficult to analysis. But in our problem pm^2+N^2 is much

Algorithm 1: General Knowledge Embedded Image Representation Learning

Require: Image-tag data with visual contents; Knowledge base data in triplet form.
Ensure: The R^3CNN model that includes tag presentations and can be used to extract image representations.
1: Initialize the weights in R^{3}CNN using the trained AlexNet model in ImageNet.
2: Fine tune R^{3}CNN using image-tag data, obtain the weights of the last layer W.
3: Initial L_p and R_p with random value.
4: **repeat**
5: Use gradient descent to find optimal L_p and R_p. w.r.t loss function (6) and gradients ((9) and (10)).
6: Fine tune the Regularized Regression CNN based on the loss function (7).
7: **until** All variables converge.

smaller than the whole scale of the parameters in CNN. Thus, optimizing the proposed Regularized Regression CNN costs comparable time to AlexNet (less than 2 times in practice). Besides, the time cost in optimizing L_p and R_p is much less than tuning the CNN. Overall, the time cost of our method is linear to the time cost of tuning an AlexNet, which is acceptable for most GPU computing environments.

E. Applications

The learned image representations can be used in the following application scenarios:

1) *Tag prediction:* In our model, we can obtain the representation of each tag and for any image, we can extract its features using the activation of fc7 layer in our R^{3}CNN model. Thus, we can use (4) to predict the probability of a tag belonging to an image.
2) *Zero-shot tag inference:* Tags of images always evolve over time. It is impossible to cover all of the tags in a given learning framework. For new tags, it takes a lot of time to learn new classifiers. It is important to investigate the problem of inferring unseen tags without training for images, that is zero-shot tagging problem. In our method, we can address this problem by exploiting knowledge graph to infer the representation of the new tags if these tags have relations to the trained tags for images, that is. For a new tag w_i, suppose that it has triplet $< i, j, p >$ in the knowledge graph and w_j has been trained in our method. From (6), we know that a converged solution should have $L_p w_i \approx R_p w_j$. Besides, the regularizer requires that: $L_p^T L_p \approx I$. Therefore, we have $w_i \approx L_p^T R_p w_j$. When w_i occurs on the right of a relation, the result is similar. By averaging all the relations, we have

$$w_i \approx \frac{1}{Z_i}\left(\sum_{e_{ij}\in E_p} L_p^T R_p w_j + \sum_{e_{ji}\in E_p} R_p^T L_p w_j\right) \quad (11)$$

where Z_i is the total number of edges that are related to i for normalization. After we calculate w_i by (11), we can infer that whether image i should have tag j by (4). This procedure can be calculated very fast without any training process.
3) *Content based image retrieval:* Although we use the image-tag as supervisory information to learn image representations, we argue that the learnt representations can also be applied into other image applications, such as typical content based image retrieval. We can extract image features using the proposed R^{3}CNN model. Due to the fact that the higher layers are more close to semantics, we use the activation of fc6 or fc7 layer as image features. Then, nearest neighbor based methods can be used for content based image retrieval task.

V. Experiments

A. Experimental Settings

We evaluate our approach in three applications, including tag prediction, zero-shot tag inference and CBIR. Due to the fact that our approach is a general representation learning method to capture image semantics, we select the state-of-the-art representation learning approaches with or without extra semantic resources as baselines.

1) Datasets: There are two datasets in our experiments. One is NUSWIDE [28] for image tagging and the other is ConceptNet [3] for knowledge base embedding.

