Multi-Modal Knowledge Representation Learning via Webly-Supervised Relationships Mining

Fudong Nian[1,2], Bing-Kun Bao[2,3,4], Teng Li[1], Changsheng Xu[2,3]
[1]Anhui University, Hefei, China
[2]National Lab of Pattern Recognition, Institute of Automation, CAS, Beijing 100190, China
[3]University of Chinese Academy of Sciences
[4]State Key Laboratory for Novel Software Technology, Nanjing University, P.R. China
nianfudong@gmail.com,bingkun.bao@ia.ac.cn,tenglwy@gmail.com,csxu@nlpr.ia.ac.cn

ABSTRACT

Knowledge representation learning (KRL) encodes enormous structured information with entities and relations into a continuous low-dimensional semantic space. Most conventional methods solely focus on learning knowledge representation from single modality, yet neglect the complementary information from others. The more and more rich available multi-modal data on Internet also drive us to explore a novel approach for KRL in multi-modal way, and overcome the limitations of previous single-modal based methods. This paper proposes a novel multi-modal knowledge representation learning (MM-KRL) framework which attempts to handle knowledge from both textual and visual modal web data. It consists of two stages, i.e., webly-supervised multi-modal relationship mining, and bi-enhanced cross-modal knowledge representation learning. Compared with existing knowledge representation methods, our framework has several advantages: (1) It can effectively mine multi-modal knowledge with structured textual and visual relationships from web automatically. (2) It is able to learn a common knowledge space which is independent to both task and modality by the proposed Bi-enhanced Cross-modal Deep Neural Network (BC-DNN). (3) It has the ability to represent unseen multi-modal relationships by transferring the learned knowledge with isolated seen entities and relations into unseen relationships. We build a large-scale multi-modal relationship dataset (MMR-D) and the experimental results show that our framework achieves excellent performance in zero-shot multi-modal retrieval and visual relationship recognition.

KEYWORDS

Webly-supervised, relationship mining, multi-modal, knowledge representation learning

1 INTRODUCTION

Knowledge representation learning (KRL), which originates from structured text representation, has been successfully utilized in various fields such as knowledge graph construction [22] and knowledge inference [39]. Typical knowledge graph usually provides a huge amount of structured information with triple facts represented as *(head entity, relation, tail entity)*, which are also abridged as *(h,r,t)*. To model them, most previous textual KRL methods [4] [18] [34] focus on textual modality and utilize translation-based technique which projects both entities and relations into a continuous low-dimensional semantic space, with relations considered as translating operations between head and tail entities. Inspired by textual KRL, visual KRL methods presented in [19] [27] [38] [42] focus on analyzing each image by detecting two objects as well as their intersection and encoding them into a semantic space. The objects and their intersection are corresponding to entities and relation in textual KRL respectively. However, all the existing visual KRL methods can only handle the predefined visual relationships and the data need to be annotated with bounding boxes. This puts a very high requirement on the data acquisition, and also restricts their applications, e.g. zero-shot multi-modal retrieval.

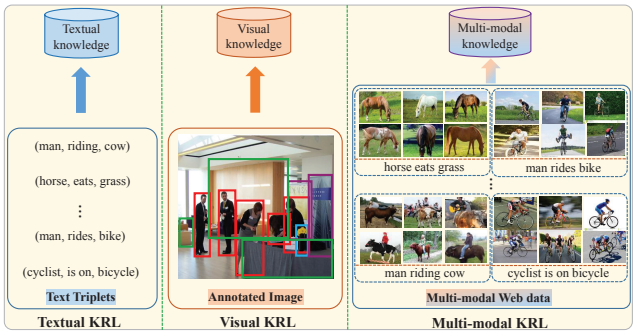


Figure 1: Illustration of the difference of conventional textual KRL, visual KRL and the Multi-modal KRL (MM-KRL).

We extend single modal KRL into multi-modal KRL (MM-KRL), in which each relationship has two modalities, that is texts and images. Multi-modal KRL is firstly proposed by Xie *et al.* [37]. As a typical multi-modal KRL work, the textual and visual modalities we consider are usually globally labeled with only one relationship for less manual labor and wider applications. Figure 1 illustrates the difference of conventional textual KRL, visual KRL and MM-KRL. The advantages of MM-KRL are as follows. Firstly, MM-KRL can reveal the *comprehensive semantics* by utilizing the complementary information from multiple modalities. For example, given two textual relationships "man rides bike" and "cyclist is on bicycle", conventional textual KRL method would deem them as not similar. However, their knowledge is extremely similar while we consider their corresponding images. Secondly, MM-KRL is beneficial for many downstream *multi-modal tasks*, such as cross-modal retrieval,

MM'17, October 23–27, 2017, Mountain View, CA, USA.
 ISBN 978-1-4503-4906-2/17/10…$15.00
DOI: http://dx.doi.org/10.1145/3123266.3123443

visual relationship recognition and image/video captioning. With the learned common space and extracted unified representation, the gap between text and image can be bridged easily. Finally, a good MM-KRL could implicitly link visual modality with textual modality on entities and relations respectively, which can be further applied to *zero-shot tasks*. For example, given "man rides cow" and "horse eats grass", the learned knowledge representation can implicitly project textual and visual modalities of "cow" into a unified representation in the common space, and so do "eat" and "grass".

Learning a good knowledge representation on multiple modalities is not a trivial matter and obviously faces several challenges, which lead to few works focus on MM-KRL. Multi-modal training data with structured relationships are hard to obtain. There is no existing large scale multi-modal relationship dataset including textual modality in (h,r,t) form and their corresponding images with globally annotated. Therefore, we have to mine structured relationship sentences and their corresponding images directly from web data. The method design is even more challenging, as the learned knowledge representation needs to be independent to task and modality, and also locally consistent cross the modalities. First of all, different from some task-dependent learning methods, which train CNN and RNN end-to-end respectively for textual relationship or sentence classification [40] [41], we need learn the knowledge only from data themselves instead of any task-driven one. Moreover, we need to seek a common knowledge space which is beyond textual and visual spaces, and extract the representation by taking cross-modal relevance into consideration. Last but not the least, the designed method should have the ability to project image region (e.g. object/intersection) and its corresponding textual word (e.g. entity/relation) into a unified representation especially from the globally annotated images and holistically structured texts, as it is a huge amount of work to annotate objects and their intersection into image regions manually.

