

Deep Attentive Multimodal Network Representation Learning for Social Media Images

FEIRAN HUANG, College of Cyber Security, Jinan University

CHAOZHUO LI, Microsoft Research at Asia, Beijing, China

BOYU GAO, College of Cyber Security, Jinan University

YUN LIU, School of Computer Science and Engineering, Beihang University

SATTAM ALOTAIBI, Head of Innovation and Entrepreneurship Center, College of Engineering, Taif University

HAO CHEN, School of Computer Science and Engineering, Beihang University

The analysis for social networks, such as the socially connected Internet of Things, has shown a deep influence of intelligent information processing technology on industrial systems for Smart Cities. The goal of social media representation learning is to learn dense, low-dimensional, and continuous representations for multimodal data within social networks, facilitating many real-world applications. Since social media images are usually accompanied by rich metadata (e.g., textual descriptions, tags, groups, and submitted users), simply modeling the image is not effective to learn the comprehensive information from social media images. In this work, we treat the image and its textual description as multimodal content, and transform other meta-information into the links between contents (such as two images marked by the same tag or submitted by the same user). Based on the multimodal content and social links, we propose a **Deep Attentive Multimodal Graph Embedding** model named **DAMGE** for more effective social image representation learning. We introduce both small- and large-scale datasets to conduct extensive experiments, of which the results confirm the superiority of the proposal on the tasks of social image classification and link prediction.

CCS Concepts: • **Information systems** → **Multimedia and multimodal retrieval**; • **Computing methodologies** → **Image representations**;

Additional Key Words and Phrases: Social image, graph convolutional network, multimodal, attention network, representation learning

This work was supported in part by the National Natural Science Foundation of China (No. 61932011, No. 61932010, No. 61906075, No. 61906074, No. 61902147, No. 62002068), Natural Science Foundation of Guangdong Province, China (No. 2019A1515011920, 2019A1515011276), and Guangdong Provincial Key R&D Plan (Grant No. 2019B1515120010, 202020022911500032, 2019B010136003).

Authors' addresses: F. Huang and B. Gao (corresponding author), College of Cyber Security, Jinan University, No. 601, West Huangpu Avenu, Guangzhou, China, 510632; emails: {huangfr, bygao}@jnu.edu.cn; C. Li, Microsoft Research at Asia, Beijing, China, Danling Street No. 5, Haidian Dist., Beijing, China, 100080; email: cli@microsoft.com; Y. Liu and H. Chen, School of Computer Science and Engineering, Beihang University, No. 37, Xueyuan Rd., Haidian Dist., Beijing, China, 100191; emails: {gz_liuyun, chh}@buaa.edu.cn; S. Alotaibi, Head of Innovation and Entrepreneurship Center, College of Engineering, Taif University, Airport Rd., Al Hawiyah Area, Taif, Saudi Arabia, 26571; email: srotaibi@tu.edu.sa.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Association for Computing Machinery.

1533-5399/2021/06-ART69 \$15.00

<https://doi.org/10.1145/3417295>

ACM Reference format:

Feiran Huang, Chaozhuo Li, BoYu Gao, Yun Liu, Sattam Alotaibi, and Hao Chen. 2021. Deep Attentive Multimodal Network Representation Learning for Social Media Images. *ACM Trans. Internet Technol.* 21, 3, Article 69 (June 2021), 17 pages.
<https://doi.org/10.1145/3417295>

1 INTRODUCTION

Social networks, such as socially networked sensors and mobile network, have changed the way we communicate, collaborate, and live in this society. The large amounts of data on these networks have raised the urgent demands of analyzing the topological and multimodal content effectively and efficiently for smart cities. Social media image is one of the most common information carriers on social media websites such as Facebook, Instagram, and Flickr. Different from independent images in the database, social media images are associated with plenty of meta-info such as textual descriptions, tags, groups, commentators, and submitted users. In order to learn effective embedding for social images, it is necessary to incorporate this auxiliary meta-information into the representation learning process. Learning the representation for social media images facilitates many network-based and content-based applications, such as social image classification [2, 32, 34, 54], cross-modal retrieval [20, 49], and link prediction [14, 47].

Recently, much research has been focused on learning the representation for social images, which can be roughly categorized into content-based and network-based approaches. The first category of methods considers the meta-information of social images as auxiliary content and integrates these data with multimodal embedding models to learn the representation [16, 17, 26, 28]. Li et al. [26] propose a collaborative embedding method to project social images and corresponding tags to a joint space by combining collaborative factor analysis and end-to-end network. Huang et al. [16] propose a joint social image embedding framework to embed social links and multimodal content jointly. It employs an attention model to learn the reflections between the social images with tags and triplet metric learning to the relations between images. Network-based methods treat social images as the nodes in the social networks and utilize network embedding approaches to learn the social image representation from the network perspective [6, 14, 18, 27]. Cui et al. [6] propose a relational CNN with regularization to integrate a multi-type relational knowledge graph to mine the interactions between concepts for social image embedding. Liao et al. [27] propose an attributed social network embedding method to combine attribute proximity and structural proximity for representation learning.

Although many approaches have been proposed to study social image embedding as mentioned above, most of them focus on one side of the representation learning, i.e., either the network side or the content side. Figure 1 shows one social image on Flickr. This image is accompanied with the description “A boy is playing with his dog,” the tags “boy” and “soccer,” and the submitted group “Anything Nature.” This image is linked to another image if the two have the same tags, are included in the same group, or are commented by the same user, and so on. One can see that the image and its description compose the multimodal content, and the links among images constitute the network side of the social images. Only from the network side, it is not capable of excavating close relations between multiple modalities of data, thereby learning effective multimodal representation. Only from the content side, we cannot fully extract the topological network information within the social links for effective node embedding. Although AMVAE [14] is proposed to integrate content information and graph structure for embedding learning, it encodes node vectors directly with a variational auto-encoder from the adjacency matrix, which results in a huge amount of



Fig. 1. A social image example on the website of Flickr. Images are accompanied with rich metadata, such as a description, several tags, and some groups. The image and the description constitute the multimodal content. The social links among social images can be established by the relations of submitted to the same group, marked by the same tag, and so on.

parameters for computation. Hence, it is difficult to implement this method on very large-scale social networks.