NUSWIDE contains 269,648 images with 425,001 tags, which are collected from Flickr. ConceptNet is a general knowledge base with multiple languages. In our experiments, we only select concepts from ConceptNet5[3] in English language, which has millions of concepts and 56 types of relations. In NUSWIDE and ConceptNet, there are 92,595 common tags (including words and phrases). To bridge them together, we sampled 1,000 overlapped tags with the highest frequency. And then, we select the images in NUSWIDE, which has at least three selected tags. After filtering the images that cannot be crawled from Flickr at present, there are 86,035 images left. Referring to [19], we first filter some relations that are intuitively not recognizable in images (e.g. *Causedby, HasSubevent*) and some negative relations (e.g. *NotIsA, Antonym*). Finally, there are 15 types of relations selected, including *IsA, HasA, RelatedTo, UsedFor, AtLocation, DefinedAs, InstanceOf, PartOf, HasProperty, CapableOf, SymbolOf, LocatedNear, ReceivesAction, MadeOf and Synonym*. Compared to [19], we add *Synonym* because we found it is positive in our experiments. There are 27,440 edges in these 15 types among the 1000 selected concepts. In summary, the scale of our dataset is shown in Table I. In our experiments, we randomly sampled 50,000 images for training and rest 36,035 images for testing.

2) Model Implementation: We train the R^{3}CNN using the training dataset following algorithm in Section IV-D. In training data, each image is resized to 224×224 with horizontal flipping. Following AlexNet [4], dropout with probability 0.5 is used in fc6 and fc7 layers in training process. When fine-tuning

[3][Online]. Available: http://conceptnet5.media.mit.edu

TABLE I
SCALE OF THE DATASET

	tags	images	relation types	relations
number	1000	86,035	15	27,440

the R^{3}CNN, we fix the weights of first 3 convolutional layers by setting the learning rate to be 0; set learning rate of the last regression layer to be 0.1; and the learning rate of other layers to be 0.01. The learning rates are decreased by 0.1 times after 10,000 iterations. When fine-tuning the Regularized Regression CNN with knowledge base embeddings, we also fix the first 3 convolutional layers and set the other learning rates to be 0.01. Our model is implemented by matlab toolbox MatConvnet [29]. It is trained on a single GeForce Tesla K40 GPU with 12 GB memory.

In the loss function (7), there are 5 parameters: α, β, γ, λ, and the dimension of projection hidden space k. In our experiments, we tune these parameters alternatively by grid search. We get an acceptable performance when $\gamma = 0.5$, $\lambda = 0.1$, $k = 100$ and $\beta = 0.02$. When k is larger than 100, the performance improves a little, but it costs more time.

B. Tag Prediction

We use image-tag associations in NUSWIDE dataset as ground truth and use $50,000/36,035$ images for training/testing. After training, we use the probability produced by (4) to rank the 1,000 tags for a given image. We use Precision and MAP (Mean Average Precision) to evaluate the ranking performance.

We use the following image tagging methods as baselines:

1) *Neighbor voting (NV) [18]:* It is a nearest neighbor based method, which evaluates tag relevance for an image w.r.t the tags of its nearest neighbors. In our experiments, we use: 1) Bag-of-Words feature provided by NUSWIDE (denoted as NV-BoW) and 2) the CNN feature extracted by AlexNet fc7 layer for nearest neighbor selection (denoted as NV-CNN). We use 100 as the number of nearest neighbors.
2) *SVM:* We use SVM as the classifiers of the tags based on CNN features from AlexNet.
3) *Regression CNN (RegCNN):* We use the Regression CNN introduced in Section IV without knowledge base embedding. It can be regarded as a fine-tuned deep model for image tagging.
4) *Linked tag (LinkTag) [21]:* This method uses DBPedia to extend the tags. In our experiments, we use ConceptNet instead.
5) *DeViSE [1]:* DeViSE first learns words embeddings and then learns a CNN model which can transfer the visual representations to word embeddings. This method is designed for ImageNet classification. When implementing the algorithm, in order to make the learned representations fitting for our data, we use the tags in NUSWIDE and ConceptNet instead. As in [1], we use the ConceptNet as an indirected similarity graph.