This paper proposes a novel MM-KRL framework which attempts to extract knowledge from both textual and visual modalities. To tackle the challenges in data acquisition and method design, the framework consists of webly-supervised multi-modal relationship mining and multi-modal knowledge representation learning, as shown in Figure 2. The first stage, webly-supervised multi-modal relationship mining, aims to mine a large-scale multi-modal relationship dataset from web data. At this stage, we firstly mine textual relationship candidates from a given text corpus and refine them by discarding the illogical and un-visualized ones, then enrich the textual relationships with images through visual representativeness qualification and noisy image filtering. At the second stage, we propose a Bi-enhanced Cross-modal Deep Neural Network (BC-DNN) to collaboratively learn textual and visual knowledge representations iteratively, and finally achieve a unified representation which is independent to both task and modality. The learned knowledge representation is supervised by the constructed knowledge triplets which jointly utilize intra-similarity and inter-difference on single modality as well as cross-modal relevance.

In the experiments, the qualitative and quantitative results on zero-shot multi-modal relationship retrieval clearly demonstrate that our framework can successfully mine the multi-modal relationships from un-annotated web data, and is able to project knowledge vector learned from different modalities into a common space as well as represent unseen multi-modal relationships. Moreover, we demonstrate the superiority of obtained visual relationship representation model from our framework on public visual relationship recognition dataset.

The main contributions of our work are summarized as follows:

- This is a pioneer study on multi-modal knowledge representation learning directly from web data, which is challenging on both data acquisition and method design. The proposed MM-KRL framework has great potential for ongoing multi-modal knowledge graph construction and knowledge driven cross-modal tasks.
- We propose BC-DNN method to project different modalities into a common knowledge vector space for a united knowledge representation. The learned representation is independent to task and modality, and also locally consistent cross the modalities. The code and obtained textual and visual knowledge representation models are released at project page (http://nlpr-web.ia.ac.cn/mmc/homepage/bkbao/publications/MM-KRL.html).
- We construct a large-scale muti-modal relationship library, called MMR-D, by utilizing the proposed webly-supervised multi-modal relationship mining method. The dataset is released at project page for academic use. Experimental results on both public and constructed datasets validate the effectiveness of the proposed MM-KRL framework.

The rest of the paper is organized as follows. In Section 2, the related work is reviewed. Section 3 introduces the details of the proposed approach. Implementation details are described in Section 4. In Section 5, we report and analyze extensive experimental results. Finally, we conclude the paper with future work in Section 6.

2 RELATED WORK

In this section, we briefly review the related work about textual KRL, visual KRL, and multi-modal learning.

2.1 Textual KRL

Textual knowledge representation learning is a traditional topic in information extraction and knowledge construction. Recently, many methods have been proposed based on translation scheme. Bordes *et al.* [4] propose a TranE model which interprets relation as translating operations between head and tail entities. It is straightforward and effective, but it cannot model 1-to-N, N-to-1 and N-to-N relations. To address it, Wang *et al.* [34] introduce a TransH model to translate on relation-specific hyperplanes. Beyond modeling entities and relations into a common space, TransR [18] interprets entities and relations in different semantic spaces, and sets a projection matrix to project entities into relation space. Besides translation model, there are some techniques utilizing deep learning. Zeng *et al.* [40] exploit a convolutional deep neural network to extract lexicon and sentence level features. Zhang *et al.* [41] propose a simple framework based on recurrent neural networks to do full supervised sentence classification. Unlike their methods which only learn knowledge from textual modality, we focus on knowledge representation from both textual and visual modalities, and our

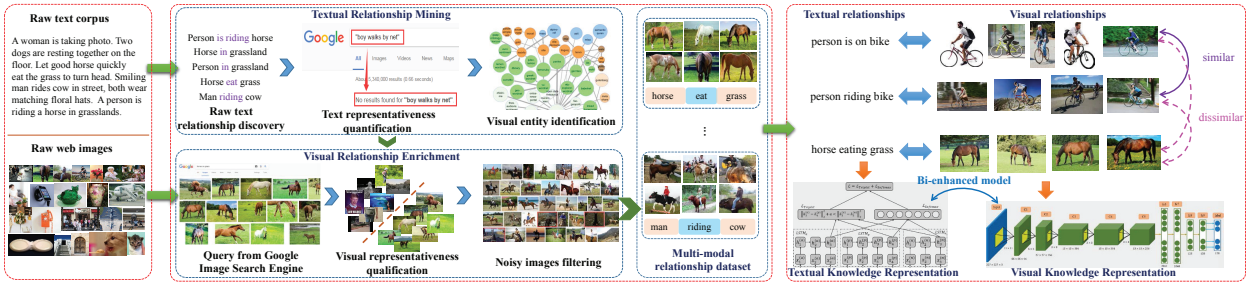


Figure 2: Proposed framework of multi-modal knowledge learning.

data are mined via webly-supervised method rather than annotated text triplets.

2.2 Visual KRL

This section reviews visual KRL for visual relationship, which aims to understand the interaction between two objects in an image. The existing work can be generally categorized into four classes according to interaction types: (1) **Object co-occurrence**. Early work, e.g. [7] [10], leverages object co-occurring statistics to measure whether two objects are appeared in an image simultaneously. (2) **Spatial interactions**. To estimate the relative location between two object classes, Galleguillos *et al.* [12] use a conditional random field (CRF) model and Gould *et al.* [13] propose a two-stage image segmentation approach for incorporating global information into local features. (3) **Human-object interactions**. Ramanathan *et al.* [24] identify interactions between people and objects by a neural network framework. Chao *et al.* [6] propose a Human-Object Region-based Convolutional Neural Networks (HO-RCNN) for human-object interactions detection. (4) **General interactions between objects**. Sadeghi *et al.* [27], Lu *et al.* [19], Li *et al.* [17] and Zhang *et al.* [42] detect and localize pairs of objects in an image, and also classify the predicate/interaction between each pair.

All the existing visual KRL methods only focus on visual modality and the data need to be locally annotated with predefined objects/interactions by bounding boxes. Our work aims at general interaction with the advantages on that 1) we focus on multiple modalities, and 2) the data are from web rather that the pre-defined visual relationships and do not need time-consuming bounding boxes annotation. Specifically, in terms of the interaction type, our work is related to [17] [19] [27] [42]. Unlike our work, [27] only handles the small-scale visual relationship detection, and does not touch multiple modalities and semantic representation learning. [17] [19] [42] improve to study large-scale case, e.g. [42] successfully learns visual knowledge representation by extending textual TransE model [4] to visual TransE. However, the above work limits on the locally annotated data and single modality relationship learning.