In this work, our aim is to learn social image embedding by incorporating the images with corresponding auxiliary metainformation for joint embedding with consideration of both the network side and the content side. It is of great challenge due to the following reasons. First, there exists a fine-grained relation between visual content and textual content of the social image. Taking Figure 1 as an example, the words “boy” and “dog” in the text description can be reflected to the special objects or areas in the corresponding image. This kind of fine-grained alignment is easily recognized by human eyes but is difficult to be automatically discovered by computer models. Second, as mentioned above, there exist many types of social links constructed by metadata (tags, groups, users, etc.) among social images. These links are usually highly nonlinear, and it is nontrivial to design an appropriate network embedding approach to integrate this information.

To deal with these challenges, a **Deep Attentive Multimodal Graph Embedding** model named **DAMGE** is proposed. It combines multi-modalities and social links for more effective social image representation learning. Specifically, we first build a multi-level visual attention network to learn joint multimodal representation by exploring the close correlation between the image and its textual description. The representation and the social links are then used to build a multimodal graph, and a **graph convolutional network (GCN)** is employed to further learn the node representation. Through joint optimization, the two networks can be assembled into a holistic framework and mutually reinforced to learn the embedding for social images. The main contributions are mainly threefold:

- We propose DAMGE to address the challenges of incorporating multimodal content and graph structure for social image embedding.
- Different from traditional social image embedding methods, our approach learns social images from both the network side and the content side with a multi-level visual attention network and a graph convolutional network. The embedding can be integrated and mutually reinforced by the two networks.

- We introduce both small- and large-scale datasets to conduct extensive experiments. The results confirm the superiority of the proposal on the tasks of social image classification and link prediction.

The rest of this work starts from related works. Section 3 presents the proposed method DAMGE. Experiments are presented in Section 4. Finally, conclusions are drawn in Section 5.

2 RELATED WORKS

2.1 Multimodal Embedding

Multimodal embedding involves learning the representation by relating information from multiple modality sources [8, 13, 19, 21, 39, 50, 52]. One typical kind of method is to embed each modality into a separate representation and use a specific constraint to coordinate them. **Canonical Correlation Analysis (CCA)** [13] and **Partial Least Squares (PLS)** [39] try to project multimodalities to a latent vector space by relation maximization between different modalities. **Kernel CCA (KCCA)** [13] uses the kernel trick to make the correlation learning nonlinear. With the advent of deep learning age, **deep CCA (DCCA)** [50] is proposed to employ deep neural networks upon canonical correlation analysis, which makes it applicable to highly dimensional embedding of the image and text. Another strategy is to fuse multiple modalities into a joint representation by neural networks or other models [9, 11, 15, 35, 41, 46]. Ngiam et al. [35] propose a multimodal autoencoder with the initialization of sparse RBMs to study cross-modal learning, the fusion of multi-modalities, and shared embedding learning. Srivastava and Salakhutdinov [41] design a deep Boltzmann machine in multimodal architecture to fuse modalities in the hidden neurons for the tasks of generating missing modalities, inferring joint representations, and classification. Huang et al. [15] design an adversarial attention framework to incorporate adversarial networks and an attention mechanism to learn robust multimodal representation to address the noise within social multimodal data.

2.2 Network Embedding

Early methods for network embedding, e.g., Laplacian Eigenmap [1] and IsoMap [42], use feature vectors to build an affinity graph and then embed the graph nodes as low-dimension vectors. Inspired by the recent works on word embedding, DeepWalk [37] and its extensions [10, 24, 25, 51] are proposed to learn latent embeddings of graph nodes with a skip-gram model to encode features after the process of random walk. Recently, deep learning has been employed to introduce deep neural techniques over the graph data [3, 12, 23, 44, 45], which are more effective and scalable compared to traditional methods. A GCN [23] is proposed to design a convolutional operation-based network with an approximation of spectral graph convolutions in the context of node classification. FastGCN [3] is proposed to improve the time efficiency of the GCN by importance sampling and generalize well from transductive learning to inductive learning. GraphSAGE [12] designs an aggregation method to fuse node information from a local neighborhood using degrees or attributes of nodes for inductive learning. GAT [44] is proposed to introduce an attention mechanism into GCNs to draw different weights on the neighborhoods' features of central nodes. DGI [45] learns unsupervised graph embedding by setting a loss function to maximize the mutual information between global graph summary and patch representation.