The comparison with these baselines can demonstrate the advantages of our methods in different aspects. The first three methods do not use external semantic resources: NV is knn-based, SVM is a shallow-structured classification model and RegCNN is a deep-structured classification model. The rest two methods exploit knowledge base, LinkTag fuse tags and knowledge base in the late stage rather than representation stage. DeViSE is a deep model based method with semantic ontology embedding for classification. However, it learns word embedding and CNN independently without joint optimization, and it is based on semantic ontology (WordNet) rather than multiple kinds of relational knowledge.

The experimental results are shown in Table II. We have the following observations:

1) Our method performs significantly better than baseline methods in all cases. Especially, it has 26% improvement in MAP.
2) The improvement of R^{3}CNN over DeVISE demonstrates that directly mapping image space to general knowledge space is inadequate, due to the inconsistency of these two spaces. The relation regularization imposed to deep model can help to solve this problem. Also, exploiting different kinds of relations rather than treating them uniformly can bring much more additional useful information to image space.
3) LinkTag performs much worse than DeVISE, indicating that the late fusion of image labels and knowledge space will bring many erroneous labels that are not related with the image content. Embedding the knowledge into image representation stage is an effective way.
4) The RegCNN performs better than SVM, NV-CNN and NV-BoW, demonstrating the advantages of end-to-end representation learning over hand-crafted features and the AlexNet representations cannot work well in wild images.

C. Zero-Shot Tag Inference

We randomly sampled 100 unseen tags with the requirement that each one should occur in both NUSWIDE and ConceptNet and have at least 5 relations with the seen tags. For each unseen tag, we use (4) to evaluate its relevance to the testing image, and give it a ranking score. The images that are eventually labeled by the unseen tag are ground truth. Then, we use MAP to evaluate the performances. Zero-shot problem cannot be solved by classification basedl methods because there is no training process for the unseen tags. Therefore, only LinkTag and DeViSE can be used as baselines.

We randomly select 20 unseen tags and report the performances of different methods on these selected tags in Fig. 3. From left to right, the unseen tags are ordered by the number of relations do they have with the trained tags. We also report the average performances for the 100 unseen tags at the last. From the results, we have the following observations.

1) Our method consistently and significantly outperform the baselines in all these unseen tags. On average, our method has more than 100% relative improvement over other baselines in MAP.
2) The more relations do the unseen tag has with the trained tags in ConceptNet, the higher MAP and relative

TABLE II
PERFORMANCE OF TAG PREDICTION TASK ON NUSWIDE DATASET

	NV-BoW	NV-CNN	SVM	RegCNN	LinkTag	DeViSE	R^3CNN
MAP	0.0651	0.1627	0.1726	0.1890	0.0932	0.1863	**0.2388**
P@5	0.1030	0.3000	0.3182	0.3854	0.1891	0.4835	**0.5587**
P@10	0.0890	0.2758	0.3163	0.2970	0.1456	0.3001	**0.3951**
P@50	0.0704	0.1206	0.1805	0.2071	0.0610	0.1900	**0.2109**