2.3 Multi-Modal Learning

Generally, multi-modal learning methods aim to project heterogeneous data into a latent space. The most widely used feature subspace methods on multiple modalities are canonical correlation analysis (CCA) [33], Partial Least Squares (PLS) [8] and Bilinear Model (BLM) [28] [32]. Recently, some methods based on deep learning are also proposed for multi-modal learning tasks. Srivastava *et al.* [29] propose a deep multi-modal restricted boltzmann machines (RBM) to model the cross-modal consistency. Feng *et al.* [11] propose a correspondence autoencoder via constructing correlations between hidden representations of two uni-modal deep autoencoders. In [20] and [35], the deep matching-based methods are used to do multi-modal learning. However, all above methods depend on fully supervised information including annotations on both modalities and the within-modality similarities, and do not emphasize on relationship representation learning.

Multi-modal knowledge representation learning is firstly proposed by Xie *et al.* [37]. In this type of work, e.g. [23] [36], text and image are consistent based on isolate text/visual concept (from WordNet or visual tags), which neglect the structured textual/visual information. Different from these researches, we aim to learn knowledge from structured textual/visual relationship data, which maintain the consistent of data structure (text triplets) used in typical knowledge graphs.

3 OUR APPROACH

Our proposed MM-KRL framework is illustrated in Figure 2, which contains two key stages: webly-supervised multi-modal relationship mining and bi-enhanced cross-modal knowledge representation. In this section, we firstly define the problem and then present the details of the proposed framework.

3.1 Problem Definition

Knowledge exists in the form of relationship implicitly. We define a multi-modal relationship sample R_i (i is the index) with two modalities: (T_i, V_i). Here T_i is the i-th structured textual relationship, which is regarded as a holistic one and does not split into entities and relation in our work. $V_i = \{v_{ik} : k = 1 : N_i\}$ is the i-th visual relationship set including corresponding N_i images. Our purpose is to mine R_i from web data and learn multi-modal relationship projection $f(R_i)$ for knowledge representation, which consists of textual relationship projection $h(T_i)$ and visual relationship projection $g(V_i)$.

3.2 Webly-supervised Multi-modal Relationship Mining

As there are no available public datasets with multi-modal relationships, we need to mine them from web data, which are easy to obtain but contain huge noises in both textual and visual modalities. Considering that mining textual relationships is easier than mining visual ones, we attempt to get textual relationships ready first, then enrich them with corresponding images.

3.2.1 Textual Relationship Mining. Given an arbitrary web text corpus, we firstly utilize Stanford OpenIE toolkit [1] to extract

candidate textual relationships. The resultant textual relationships are then normalized into the lower case and constituted into the raw textual relationship set. Furthermore, we delete the numeral words (e.g. *two, three, many*) and replace the personal pronouns with its general gender appellation (e.g. *he* → *man, she* → *woman*). Considering that the textual relationship should be short in length as it is structured with triple facts represented as *(h,r,t)*, we remove those with more than eight words (e.g. *"Major League Baseball game with player from Pittsburgh Pirates crossing"*).

However, due to shortcomings of the OpenIE algorithm, the raw textual relationship set contains amount of illogical relationships and non-visualized entities. For example, *"altitude house looking over herd"* is an illogical textual relationship. Another example is, there is no corresponding images to describe *"Europe"*. Due to these observations, two steps, text representativeness quantification and visual entity identification, are employed to construct the final textual relationship set. In text representativeness quantification, we assume that a representative textual relationship should be used commonly, that is, we can find it from web. Hence, we utilize Google Text Search Engine to filter non-representative relationships which do not receive any returned web links. In visual entity identification, we utilize WordNet [21], which is a lexical database for the English language that groups English words into sets of synonyms. It is congenial with reason and common sense that word belonging to *"animal", "person", "plant", "artifact", "natural object", "substance", "body", "food", "group"* can be well visualized by image.

3.2.2 Visual Relationship Enrichment. To enrich the textual relationships with visual modality, we retrieve the top returned images from Google Image Search Engine individually. Figure 3(a) is the desired returned images which can constitute into a good relationship dataset. By observing the returned images, we find that there are two more tasks that need to be done to obtain a good dataset. One is to remove the returned images and their descriptions which have no common semantics, as shown in Figure 3(b). The other is to remove noisy returned images which are irrelevant to the corresponding relationship, like the images in red boxes in Figure 3(c).

Visual Representativeness Qualification: Inspired by tag representativeness [30], we define visual representativeness of relationship to remove the non-visualized relationships with the following assumptions:

1. If a relationship is visual representativeness, then most of its retrieved images are semantically related to the corresponding textual relationship. In other words, most images should be semantically similar to each other.

2. Distance derived from image representations reflects visual semantic similarity. That is, distance between images sharing similar visual semantic content is smaller than distance between images not sharing similar visual semantic content.

The representativeness of visual relationship is computed as in Eq. (1).

$$\Phi(V_i) = \frac{1}{N_i}\sum_{k=1}^{N_i} dist(v_{ik}, Cent(V_i)) \tag{1}$$

where, V_i is the image set corresponding to the i-th relationship, N_i is the number of these images and $Cent(V_i)$ is the centroid of V_i. We remove relationships whose $\Phi()$ are greater than a threshold.

(a) (b) (c)

Figure 3: Illustration of three types of images returned by searching textual relationships with Google Image.

Distance function $dist()$ can be computed using deep semantic features. For pretrained CNN model, we fine-tune the ResNet-152 [14] parameters of the released CNN model on ImageNet with the images from Visual Genome [19]. The reason of introducing Visual Genome for fine-tuning is that images of our dataset are multi-object and some objects are not listed in ImageNet. Considering that images in Visual Genome are multi-object and the class number is large, we utilize cross entropy as the loss function during fine-tuning.