3 SOCIAL IMAGE REPRESENTATION LEARNING

In this section, we first define the studied problem and present the framework of DAMGE. Then we present the multi-level visual attention network and graph convolutional network to learn

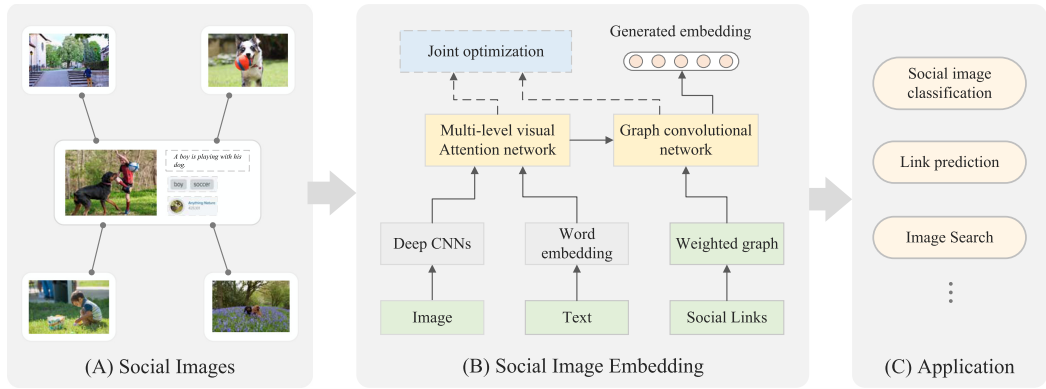


Fig. 2. The framework of DAMGE. (a) Social images containing rich metainformation of multimodal content and social links. (b) The proposed approach for social image representation learning. (c) Applications based on the learned representation.

multimodal embedding and graph embedding, respectively. Based on the two networks, we propose a joint optimization framework to learn social image embedding.

3.1 Problem Statement

Given social image set M , we obtain the corresponding textual description set T and social links from the metainformation of social images. These social links are established if two images are included in the same group, marked by the same tag, uploaded by the same user, or commented by the same commentator. Based on these social links, we build a weighted graph $G = (V, E, M, T)$, where V denotes the set of nodes, E is the edges (or links), and M and T are the image and text sets on nodes, respectively. Since multiple types of links are constructed by different metadata, each edge $e \in E$ between two nodes is related with a value of k which represents how many types of links between these two nodes. Our aim is to learn a deep embedding model to embed the graph $G = (V, E, M, T)$, such that each node $v_i \in V$ can be represented as a high-level representation h_i . These representations can well exploit the multimodal content M and T and social link information E , and then be utilized for downstream tasks, e.g., social image classification and link prediction.

The framework of the proposed DAMGE is illustrated in Figure 2. For each social image and its textual description, we first propose a multi-level visual attention model to learn the multimodal embedding by merging these two modalities with deep fusion. This model well explores the fine-grained relation between the image and text by two types of visual level attentions, i.e., object level and region level. The multimodal embedding and multiple types of social links are then used to build a graph with weighted edges. An unsupervised GCN is employed to address the weighted graph to learn graph embedding by maximization of the mutual information between whole graph and patch presentation. Finally, the attention network and GCN are integrated into a holistic model with joint optimization to learn the embedding for social images.

3.2 Multimodal Embedding

For the sake of learning the representation for the multimodal content of social images, we need to mine the relation between the image and text such that the complementary information can be fully fused. Recently, many approaches have adopted an attention mechanism applying to language and vision-related tasks, such as visual question answering [29–31], image caption [48, 53],

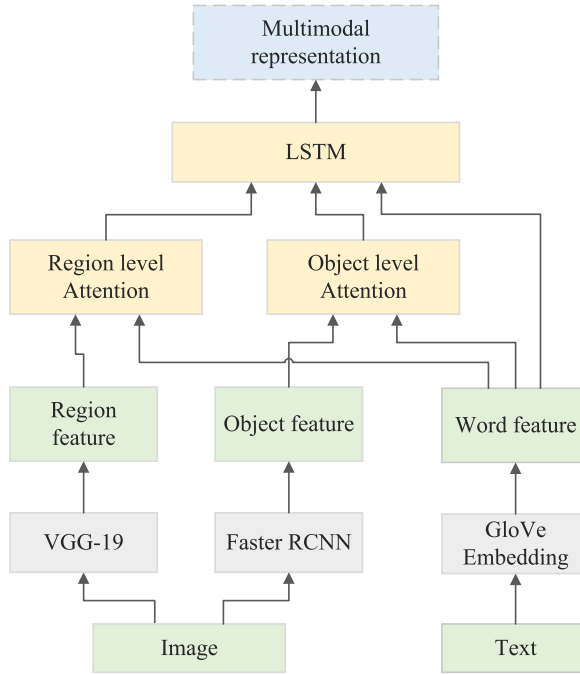


Fig. 3. The structure of the proposed multi-level visual attention model.

and machine translation [33, 43], which have achieved satisfactory performance. An attention mechanism can selectively focus on small areas or certain important words in visual or textual content to extract salient features as well as reduce the redundant information to be processed. Different from other attention models, we employ a multi-level visual attention network to capture the fine-grained correlation of both object-word level and region-word level to learn multimodal embedding. Figure 3 presents the detailed structure of our multi-level visual attention network for multimodal representation learning.

Given an image $m \in M$ and its description $t \in T$, we first conduct preprocessing for feature extraction. For the text t , we employ pretrained word embedding of GloVe [36] to represent each word in the sentence and denote it as $t = \{w_1, w_2, \dots, w_{n_w}\}$, where n_w is the length of the sentence. For the image, we use two types of deep CNNs to extract deep features of the image, i.e., region-level feature $m^{(1)} = \{r_1, r_2, \dots, r_{n_r}\}$ by VGG-19 [40] and object-level feature $m^{(2)} = \{o_1, o_2, \dots, o_{n_o}\}$ by Faster RCNN [38].