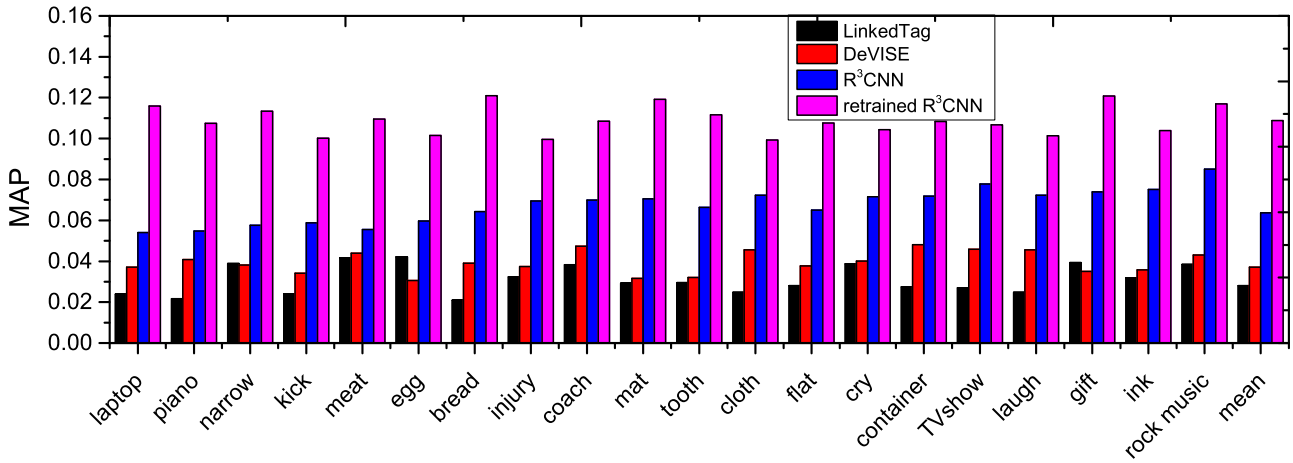

Fig. 3. MAP of zero-shot tag inference task, including 20 example tags and the average of 100 unseen tags.

improvement can be achieved by our method. This demonstrates the values of general knowledge in benefiting image representation learning.

3) In addition, we retrain the R^3CNN model with the data including unseen tag information. Then we use the retrained R^3CNN to predict the unseen tags for images, and report the prediction performance in the last column in each group. Here we use this performance as the higher bound of unseen tag inference. We can see that, in average, the performance in inferring unseen tags by our method can get about 70% of the performance in predicting trained tags. Also, the more relations an unseen tag has with trained tags in ConceptNet, the smaller the margin becomes between unseen tag inference and trained tag prediction in our method. This once again demonstrates the importance of introducing general knowledge graph into image representation learning.

D. Content-Based Image Retrieval

In order to demonstrate that the learned representations can be widely used in multiple application scenarios with good generalization ability, we evaluate the proposed approach in content-based image retrieval task.

Holidays [32] is a commonly used dataset in content based image retrieval tasks, which contains 1491 high resolution personal holiday photos. 500 images are used as queries and the remaining 991 images are labeled as ground truth. For each query image, we use Euclidean distance as the metric to rank the testing images. Then, we evaluate the ranking results by MAP. There are four baselines in this experiment, including two typical representations, including BoW (Bag-of-Words based on SIFT discriptors) and AlexNet (CNN feature extracted from AlexNet), RegCNN in Section IV-A for multi-label image representation learning (note that RegCNN is eventually a fine-tuned AlexNet by NUS-WIDE in a multi-label setting), and DeViSE (the representations derived from DeVISE). In our method, we have two variants: R^3CNN-fc6 and R^3CNN-fc7. R^3CNN-fc6 corresponds to fc6 layer (the layer before the last fully connected layer) and R^3CNN-fc7 corresponds to the last fully connected fc7 layer.

TABLE III
PERFORMANCE OF CBIR ON HOLIDAYS

	BoW	AlexNet	RegCNN	DeViSE	R^3CNN-fc6	R^3CNN-fc7
MAP	0.540 [30]	0.642 [31]	0.6709	0.6876	0.7254	0.7445

The experimental results are shown in Table III. The results of BoW and AlexNet are respectively reported by [30] and [31]. We can see that both of the R^3CNN-fc6 and R^3CNN-fc7 representations perform better than baseline representations. This demonstrates that the representation learned from our model has good generalization ability in different datasets and different application scenarios. The result that R^3CNN-fc7 performs better than R^3CNN-fc6 indicates that the higher-level representations perform better when transferring to a different dataset. A plausible reason is that the higher-level representations are closer to semantics and thus have better transferability across these two datasets.

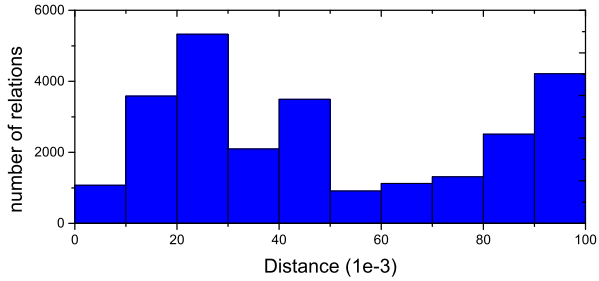

Fig. 4. Histogram of the relations on distance regions.