Noisy Images Filtering: We follow the observation that if the relationship's visual modality is representativeness, most images are desired and only a few are noisy. As there are at least two visual objects contained in each image, the conventional image filtering method, exemplar-LDA based concept detectors [7], is not suitable for our case. Thus, we filter noisy image by its noisy score, which is calculated by summarizing the distances of this image with all others, as shown in Eq. (2).

$$S_{ik}^{noisy} = \sum_{j=1, j\neq k}^{N_i} dist(v_{ik}, v_{ij}) \tag{2}$$

where, S_{ik}^{noisy} is the noisy score of the k-th image in the V_i relationship image set, $dist()$ is the distance of an image pair (v_{ik}, v_{ij}) computed by low level visual features. The image is regarded as noisy image if its S_{ik}^{noisy} is greater than a threshold. Inspired by [26], we use the distance measurement method through Fourier transformation. We denote the feature vector of the k-th image in the V_i relationship image set as $\mathbf{x}_{ik}$, which is obtained using Bag-of-word (BOW) and Spatial Pyramid Matching (SPM). As shown in Eq. (3), we use the 1-dimensional circulate encoding method to achieve the visual distance of an image pair (v_{ik}, v_{ij}).

$$dist(v_{ik}, v_{ij}) = \frac{1}{\mathcal{F}^{-1}(\frac{\mathcal{F}(\mathbf{x}_{ik})^* \odot \mathcal{F}(\mathbf{x}_{ij})}{\mathcal{F}(\mathbf{x}_{ik})^* \odot \mathcal{F}(\mathbf{x}_{ik}) + \lambda})} \tag{3}$$

where $*$ denotes the conjugation, $\odot$ denotes the element-wise multiplication, $\mathcal{F}$ is the 1D discrete fourier transformation, $\mathcal{F}^{-1}$ is its inverse, $\mathbf{x}_{ik}$ and $\mathbf{x}_{ij}$ are the feature vectors of image pair (v_{ik}, v_{ij}), and λ is the regularization coefficient which ensures the stability of the filter.

Note that we use low level features here instead of fine-tuned CNN feature introduced in the previous section with the following two reasons. One is that the computation of measuring distance on Fourier transformation with low level features is much faster than that on L_2 distance with CNN features. The other is that it is more welcomed to use local feature, e.g. SPM, than the global CNN one, as the visual relationship may only occupy a region of an image. The result of noisy image filtering using the proposed method is shown in Figure 3(c), two images with blue boundingbox are regarded as noisy images by our method.

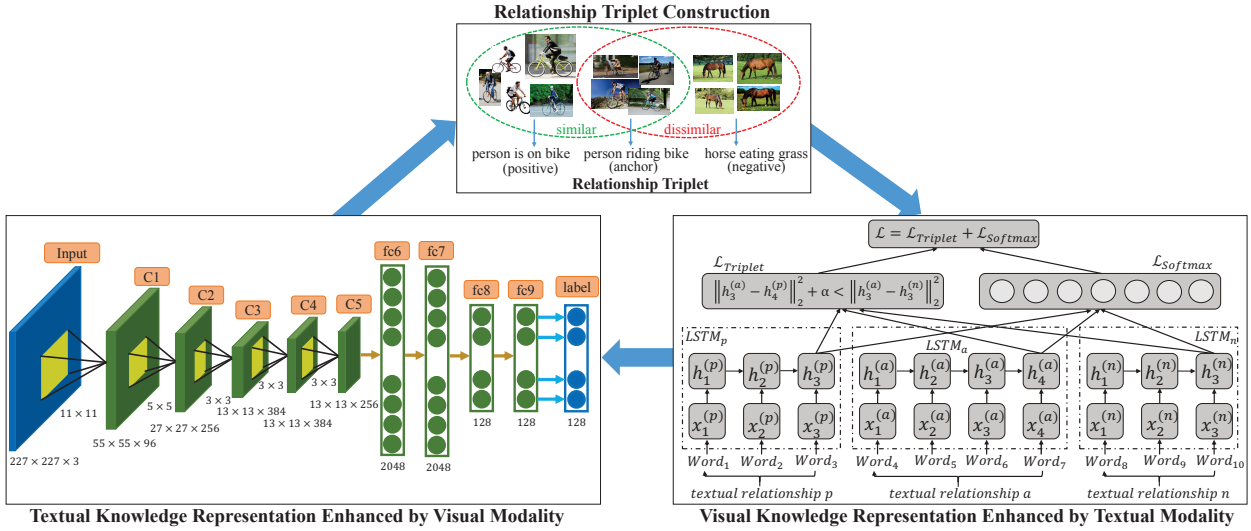


Figure 4: Bi-enhanced cross-modal knowledge representation.

3.3 Bi-enhanced Cross-modal Knowledge Representation

This section aims to fuse textual and visual modalities to enhance a comprehensive knowledge representation. We propose a Bi-enhanced Cross-modal Deep Neural Network (BC-CNN) to learn a common knowledge space beyond task and modality by introducing the supervised information from visual/textual modality to textual/visual network. That is, as shown in Figure 4, the knowledge learned from textual modality is supervised by a relationship triplet measured by visual similarity, while the knowledge learned from visual modality is supervised by the belonging textual relationship.

3.3.1 Textual Knowledge Representation Enhanced by Visual Modality. Different from previous texture knowledge learning work depending on the tasks, we target to construct a general knowledge learning method that adapts to any cross-modal relationship tasks. The supervised information on textual knowledge representation utilizes the similarities between the corresponding visual relationships. Specifically, we employ deep neural network (especially, RNN), and introduce relationship triplet measured by visual similarity as the supervised information.

The overall architecture of the textual network is a LSTM based recurrent neural network which obtains a learned fixed-length knowledge representation vector at the last hidden state. The supervised information is constructed with a set of relationship triplets as follows. Given an **anchor relationship** S_i^a, the rest relationships are regarded as **positive relationship** S_i^p which have similar visual content with S_i^a, and **negative relationship** S_i^n which is not similar to S_i^a in the corresponding visual modality. The learned representation should maintain all the relationship triplets, that is,

$$\|h(S_i^a) - h(S_i^p)\|_2^2 + \alpha < \|h(S_i^a) - h(S_i^n)\|_2^2 \quad (4)$$
$$\forall((S_i^a), (S_i^p), (S_i^n)) \in \Omega$$

where α is a margin that is enforced between positive and negative pairs, $h(S_i)$ is knowledge representation of textual relationship S_i, Ω is the set of all possible triplets in the training set and has cardinality C.