Region-word attention. In the region-word attention module, for each word w_i , an attentive score $\alpha_{i,j}$ is assigned to each region r_j in the image to discover the relation between the word with all regions. By introducing a bi-linear function, the attention map over the n_r regions can be calculated through a softmax activation:

$$\alpha_{i,j} = \text{softmax}(\tanh(w_i^T U^{(r)} r_j + b^{(r)})), \quad (1)$$

where $U^{(r)}$ and $b^{(r)}$ are weight and bias parameters, respectively. This attention score automatically calculates the relevance between word w_i and region r_j by parameter learning. $\tanh$ is an activator to make the network nonlinearity and softmax is the activation to normalize the attention score between 0 and 1. Then the attended region features are computed by weighted summation of the

original region features based on attention scores as follows:

$$r^i = \sum_{j=0}^{n_r} \alpha_{i,j} r_j. \quad (2)$$

Object-word attention. Likewise, for each word w_i , an attentive score $\beta_{i,j}$ is assigned to each object o_j in the image to discover the relation between the word with all objects. The attention map over the n_o objects can be calculated as:

$$\beta_{i,j} = \text{softmax}(\tanh(w_i^T U^{(o)} o_j + b^{(o)})), \quad (3)$$

where $U^{(o)}$ and $b^{(o)}$ are weight and bias parameters, respectively. This attention score automatically calculates the relevance between word w_i and object o_j by parameter learning. Intuitively, we consider that the original region feature should be related with the attention score at each corresponding visual region. Then the attended object features are computed by weighted summation of the original region features based on attention scores as follows:

$$o^i = \sum_{j=0}^{n_o} \beta_{i,j} o_j. \quad (4)$$

Compared to the original region and object features, the attended region features r^i and object features o^i can reflect the important regions and objects regarding the corresponding word more effectively. To learn the multimodal representation, we concatenate attended region feature r^i , object feature o^i , and word feature w_i as a vector $c_i = [r^i; o^i; w_i]$, and feed it into an LSTM in each timestep. In this way, the image and text features can be fused as a multimodal sequence $\{c_1, c_2, \dots, c_m\}$ to the LSTM for semantic learning. The last cell's output x is then obtained as the multimodal representation:

$$x = \text{att}(m, t), \quad (5)$$

where $\text{att}(\cdot)$ is the function of visual attention network. Although the embedding x is obtained, there is no objective function to help learn the parameters within the attention model. Inspired by the recent work on image-text retrieval, we introduce a siamese metric to learn the representation by making a comparison over the matching of the image and text. To achieve this goal, we need to make the model distinguish the relevance and difference between an image and a text. Therefore, given the image m and its corresponding text t , we also randomly sample negative text, which is not related to the image m . Then both positive pair (m, t) and negative pair (m, t^-) are input to the attention network. We employ a pairwise ranking loss to learn the representation:

$$L_1 = \sum_{M, T} \max(0, \text{Margin} - f(\text{att}(m, t)) + f(\text{att}(m, t^-))), \quad (6)$$

where $f(\cdot)$ is the function to learn the matching score based on the output $\text{att}(m, t)$ and $\text{att}(m, t^-)$ of the LSTM. By this loss, it is expected that the matching score obtained by $f(\cdot)$ is not less than the margin value *Margin* for the positive pair (m, t) compared with the negative pair (m, t^-) . Specifically, a two-layer fully connected network with the activation of $\tanh$ is employed to learn the matching score.

3.3 Graph Embedding

Apart from the multimodal content information, the social links among social images also have abundant auxiliary information which can be learned for more comprehensive representation learning. However, the social links provide more abstract and higher-level information, which is hard to be exploited by traditional methods. Recently, GCNs [12, 23, 44] have been proposed for

network embedding by deep learning techniques. Motivated by the deep graph infomax [45], we build the GCNs on the weighted network with an unsupervised manner.

Given the graph $G = (V, E, M, T)$ of social images, the multimodal features can be obtained from the attention model as $X = att(M, T)$. Then the given graph can be converted to $G = (V, E, X)$. The basic GCN uses the first-order convolution to collect information from one-hop neighbors and generate the hidden representation of a higher layer. The k -th convolutional layer is denoted as $H^{(k)}$ and the 0-th layer is the input multimodal features:

$$H^{(0)} = X. \quad (7)$$

Given $H^{(k)}$ as the hidden representation of the k -th layer, the $(k + 1)$ -th layer is then calculated as follows:

$$H^{(k+1)} = \sigma(\hat{A}H^{(k)}W^k), \quad (8)$$

where $\hat{A}$ is denoted the adjacency matrix with normalization and self-loops, W^k is the parameter matrix to be learned, and σ is the activator to make the process nonlinear. However, the matrix multiplication of the original GCN is very time-consuming and the convolution operation cannot be directly generalized to unobserved nodes. Therefore, we employ the mean aggregator of GraphSAGE [12] to generalize the original GCN for unsupervised learning for social images.

Let $H = \{h_1, h_2, \dots, h_n\}$ be the generated embeddings of each node. The operation of aggregation results in that the node embeddings, h_i , summarize a patch of the graph around the central node i instead of the node i itself. Hence, the embedding h_i is called *patch representation*. On the other hand, we find another embedding s to capture the global information of the entire graph, which is called *summary representation*. To simplify the process, we average the patch representation of all nodes to obtain summary representation:

$$s = \text{sigmoid}\left(\frac{1}{n} \sum_{i=1}^n h_i\right). \quad (9)$$

Following [45], we define a fully connected network $g(\cdot)$ to learn the score of mutual information on each node as $g(h_i, s)$. To make a comparison, we also build a corrupted graph $\tilde{G} = (V, E, \tilde{X})$ to calculate the contrastive mutual information score. The graph $\tilde{G}$ has the same edges and structure with the original graph G , but the input vector feature $\tilde{X}$ is obtained by shuffling the row order of X . In other words, these two graphs have the same nodes but the nodes are located in different places, and thus generate different patch representations.