Useless relations:			Omissive relations:		
Concept1	Concept2	Relation	Concept1	Concept2	Relation
holiday	book	InstanceOf	love	action	RelatedTo
nature	artist	IsA	cat	dog	isA
country	field	RelatedTo	Paris	capital	AtLocation
stop	fast	RelatedTo	airplane	large	HasProperty
space	cold	HasProperty	police	tourist	RelatedTo
sheep	female	RelatedTo	apple	phone	isA
fence	metal	MadeOf	job	work	InstanceOf
cat	feline	isA	sunset	sky	RelatedTo
cemetery	house	part of	park	warm	HasProperty

Fig. 5. Showcase of "useless relations" and "omissive relations".

E. Insight Analysis

Till now, we have demonstrated the advantages and effectiveness of embedding general relational knowledge into image representation learning. In this section, we will further analyze how the image representation learning and knowledge graph embedding reenforce each other, and what kinds of relational knowledge can benefit the image representation learning.

In our method, the image representation learningand knowledge graph embedding are jointly optimized. Thus for any pair of concepts i and j, there are two sources of constraints on the similarity of their representation. If they are related in ConceptNet with relation type p, they should share similar representation in space p through knowledge graph embedding. However, if the images with tag i and images with tag j are visually different, their representations should be quite different in image representation learning. After the model converges, the distance of the output representations of w_i and w_j in $p-th$ space d_{ij}^p can be calculated by equation 5, and the distance can tell us: 1) If concept i and j are related in ConceptNet with relation type p, but d_{ij}^p is relatively large, that means the knowledge $< i, j, p >$ cannot be reflected or supported in image space. 2) If i and j are not related in ConceptNet with relation type p, but d_{ij}^p is relatively small, that means the knowledge $< i, j, p >$ might be an image-specific knowledge and can possibly supplement ConceptNet. Thus, we temporarily define that:

1) if the distance $d_{ij}^p > \delta$, and relation $\langle c_i, c_j, r_p \rangle$ exists in ConceptNet, this is a "useless" relation in image representation learning;
2) if the distance $d_{ij}^p < \epsilon$, and relation $\langle c_i, c_j, r_p \rangle$ does not exist in ConceptNet, this is an "omissive" relation in the knowledge base.

Here δ and ϵ are two parameters, which are determined according to the distribution of the distances.

For the 27,440 relations in our training data, we calculate their histogram on distance regions, which is shown in Fig. 4. According to the distribution of the distances, we define $\delta = 0.08$ and $\epsilon = 0.03$. Then, we select some representative "useless" relations and "omissive" relations as showcase in Fig. 5. We can observe that useless relations are usually caused by: 1) word ambiguity, such as "holiday is an instance of book"; 2) abstract or invisible concepts, such as "stop is related to fast"; 3) true but rarely-used tags, such as "cat is a feline". For omissive relations, we just select the space which makes them the nearest. From the results, we can observe that, except several failed cases, most of the pairs of tags in omissive relations are truly related, and there are many reasonable relations that can supplement ConceptNet.

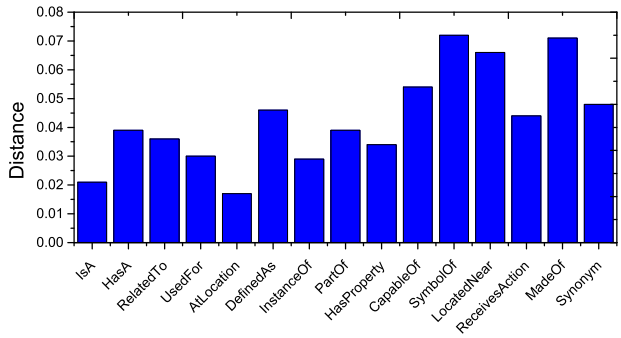

Fig. 6. Average distance for each relation type.