Then the loss is defined as

$$\mathcal{L}_{Triplet_i} = [\|h(S_i^a) - h(S_i^p)\|_2^2 - \|h(S_i^a) - h(S_i^n)\|_2^2 + \alpha]_+ \quad (5)$$

Considering that two visually similar relationships (e.g. "*man riding bike*" and "*man is on bike*") impossibly have the same text appearance, we need to avoid the trivial solution that $h(S_i^a) = h(S_i^p)$. Therefore, we add an additional classification loss to force $h(S_i^a) \neq h(S_i^p) \neq h(S_i^n)$ in Eq. (6):

$$\mathcal{L}_{Softmax_i} = -\log \frac{e^{(W_{ya}^T + b_{ya}) + (W_{yp}^T + b_{yp}) + (W_{yn}^T + b_{yn})}}{(\sum_{c=1}^{C} e^{C_c^T + b_c})^3} \quad (6)$$

where W and b are parameters in softmax layer.

Combining Eq. (5) and Eq. (6), the final optimization problem is given by Eq. (7):

$$\min \sum_{i=1}^{C} (\mathcal{L}_{Triplet_i} + \mathcal{L}_{Softmax_i}) \quad (7)$$

Here, we use Adaptive Moment Estimation (Adam) to optimize Eq. (7).

For generating $(h(S_i^a), h(S_i^p), h(S_i^n))$, we calculate the L_2 distance between centroid of two visual relationship sets to measure the similarity between them. Ideally, we should utilize visual relationship knowledge representation vector as image feature. However, visual relationship knowledge learning network is not trained at present. Therefore, we utilize the fine-tuned CNN features extracted in Section 3.2.2 as visual features, and then update the relationship triplet after visual relationship knowledge network is trained. In addition, note that three LSTM networks shown in Figure 4 are sharing parameters.

3.3.2 Visual Knowledge Representation Enhanced by Textual Modality. This part is to learn knowledge from visual modality and ensure knowledge learning from different modalities in a common space. The supervised information is chosen as textual relationship.

Considering that CNN has been successfully applied to advanced and semantic visual perception tasks, we design a CNN model to learn knowledge from image as shown in Figure 4. It contains five convolution layers and four fully connected layers. The dimension

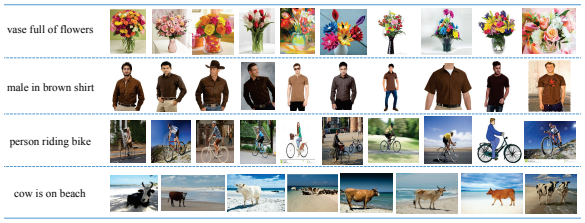


Figure 5: Illustration for a part of the constructed multi-modal relationship dataset (MMR-D).

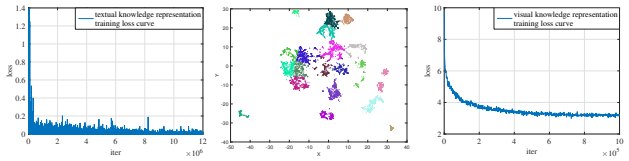
of the last fully connected layer is equal to the dimension of knowledge learning from textual modality and we deem the output of the last fully connected layer as visual knowledge. Hence, the complicated problem of knowledge learning from visual modality is converted to a multivariate regression issue. To optimize the model in Figure 4, we have objection function in Eq. (8):

$$\min_{W,b} \sum_{i=1}^{M} \sum_{k=1}^{N_i} \|g(v_{ik}) - h(T_i)\| + \lambda \|W\|^2 \tag{8}$$

where, M is the number of multi-modal relationships in training set, v_{ik} is the k-th image in relationship i, N_i is the number of image in multi-modal relationship i, $g()$ is image's visual knowledge representation, $h(T_i)$ is the textual knowledge representation learning from the i-th textual relationship, W and b are parameters in proposed CNN model, and $\lambda\|W\|^2$ is regularization term. When Eq. (8) reaches an optima, we obtain the visual knowledge, meanwhile, knowledge representations learning from textual and visual modalities are in a common space. We use Mini-batch stochastic gradient descent (SGD) to optimize Eq. (8).

3.3.3 Multi-modal knowledge representation Optimization. Recalling the triple $((S_i^a), (S_i^p), (S_i^n))$ selected method in Section 3.3.1, the visual features we used are CNN pre-trained on Visual Genome. However, the data content and distribution on Visual Genome are significantly different from our multi-modal relationship dataset. Thus, we need to process the previous two steps with the learned visual knowledge representation. Naturally, we utilize visual knowledge learning from Section 3.3.2 as image feature to measure the similarity between different visual relationships and further select new triple $((S_i^a), (S_i^p), (S_i^n))$ to re-optimize Eq. (7). Then we utilize the new output of our textual knowledge representation learning network to re-optimize Eq. (8). Based on the above analysis, in order to enhance the knowledge learning from textual modality, the learned knowledge from visual modality is utilized to modify triple samples. In order to enhance the visual knowledge learning from visual modality, the new textual knowledge is utilized to re-train our visual knowledge representation learning network.

For unseen visual relationship, as we utilize RNN to learn textual knowledge representation, every word in a textual relationship could be converted to a fixed length vector by the gates of the LSTM unit. Naturally, our model could represent unseen textual relationship while we replace one word (may correspond to different entity or relation) to another. Meanwhile, since our textual and visual knowledge representation models are bi-enhanced, the proposed visual knowledge representation model can also represent unseen visual relationships.



(a) (b) (c)

Figure 6: Visualization of training characteristics in our MM-KRL model.

4 DATASET CONSTRUCTION AND IMPLEMENTATION DETAILS

For raw textual relationship discovery, the text corpus consists of all sentences on MSCOCO dataset. All text preprocessing steps are based on Python API in NLTK [3]. The first 50 images are retrieved from Google Image search engine for each textual relationship query to construct raw image dataset. To obtain deep semantic image feature extractor for quantifying visual representativeness of relationship, we adopt pre-train ResNet-152 model and utilize Caffe for fine-tuning. Examples of the constructed multi-modal relationship dataset are shown in Figure 5, where each multi-modal relationship contains one textual relationship instance and many visual relationship instances. The statistics of the dataset is as follows: 20726 multi-modal relationships (20726 textual relationship instances and 687784 visual relationship instances).