Then our aim is to maximize the mutual information between the patch representation h_i and summary representation s in the original graph, while minimizing the mutual information of the patch representation $\tilde{h}_i$ and summary representation s in the corrupted graph. That is to say, we need to enlarge $g(h_i, s)$ and reduce $g(\tilde{h}_i, s)$ to encourage the model to make the natural graph globally relevant. Note that the original and corrupted graphs have the same summary representation s since the input feature vectors have the same values but in different order. Similar to [45], we employ a binary cross-entropy loss between positive samples and negative samples:

$$L_2 = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_G \log(g(h_i, s)) + \sum_{i=1}^n \mathbb{E}_{\tilde{G}} \log(1 - g(\tilde{h}_i, s)), \quad (10)$$

where $\log(\cdot)$ is the cross-entropy function. With this loss, the node embedding is trained to preserve mutual information with the graph summary.

ALGORITHM 1: The social image representation learning algorithm of DAMGE.**Input:**

The weighted graph $G = (V, E, M, T)$ constructed by social image set and corresponding metainformation;

Hyperparameters: balance parameter α , embedding size d .

Output: Learned node embedding.

```

1: repeat
2:   for  $v$  in  $V$  do
3:     Obtain image  $m \in M$  and text description  $t \in T$ .
4:     Randomly sample the negative text  $t^-$  from  $T$ .
5:     Calculate the hinge-ranking loss  $L_1$  fed with  $(m, t, t^-)$  by Equation (6).
6:     Calculate the cross-entropy loss  $L_2$  by Equation (10).
7:     Calculate the whole loss  $L$  by Equation (11).
8:     Update the parameters of DAMGE via SGD on  $L$ .
9:   end for
10: until DAMGE converges

```

3.4 Joint Embedding

As mentioned above, a multi-level visual attention model is proposed to learn the embedding of multimodal content and a GCN is proposed to learn the node representation on the social weighted graph. In order to effectively learn the embedding for social images, it is necessary to integrate these two models to exploit both content information and network structure for joint embedding. A straightforward way to unify these two models is to learn the embedding with two steps, i.e., multimodal content embedding learning step with ranking loss (Equation (6)), and then graph embedding learning with cross-entropy loss (Equation (10)). However, this type of combining manner cannot reinforce these two models to learn from each other for joint embedding learning.

Therefore, in the proposed deep attentive multimodal graph embedding model DAMGE, we conduct the multimodal content embedding and graph embedding simultaneously to find the intrinsic relations. DAMGE's basic design concept is that the semantic content embedding network and the graph embedding network can constitute a learning framework of mutual reinforcement. Based on such analysis, the loss of DAMGE is formulated as the summation of the margin ranking loss (Equation (6)) embedding and the cross-entropy loss (Equation (10)) for joint social image embedding:

$$L = L_1 + \alpha L_2, \quad (11)$$

where the hyperparameter α is used to balance the weight of the second loss. To show the complete social image representation learning process, we present the details of the proposed DAMGE in Algorithm 1.

The proposed DAMGE is trained by minimizing this joint loss with the optimizer of mini-batch stochastic gradient descent.

4 EXPERIMENTS

In this section, we first present the descriptions of the datasets, experimental settings, and compared methods. Then the proposed DAMGE is evaluated on the tasks of social image classification and link prediction.

Table 1. Dataset Statistics

	PASCAL	MIR	CLEF	NUS-WIDE
#node	5,912	5,193	3,930	121,152
#edge	74,231	111,520	78,356	19,216,795
#link	181,812	301,782	133,214	56,811,268

4.1 Datasets

We conduct the experiments on four real-world datasets of social images including MIR, CLEF, PASCAL, and NUS-WIDE. These datasets are all gathered from the website of Flickr and are annotated with multi-labels. Based on these datasets, the metadata of social images in the datasets is gathered in [34] and shared in the Stanford Network Analysis Project.¹ The detailed information of these datasets is presented as follows:

- PASCAL [7] is a widely used dataset, which has been built and updated for image recognition. 9,963 images are included in PASCAL 2007, but only 9,474 of them are available on Flickr during experiments.
- MIR [22] dataset consists of a million images downloaded from the Flickr. Among them, 25,000 images are associated with manual annotations, but only 13,368 images are available on the Flickr during experiments.
- CLEF [5] is a subset of the MIR dataset, which contains 18,000 images, but only 8,000 images have the correspondence. Among them, 4,179 of the annotated images are available during experiments.
- NUS-WIDE [4] dataset is a social image database for image classification and retrieval, which consists of 269,648 annotated images. Among them, only 226,912 are available during experiments.

Besides images, we also collect the corresponding descriptions, tags, locations, groups, users, and commentators. The images having no textual content or with meaningless texts are removed. As for the social links for building networks, five types of relations are utilized, i.e., included in the same group, taken from the same place, commented by the same commentator, having the same tag, and uploaded by the same user. Based on these links, we establish an edge once at least two types of links exist between two nodes and the weight score of edges are set as the number of links. Table 1 shows the summary of the datasets utilized in the experiments.

4.2 Experimental Settings

For images, we resize each picture to the shape of 224×224 and then employ two types of deep CNNs pretrained on Imagenet² to extract visual features: (1) We employ VGG-19 [40] networks and extract the region feature at the layer of *conv5_4*, of which the output shape is $512 \times 14 \times 14$. This means that there are 14×14 regions that need to be attended on an image and each region has 512-dimensional features. (2) We employ Faster RCNN [38] for object detection in the image and obtain the top-10 ranked objects in the region proposal network. The output of the layer *fc7* is selected as object features with the dimension of 2,048 to the attention model. For texts, we employ the word embedding of GloVe [36] pretrained on Wikipedia data, and each word in the description is embedded as a 512-dimensional vector. The LSTM is set to contain 256 hidden neurons in each

¹<https://snap.stanford.edu/data/web-flickr.html>.