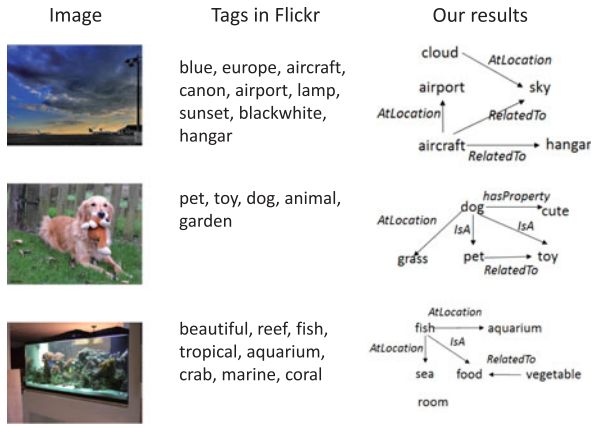

Fig. 7. Showcase of structured image tagging results. The first two rows are successful cases and the last row is a failure case.

In addition, we also calculate the average distance of each relation type (as shown in Fig. 6 to observe which type is the most helpful in image representation learning. From the results, we can observe that "AtLocation", "isA" and "InstanceOf" are the most important relation types in image representation learning, which is quite reasonable. In contrast, "MadeOf" and "SymbolOf" are the most useless relation types. The main reason is that such relations can hardly be reflected in image space. For example, "paper is made of wood" is a common knowledge. However, we rarely tag "wood" on a picture of papers.

By properly embedding the general relational knowledge into image representation learning, the learnt representation is endowed with the ability of reflecting the structured semantics. We further explore this merit and evaluate whether the learnt representation can generate structured tags for images. We first predict tags for an image as in section 5.2. Then, for the top tags, we use the previous definition: if $d_{ij}^p < 0.03$, we think tag i and tag j have relation p. Fig. 7 is a showcase of our results. The first two rows are successful cases and the last row is a failed

case. From these examples, we can see the contrast between traditional tags and the structured tags generated by our method. It is obvious that the structured tags with relational knowledge can significantly help to understand the image content. Also, it can well support reasoning on the images, which is crucial in various image applications.

VI. Conclusion

In this paper, we propose a novel approach, which embeds the relations in general knowledge graph into multi-label image representation learning task to make the learnt representation involving more precise and complete semantics of images. A Relational Regularized Regression CNN (R^3CNN) model is designed to jointly optimize the image representation learning problem and knowledge graph embedding problem. In image representation learning problem, a Regression CNN model is adopted to learn image representation from multi-label image-tag data. In knowledge graph embedding task, we map a concept in knowledge graph into multiple hidden spaces and consider the asymmetric relations. The experimental results in tag prediction and zero-shot tag inference tasks demonstrate that our approach can better embed the general knowledge in representation learning. Experiments in CBIR show that the learnt representation is transferable to other semantic related tasks and performs better than traditional deep features. The observations of the learnt tag relations demonstrate the potential of our approach in building vision-based concept relationships.

In this work, we just give some simple observations of the learnt relationships. In the future, we can go further in building vision-based concept relations based on the user tagging behavior and discover the difference between language-level knowledge and vision-level knowledge.

Acknowledgment

The authors would like to thank the research fund of Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology and the Young Elite Scientist Sponsorship Program by CAST.

References

[1] A. Frome *et al.*, "Devise: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2121–2129.

[2] X. Li, S. Liao, W. Lan, X. Du, and G. Yang, "Zero-shot image tagging by hierarchical semantic embedding," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2015, pp. 879–882.