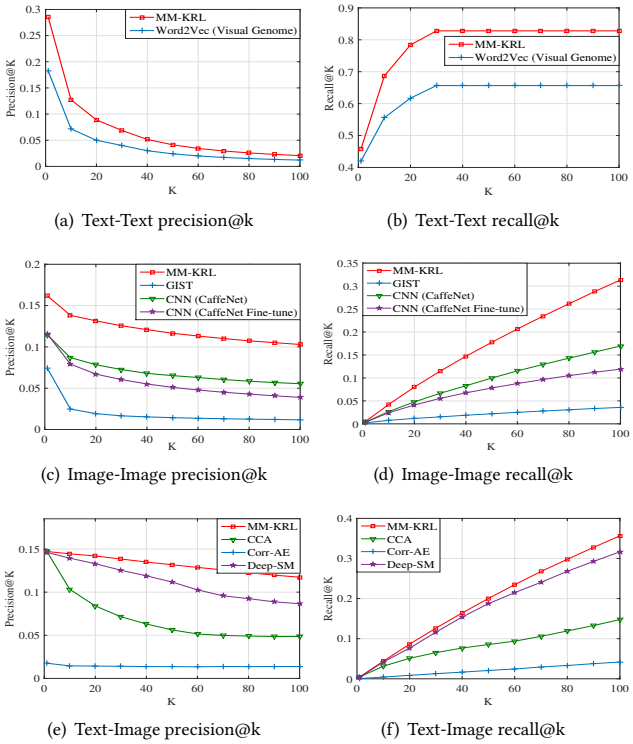
To optimize Eq. (7), we use Theano framework. To optimize Eq.(8), we use Caffe framework. The learning rate is set to $1e-3$ initially and reduced 30% for each 10 epochs, the momentum is set to 0.9, and the weight decay is set to $5e-4$. The dimension d of our multi-modal knowledge representation vector is set to 128. Figure 6(a) and Figure 6(c) show the convergence of our textual and visual knowledge representation model respectively. We also use LargeVis [31] to visualize the textual knowledge representation in training set. As shown in Figure 6(b), the learned textual knowledge representation is discriminative and similar knowledge are clustered together. This phenomenon clearly indicates the effective of our MM-KRL model.

5 EXPERIMENTS

This section demonstrates the effectiveness of the learned multi-modal knowledge. As knowledge is a high abstracted concept, it is difficult to verify directly. Therefore, we employ zero-shot multi-modal retrieval to demonstrate that our method can project different modalities into a common knowledge space, and the learned knowledge can represent unseen relationships. Additionally, we employ a public visual relationship dataset to show the superiority of our method on visual relationship recognition.

5.1 Zero-shot Multi-modal Retrieval

In this section, we employ three zero-shot multi-modal retrieval applications on our multi-modal relationship set, including text-text retrieval, image-image retrieval and text-image retrieval. We adopt 18000 multi-modal relationships (18000 textual relationship instances and 597299 visual relationship instances) as training data to learn multi-modal knowledge representation, and the remaining 2726 textual relationships and their corresponding 90690 visual relationships constitute test set. Noted that all test multi-modal



(a) Text-Text precision@k (b) Text-Text recall@k

(c) Image-Image precision@k (d) Image-Image recall@k

(e) Text-Image precision@k (f) Text-Image recall@k

Figure 7: Comparison of zero-shot multi-modal retrieval in terms of precision and recall.

relationships, including texts and images, are not appeared in training set, it means that we are doing zero-shot multi-modal retrieval experiments.

5.1.1 Text-Text Retrieval. All textual relationship knowledge vectors are extracted from the textual knowledge representation model trained on Section 3.3.1, and the similarity is measured by L_2 distance.

We evaluate the effectiveness of our method by comparing it with Word2Vec based textual relationship representation technique [15], which uses word2vec to convert each word in a textual relationship to a 300-dimensional vector, then removes stop words and averages the remaining words to get a vector representation of the whole textual relationship.

5.1.2 Image-Image Retrieval. An important task in computer vision is image retrieval. An improved retrieval model should be infer to the knowledge expressed by images. Recall that our test set contains 90690 visual relationship instances corresponding to 2726 multi-modal relationship types. For each multi-modal relationship we random select one visual relationship as query and rank the remaining 90689. All visual relationship knowledge vectors are extracted from the visual knowledge representation model trained on Section 3.3.2.

For comparison, we use three image descriptors that are commonly used in image retrieval: 1) *GIST*. The GIST descriptor was initially proposed in [16]. The idea is to develop a low dimensional representation of the scene, which does not require any form of segmentation. We use the Douze's implementation [9] to extract images' GIST descriptors. 2) *CNN (CaffeNet)*. We directly use CaffeNet which pre-trained on ImageNet2012 as image feature extractor that was first utilized in [2]. Same as [2], we deem the output of fc7 layer in CaffeNet as image descriptor. 3) *CNN (CaffeNet Fine-tune)*. [2] also introduces a method to improve the retrieval accuracy by re-training the CNN on the data corresponding to test set. Thus, we fine-tune the CaffeNet on our training dataset. The only difference is that we use relationship classes to replace the original 1000-class softmax layer in CaffeNet and the output of fc7 layer in the fine-tune CaffeNet is deemed as image descriptor. We rank results for a query using the L_2 distance from the query image.

5.1.3 Text-Image Retrieval. Recall that each relationship has two modalities on MMR-D, that is each textual relationship contains many visual relationships representing common knowledge as in training set. Every query uses 1 of these 2726 textual relationships and ranks all 90690 images. The textual and visual descriptors are knowledge extracted from model trained on Section 3.3.1 and Section 3.3.2 respectively.

We evaluate the effectiveness of our method by comparing it with three typical cross-modal retrieval methods: 1) *CCA*. Canonical Correlation Analysis (CCA) is a way of making sense of cross-covariance matrices. Refer to the framework proposed in [25], we use CNN feature (CaffeNet pre-trained on ImageNet2012) and the average word2vec feature [15] as inputs to calculate the projection matrices. In the latent space, we use L_2 distance to measure the similarity between two modalities. 2) *Corr-AE*. Correspondence autoencoder (Corr-AE) is proposed in [11]. Two autoencoders are used to reconstruct the input hand-crafted features. Each auto-encoder in this method is for single modality. The middle layer is used for features in latent space, and the modality similarity is measured by L_2 distance. 3) *Deep-SM*. [35] proposes a simple but effective deep-SM method to address the cross-modal retrieval problem with respect to samples which are annotated with one or multiple labels.