²<http://www.image-net.org/>.

timestep. The MLP for pairwise ranking learning is a fully connected network with the neural structure of $256 \times 128 \times 1$ through *tanh* activations. The margin value in the siamese network is set to be 0.04. The GCN is set to have two hidden layers, and the nodes are embedded as 256-dimensional representations. The balance parameter in the joint loss function (Equation (11)) is set to be 0.8.

4.3 Compared Methods

DAMGE is compared with five different embedding methods, including two network embedding methods (DeepWalk, node2vec), three multimodal content embedding methods (Bimodal-AE, M-DBM, DCCA), and three mixed embedding methods (TADW, DGI, AMVAE).

- **DeepWalk** [37] embeds social relations by using random walk to generate node sequence and then employing a skip-gram algorithm to learn the representation.
- **node2vec** [10] uses biased random walk based on DeepWalk to balance between global and local properties of a graph.
- **Bimodal-AE** [35] is a multimodal autoencoder with the initialization of RBM for representation learning.
- **M-DBM** [41] uses a multimodal deep Boltzmann machine method to learn data representations of multimodalities.
- **DCCA** [50] is proposed to employ deep neural networks upon canonical correlation analysis to embed the content of image and text.
- **TADW** [51] incorporates textual features and network structure for network embedding with a **matrix factorization (MF)** approach.
- **DGI** [45] learns unsupervised graph embedding by maximizing the mutual information between global graph summary and patch representation.
- **AMVAE** [14] integrates multimodal content and network structure with attention-based variational autoencoder for embedding learning.

To make a fair comparison, the visual and textual feature for the compared methods with content input are the output of the last average pooling layer of VGG-19 [40] and GloVe [36] embeddings, respectively.

4.4 Social Image Classification

To conduct classification, we first use different embedding methods to generate the representation for each image. Then we use a simple common classification, namely, Logistic Regression, for all the compared methods. The dataset is randomly split as training and testing set with the proportion of $p\%$ and $1 - p\%$, and calculates the results of the **mean average precision (mAP)** score on the testing sets. The whole process is repeated 10 times, and we present the average value of mAP.

The experimental results are shown in Table 2 on four datasets. $p\%$ denotes the proportion of the training set in each dataset. “-” means unreported results because the method of TADW is not capable of addressing large-scale network embedding of NUS-WIDE due to its matrix factorization. One can see that the proposed DAMGE performs consistently better than compared methods on all datasets. DeepWalk and node2vec are embedding methods that totally depend on the network structure, which results in they cannot learn the content information from the multimodalities of the image and text. Hence, these two methods show relatively low performance. Similarly, Bimodal-AE, M-DBM, and DCCA are content-based methods which do not consider the social link information within social images, thus obtaining unsatisfactory performance. HNE, TADW, and AMVAE are the state-of-the-art approaches which integrate content information and graph structure for embedding learning. It can be seen that they outperform the approaches depending on

Table 2. The Results of Social Image Classification on Four Datasets

Dataset	PASCAL			MIR			CLEF			NUS-WIDE		
$p\%$	10%	20%	30%	10%	20%	30%	10%	20%	30%	10%	20%	30%
DeepWalk	32.0	40.5	44.7	51.0	56.6	57.5	29.3	32.3	34.3	39.0	38.0	39.6
node2vec	31.8	36.9	42.5	50.4	55.5	57.0	30.3	32.9	33.4	36.3	41.3	42.8
Bimodal-AE	34.4	37.2	41.6	51.8	55.8	59.5	27.4	32.5	31.7	35.9	41.3	41.4
M-DBM	37.1	39.6	42.1	52.5	55.8	55.9	32.4	32.0	34.5	39.2	39.7	42.3
DCCA	39.9	39.6	43.5	54.9	60.1	58.5	31.4	32.1	36.5	39.5	39.4	42.8
HNE	40.2	45.5	49.5	61.9	67.8	68.9	36.6	38.4	40.3	46.7	48.7	50.4
TADW	42.4	47.3	49.7	62.1	66.0	68.4	35.7	38.2	39.1	-	-	-
AMVAE	50.5	53.9	58.4	70.2	73.0	76.3	40.2	43.2	44.9	55.0	57.6	59.7
DAMGE	53.7	55.5	61.0	72.5	75.7	79.6	42.7	45.8	47.5	58.8	60.4	62.5

one-side information with large margins, which demonstrates the importance to combine the rich meta-information from social images for representation learning. Compared to these three mixed embedding methods, the proposed DAMGE still obtains a firm improvement on all four datasets. It is because our DAMGE model integrates the multi-level visual attention network and GCN into a joint optimization framework, which can fully exploit multimodal content and social links to learn the representation for social images.

4.5 Link Prediction

The learned social image embeddings can also be utilized for link prediction, of which the objective is to predict missing links or links probably to be established in the future. Link prediction is very useful in recommendation tasks such as predicting preference between user and items, and inferring the friends connections between users in social networks. Therefore, the task of link prediction is capable of evaluating the learned embeddings by different methods. The experiments are conducted in the following procedures. First, we randomly delete half the original edges among the network of social images and treat the removed edges as the target links to be predicted. Then, the social images with the remaining network are fed into embedding approaches to generate embeddings of social images. The original node pairs which have the removed edges are labeled as positive link samples. For comparison, we also generate the same amount of negative link samples by randomly sampling unlinked node pairs. For all the positive and negative node pairs, we calculate the cosine similarity between node pairs based on the embeddings learned by different methods. **Area Under Curve (AUC)** is employed to measure the consistency between similarity scores and real link states. The experimental results of link prediction for the four datasets are shown in Table 2. One can see that the proposed DAMGE outperforms the compared approaches consistently. DAMGE goes far beyond the mixed embedding methods of HNE and TADW, and makes solid improvements over the state-of-the-art approach AMVAE. It validates that our proposal is effective to incorporate multimodalities and social links for social image representation learning.