[3] R. Speer and C. Havasi, "Conceptnet 5: A large semantic network for relational knowledge," in *Proc. Peoples Web Meets NLP*, 2013, pp. 161–176.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR,* 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residuallearning for image reognition," *CoRR,* 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[7] H. Izadinia, B. C. Russell, A. Farhadi, M. D. Hoffman, and A. Hertzmann, "Deep classifiers from image tags in the wild," in *Proc. ACM Workshop Community-Organized Multimodal Mining: Opportunities Novel Solutions*, 2015, pp. 13–18.

[8] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, "Deep convolutional ranking for multilabel image annotation," *CoRR,* 2013. [Online]. Available: http://arxiv.org/abs/1312.4894.

[9] F. Zhao, Y. Huang, L. Wang, and T. Tan, "Deep semantic ranking based hashing for multi-label image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 1556–1564.

[10] A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia, "Multi-task CNN model for attribute prediction," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1949–1959, Nov. 2015.

[11] H. Zhang, X. Shang, H. Luan, Y. Yang, and T.-S. Chua, "Learning features from large-scale, noisy and social image-tag collection," in *Proc. ACM Int. Conf. Multimedia.*, 2015, pp. 1079–1082.

[12] D. Wang, P. Cui, M. Ou, and W. Zhu, "Deep multimodal hashing with orthogonal regularization," in *Proc. 24th Int. Conf. Artif. Intell*., 2015, pp. 2291–2297.

[13] D. Wang, P. Cui, M. Ou, and W. Zhu, "Learning compact hash codes for multimodal representations using orthogonal deep structure," *IEEE Trans. Multimedia*, vol. 17, no. 9, pp. 1404–1416, Sep. 2015.

[14] X. Yang *et al.*, "Deep relative attributes," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1832–1842, Sep. 2016.

[15] S. Liu, P. Cui, W. Zhu, and S. Yang, "Learning socially embedded visual representation from scratch," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 109–118. [Online]. Available: http://doi.acm.org/10.1145/2733373.2806247

[16] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 819–826.

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

[18] X. Li, C. G. Snoek, and M. Worring, "Learning social tag relevance by neighbor voting," *IEEE Trans. Multimedia*, vol. 11, no. 7, pp. 1310–1322, Nov. 2009.

[19] L. Xie and X. He, "Picture tags and world knowledge: Learning tag relations from visual semantic sources," in *Proc. 21st ACM Int. Conf. Multimedia.*, 2013, pp. 967–976.

[20] S. Auer *et al.*, *Dbpedia: A Nucleus for a Web of Open Data*. New York, NY, USA: Springer, 2007.

[21] D.-H. Im and G.-D. Park, "Linked tag: Image annotation using semantic relationships between image tags," *Multimedia Tools Appl.*, vol. 74, no. 7, pp. 2273–2287, 2015.

[22] X. Zhang *et al.*, "Enhancing video event recognition using automatically constructed semantic-visual knowledge base," *IEEE Trans. Multimedia*, vol. 17, no. 9, pp. 1562–1575, Sep. 2015.

[23] W. Lu *et al.*, "Web multimedia object classification using cross-domain correlation knowledge," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1920–1929, Dec. 2013.

[24] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2787–2795.

[25] K. Guu, J. Miller, and P. Liang, "Traversing knowledge graphs in vector space," *CoRR,* 2015. [Online]. Available: http://arxiv.org/abs/1506.01094.

[26] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual translation embedding network for visual relation detection," *CoRR*, 2017. [Online]. Available: http://arxiv.org/abs/1702.08319

[27] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Proc. Neural Networks: Tricks Trade*, 2012, pp. 9–48.

[28] T.-S. Chua *et al.*, "Nus-wide: A real-world web image database from National University of Singapore," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2009, Art. no. 48.

[29] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for MATLAB," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 689–692.

[30] R. Arandjelovic and A. Zisserman, "All about VLAD," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 1578–1585.

[31] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2014, pp. 806–813.

[32] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *Proc. 10th Eur. Conf. Comput. Vis*. 2008, pp. 304–317.

Authors' photographs and biographies not available at the time of publication.