5.1.4 Quantitative Evaluation. To conduct quantitative evaluation of zero-shot multi-modal retrieval, *Recall@k* and *Precision@k* are used as the evaluation metrics which are standard quantitative measures in information retrieval literature. Nine annotators were asked to rank results for each of the queries on all experiments.

Denote $O_k(i)$ as the top k returned results of query i in test multi-modal relationship set, Q as the set of query relationships, and $O_{truth}(i)$ as the ground-truth annotated by annotators. The evaluation metrics are calculated by

$$Precision@k = \frac{1}{|Q|}\sum_{i\in Q}\frac{|O_k(i)\cap O_{truth}(i)|}{k} \tag{9}$$

$$Recall@k = \frac{1}{|Q|}\sum_{i\in Q}\frac{|O_k(i)\cap O_{truth}(i)|}{|O_{truth}(i)|} \tag{10}$$

The results for all above methods are shown in Figure 7. Based on the results we can make the following conclusions: **(1)** The MM-KRL model achieves superior performance on all tasks. **(2)** The superior performance on text-text and image-image retrieval demonstrates the knowledge learned from proposed MM-KRL model is able to represent unseen multi-modal relationships effectively. **(3)** The proposed MM-KRL model achieves better performance than CCA, Corr-AE, Deep-SM on text-image relationship retrieval task, which shows that MM-KRL model can project the knowledge learned from different modalities into a common space.

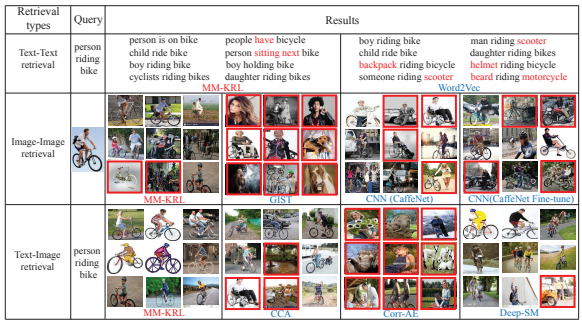


Figure 8: Illustration of sample results of zero-shot multi-modal relationship retrieval (wrong results are annotated in red font or by red bounding boxes).

5.1.5 Qualitative Evaluation. We qualitatively demonstrate the effectiveness of our MM-KRL by illustrating the sample results of zero-shot multi-modal retrieval. Figure 8 shows the zero-shot multi-modal retrieval results of the proposed method and all baselines using three type queries. On text-text retrieval task, when the query textual relationship is "person riding bike", Word2Vec-based approach cannot return "person is on bike" and "boy holding bike" though their knowledge is close to the same. The reason is that Word2Vec only learns from text and does not consider the visual information. On image-image retrieval task, results from baselines only represent visual similar not knowledge similar. We can see that when the query is the image containing a person riding bike, the baselines return many images containing person and bike which may be "person next the bike" or "person repairing the bike" while no "riding" concept is described. On text-image retrieval task, we can find that only the results of the proposed method are compact. Moreover, we observe that GIST and Corr-AE techniques are not suitable for large scale multi-modal retrieval tasks. In summary, from the results, we can see that the multi-modal knowledge representation learned from the proposed framework significantly improves the performance of zero-shot multi-modal relationship retrieval. The major reason is that our knowledge is learned from multi-modal data by jointly considering the intra-similarity, inter-difference and inter-similarity of different modalities.

5.2 Visual Relationship Recognition on Existing Dataset

In this experiment, we aim to evaluate the generalization of the learned knowledge on visual relationship recognition task. Two of the most famous visual relationship datasets are Visual Phrases and Visual Genome. Since we have utilized Visual Genome to fine-tune our image feature extractor in Section 3.2.2, we run additional experiments on the Visual Phrases dataset to recognize visual relationships.

5.2.1 Setup. The Visual Phrase dataset contains 17 phrases. However, several phrases are similar as *"dog jumping"*, we evaluate the learned multi-modal knowledge for visual relationship recognition on 12 of these phrases that can be represented as a *<entity1, relation, entity2>* relationship. Noted that visual relationships in Visual Phrases dataset are annotated by bounding boxes. Thus, in order to do visual recognition, the input of this experiment is the image region containing visual relationship and the output is its relationship type. Specifically, we utilize the proposed knowledge representation model to extract all knowledge vectors from these relationship regions and use multi-class SVM [5] to train the visual relationship recognition model. For comparison, we use Visual Phrases [27] and CaffeNet learning from scratch as two baselines.

Methods	CNN feature [16]	Visual Phrases [27]	MM-KRL
Accuracy	0.324	0.381	**0.423**

Table 1: Accuracy scores for recognizing all relationships on Visual Phrases dataset.

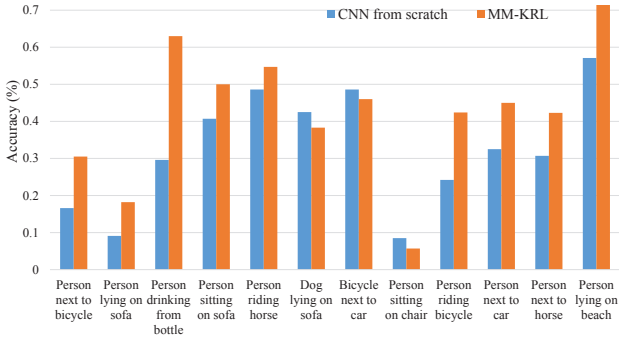


Figure 9: Detailed performance comparison for visual relationship recognition on Visual Phrases dataset.

5.2.2 Results. We use recognition accuracy as evaluation metric and report the experiment results in Table 1 and Figure 9, from which we have the following observations and conclusions: **(1)** Using proposed knowledge as feature descriptor significantly outperforms all baselines. **(2)** Since Visual Phrases [27] employ elaborate handcraft features, we mainly focus on the comparison between our knowledge descriptor and the CNN learned from scratch. Table 1 shows that we get a gain of accuracy about 10%, and the detail performance in Figure 9 shows that our method performs better than CNN learned from scratch on 9 of 12 visual phrases significantly and performs similarly to the baseline on 3 visual phrases. It is clear that the proposed multi-knowledge representation model has well generalization performance on traditional visual relationship recognition task.