4.6 Parameter Sensitivity

In the proposed DAMGE, the dimension of the generated representation and the value of balance parameter in the joint loss are two important factors which could affect the performance of DAMGE. In this section, we progressively increase the embedding dimension from

Table 3. The Results of Link Prediction on Four Datasets

Datasets	PASCAL	MIR	CLEF	NUS-WIDE
DeepWalk	0.555	0.660	0.600	0.625
node2vec	0.562	0.670	0.634	0.632
Bimodal-AE	0.556	0.648	0.597	0.643
M-DBM	0.567	0.686	0.654	0.680
DCCA	0.611	0.661	0.661	0.656
HNE	0.707	0.715	0.732	0.733
TADW	0.729	0.694	0.725	-
AMVAE	0.772	0.864	0.845	0.840
DAMGE	0.809	0.887	0.872	0.878

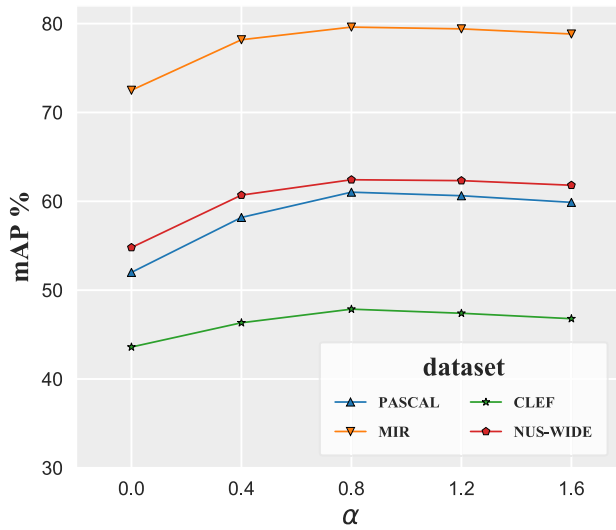


Fig. 4. Parameter sensitivity study on the embedding dimension and balance parameter.

[32, 64, 128, 256, 512, 1,024], and balance parameter α from [0,0.4,0.8,1.2,1.6] to evaluate the effectiveness of DAMGE under different parametrization.

Embedding size: First, we set the balance parameter to be 0.8 and vary the embedding size to show how the performance of DAMGE changes. The results of social image classification are presented in Figure 4. It can be seen that when the size of embedding becomes larger, the mAP results increase at the beginning as more information can be encoded with more bits. Our approach achieves the best performance when the dimension is around 256. This suggests that it is not necessary to set very large embedding size.

Balance parameters: We also evaluate DAMGE with different values of balance parameter α as shown in Figure 5. One can see that DAMGE behaves relatively bad when $\alpha = 0$. This is because under this circumstance, there is no loss to help the model learn the information from network structure for embedding. However, too big α also degrades the performance as too much attention drawn on the graph embedding will neglect the representation learning from the multimodal content.

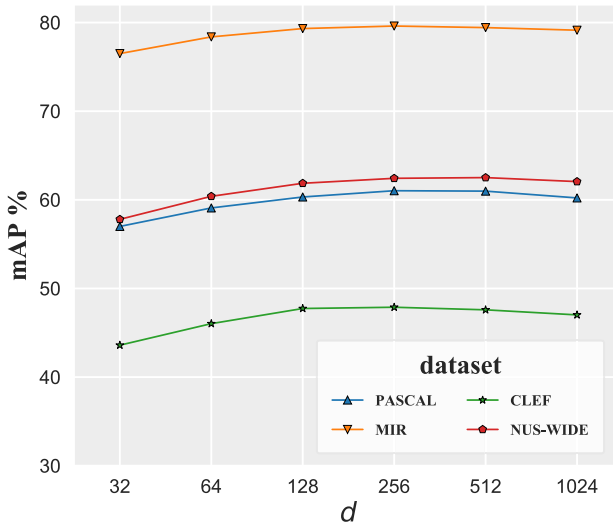


Fig. 5. Parameter sensitivity study on the balance parameter.

5 CONCLUSION

In this work, we make full use of the rich metadata within social images for embedding learning. We exploit the meta-information from both the multimodal content side and the social network side, and propose a Deep Attentive Multimodal Graph Embedding model named DAMGE for effective social image embedding. Specifically, the multi-level visual attention network well excavates the fine-grained relation between the image and text to learn multimodal representation. Then we transform social links into a weighted graph and employ a GCN to learn the node embedding. Finally, these two networks are reinforced mutually by a joint loss with a holistic framework. The experimental results on four real-world web datasets confirm the effectiveness of DAMGE on the tasks of social image classification and link prediction. Different from existing approaches of social image embedding, the proposed DAMGE can integrate multimodalities and social links for joint representation learning. The future work involves exploiting the heterogeneous relation among users, images, tags, and so on, to learn the representation of social images more effectively.