6 CONCLUSION

In this paper, we have proposed a novel multi-modal knowledge representation learning model via webly-supervised relationships mining. To achieve this goal, we first do automatic multi-modal relationship mining from web data. Then we propose a systematic solution jointly utilizing intra-similarity, inter-difference and inter-similarity to learn multi-modal knowledge from multi-modal relationships. In the experiments, the qualitative and quantitative results clearly demonstrate that the proposed MM-KRL model is able to represent unseen multi-modal relationships and has well generalization performance. In the future, we will (1) extend the raw relationship extraction step from user annotated sentences to Google Books Corpora to obtain more general and wide relationships, (2) integrate the construction of multi-modal knowledge graph.

7 ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (No. 61432019, No. 61572503, No. 61620106003, No. 61572029), by Key Research Program of Frontier Sciences, CAS (No. QYZDJ-SSW-JSC039) and by the Beijing Natural Science Foundation (No. 4152053).

REFERENCES

[1] Gabor Angeli, Melvin Johnson Premkumar, and Christopher D Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Annual Meeting of the Association for Computational Linguistics*.

[2] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. 2014. Neural codes for image retrieval. In *European Conference on Computer Vision*. 584–599.

[3] Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*. 69–72.

[4] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*. 2787–2795.

[5] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2, 3 (2011), 27.

[6] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. 2017. Learning to Detect Human-Object Interactions. *arXiv preprint arXiv:1702.05448* (2017).

[7] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. 2013. Neil: Extracting visual knowledge from web data. In *IEEE International Conference on Computer Vision*. 1409–1416.

[8] Yongming Chen, Liang Wang, Wei Wang, and Zhang Zhang. 2012. Continuum regression for cross-modal multimedia retrieval. In *IEEE International Conference on Image Processing*. 1949–1952.

[9] Douze. 2003. GIST descriptors. http://people.csail.mit.edu/torralba/code/spatialenvelope/. (2003).

[10] Quan Fang, Changsheng Xu, Jitao Sang, M Shamim Hossain, and Ahmed Ghoneim. 2016. Folksonomy-Based Visual Ontology Construction and Its Applications. *IEEE Transactions on Multimedia* 18, 4 (2016), 702–713.

[11] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. 2014. Cross-modal retrieval with correspondence autoencoder. In *ACM international conference on Multimedia*. 7–16.

[12] Carolina Galleguillos, Andrew Rabinovich, and Serge Belongie. 2008. Object categorization using co-occurrence, location and appearance. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.

[13] Stephen Gould, Jim Rodgers, David Cohen, Gal Elidan, and Daphne Koller. 2008. Multi-class segmentation with relative location prior. *International Journal of Computer Vision* 80, 3 (2008), 300–316.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li Jia Li, and David A. Shamma. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* (2016), 1–42.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[17] Yikang Li, Wanli Ouyang, and Xiaogang Wang. 2017. ViP-CNN: A Visual Phrase Reasoning Convolutional Neural Network for Visual Relationship Detection. *arXiv preprint arXiv:1702.07191* (2017).

[18] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning Entity and Relation Embeddings for Knowledge Graph Completion. In *Association for the Advancement of Artificial Intelligence*. 2181–2187.

[19] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In *European Conference on Computer Vision*. 852–869.

[20] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. 2015. Multimodal convolutional neural networks for matching image and sentence. In *IEEE International Conference on Computer Vision*. 2623–2631.

[21] George Miller and Christiane Fellbaum. 1998. Wordnet: An electronic lexical database. (1998).

[22] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2016. A review of relational machine learning for knowledge graphs. *Proc. IEEE* 104, 1 (2016), 11–33.

[23] Pietro Perona. 2010. Vision of a Visipedia. *Proc. IEEE* 98, 8 (2010), 1526–1534.

[24] Vignesh Ramanathan, Congcong Li, Jia Deng, Wei Han, Zhen Li, Kunlong Gu, Yang Song, Samy Bengio, Chuck Rossenberg, and Li Fei-Fei. 2015. Learning semantic relationships for better action retrieval in images. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1100–1109.

[25] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *ACM international conference on Multimedia*. 251–260.

[26] Jérôme Revaud, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. 2013. Event retrieval in large video collections with circulant temporal encoding. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2459–2466.

[27] Mohammad Amin Sadeghi and Ali Farhadi. 2011. Recognition using visual phrases. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1745–1752.

[28] Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W Jacobs. 2012. Generalized multiview analysis: A discriminative latent space. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2160–2167.

[29] Nitish Srivastava and Ruslan R Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*. 1967 – 2006.

[30] Aixin Sun and Sourav S Bhowmick. 2010. Quantifying tag representativeness of visual content of social images. In *ACM international conference on Multimedia*. 471–480.

[31] Jian Tang, Jingzhou Liu, Ming Zhang, and Qiaozhu Mei. 2016. Visualizing Large-scale and High-dimensional Data. In *International Conference on World Wide Web*. 287–297.

[32] J. B. Tenenbaum and W. T. Freeman. 2000. Separating Style and Content with Bilinear Models. *Neural Computation* 12, 6 (2000), 1247–1283.

[33] Henri Theil and Ching-Fan Chung. 1988. Relations between two sets of variates: The bits of information provided by each variate in each set. *Statistics & probability letters* 6, 3 (1988), 137–139.

[34] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge Graph Embedding by Translating on Hyperplanes. In *Association for the Advancement of Artificial Intelligence*. 1112–1119.

[35] Y. Wei, Y. Zhao, C. Lu, and S. Wei. 2016. Cross-Modal Retrieval With CNN Visual Features: A New Baseline. *IEEE Transactions on Cybernetics* 47, 2 (2016), 1–12.

[36] Lexing Xie and Xuming He. 2013. Picture tags and world knowledge: learning tag relations from visual semantic sources. In *ACM international conference on Multimedia*. 967–976.

[37] Lexing Xie and Haixun Wang. 2015. Learning Knowledge Bases for Multimedia in 2015. In *ACM international conference on Multimedia*.

[38] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. 2017. Scene Graph Generation by Iterative Message Passing. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[39] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575* (2014).

[40] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, and others. 2014. Relation Classification via Convolutional Deep Neural Network. In *International Conference on Computational Linguistics*. 2335–2344.

[41] Dongxu Zhang and Dong Wang. 2015. Relation Classification via Recurrent Neural Network. *Computer Science* (2015).

[42] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. 2017. Visual Translation Embedding Network for Visual Relation Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*.