REFERENCES

- [1] Mikhail Belkin and Partha Niyogi. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* 15, 6 (2003), 1373–1396. <https://doi.org/10.1162/089976603321780317>
- [2] Jiuwen Cao, Kai Zhang, Minxia Luo, Chun Yin, and Xiaoping Lai. 2016. Extreme learning machine and adaptive sparse representation for image classification. *Neural Netw.* 81 (2016), 91–102. <https://doi.org/10.1016/j.neunet.2016.06.001>
- [3] Jie Chen, Tengfei Ma, and Cao Xiao. 2018. FastGCN: Fast learning with graph convolutional networks via importance sampling. In *6th International Conference on Learning Representations (ICLR'18), Conference Track Proceedings*. OpenReview.net. <https://openreview.net/forum?id=rytstxWAW>.
- [4] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. NUS-WIDE: A real-world web image database from National University of Singapore. In *Proceedings of the 8th ACM International Conference on Image and Video Retrieval (CIVR'09)*, Stéphane Marchand-Maillet and Yiannis Kompatsiaris (Eds.). ACM. <https://doi.org/10.1145/1646396.1646452>
- [5] Paul D. Clough, Michael Grubinger, Thomas Deselaers, Allan Hanbury, and Henning Müller. 2006. Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. In *Evaluation of Multilingual and Multi-Modal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum (CLEF'06), Revised Selected Papers*, Lecture

- Notes in Computer Science, Vol. 4730, Carol Peters, Paul D. Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, and Maximilian Stempfhuber (Eds.). Springer, 579–594. https://doi.org/10.1007/978-3-540-74999-8_71
- [6] Peng Cui, Shaowei Liu, and Wenwu Zhu. 2018. General knowledge embedded image representation learning. *IEEE Trans. Multimedia* 20, 1 (2018), 198–207. <https://doi.org/10.1109/TMM.2017.2724843>
 - [7] Mark Everingham, Luc J. Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. 2010. The PASCAL visual object classes (VOC) challenge. *Int. J. Comput. Vis.* 88, 2 (2010), 303–338. <https://doi.org/10.1007/s11263-009-0275-4>
 - [8] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc’Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*, Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (Eds.). 2121–2129. <http://papers.nips.cc/paper/5204-devise-a-deep-visual-semantic-embedding-model>.
 - [9] Yue Gao, Yi Zhen, Haojie Li, and Tat-Seng Chua. 2016. Filtering of brand-related microblogs using social-smooth multiview embedding. *IEEE Trans. Multimedia* 18, 10 (2016), 2115–2126. <https://doi.org/10.1109/TMM.2016.2581483>
 - [10] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu Aggarwal, Dou Shen, and Rajeev Rastogi (Eds.). ACM, 855–864. <https://doi.org/10.1145/2939672.2939754>
 - [11] Zepeng Gu, Bo Lang, Tongyu Yue, and Lei Huang. 2017. Learning joint multimodal representation based on multi-fusion deep neural networks. In *24th International Conference on Neural Information Processing (ICONIP’17), Part II*, Lecture Notes in Computer Science, Vol. 10635, Derong Liu, Shengli Xie, Yuanqing Li, Dongbin Zhao, and El-Sayed M. El-Alfy (Eds.). Springer, 276–285. https://doi.org/10.1007/978-3-319-70096-0_29
 - [12] William L. Hamilton, Zitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 1024–1034. <http://papers.nips.cc/paper/6703-inductive-representation-learning-on-large-graphs>.
 - [13] David R. Hardoon, Sándor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.* 16, 12 (2004), 2639–2664. <https://doi.org/10.1162/0899766042321814>
 - [14] Feiran Huang, Xiaoming Zhang, Chaozhuo Li, Zhoujun Li, Yueying He, and Zhonghua Zhao. 2018. Multimodal network embedding via attention based multi-view variational autoencoder. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval (ICMR’18)*, Kiyoharu Aizawa, Michael S. Lew, and Shin’ichi Satoh (Eds.). ACM, 108–116. <https://doi.org/10.1145/3206025.3206035>
 - [15] Feiran Huang, Xiaoming Zhang, and Zhoujun Li. 2018. Learning joint multimodal representation with adversarial attention networks. In *2018 ACM Multimedia Conference on Multimedia (MM’18)*, Susanne Boll, Kyoung Mu Lee, Jiebo Luo, Wenwu Zhu, Hyeran Byun, Chang Wen Chen, Rainer Lienhart, and Tao Mei (Eds.). ACM, 1874–1882. <https://doi.org/10.1145/3240508.3240614>
 - [16] Feiran Huang, Xiaoming Zhang, Zhoujun Li, Tao Mei, Yueying He, and Zhonghua Zhao. 2017. Learning social image embedding with deep multimodal attention networks. In *Proceedings of the Thematic Workshops of ACM Multimedia 2017*, Wanmin Wu, Jianchao Yang, Qi Tian, and Roger Zimmermann (Eds.). ACM, 460–468. <https://doi.org/10.1145/3126686.3126720>
 - [17] Feiran Huang, Xiaoming Zhang, Zhoujun Li, Zhonghua Zhao, and Yueying He. 2018. From content to links: Social image embedding with deep multimodal model. *Knowl.-Based Syst.* 160 (2018), 251–264. <https://doi.org/10.1016/j.knsys.2018.07.020>
 - [18] Feiran Huang, Xiaoming Zhang, Jie Xu, Chaozhuo Li, and Zhoujun Li. 2019. Network embedding by fusing multimodal contents and links. *Knowl.-Based Syst.* 171 (2019), 44–55. <https://doi.org/10.1016/j.knsys.2019.02.003>
 - [19] Feiran Huang, Xiaoming Zhang, Jie Xu, Zhonghua Zhao, and Zhoujun Li. 2019. Multimodal learning of social image representation by exploiting social relations. *IEEE Transactions on Cybernetics* 51, 3 (2021), 1506–1518.
 - [20] Feiran Huang, Xiaoming Zhang, Zhonghua Zhao, and Zhoujun Li. 2019. Bi-directional spatial-semantic attention networks for image-text matching. *IEEE Trans. Image Processing* 28, 4 (2019), 2008–2020. <https://doi.org/10.1109/TIP.2018.2882225>
 - [21] Feiran Huang, Xiaoming Zhang, Zhonghua Zhao, Zhoujun Li, and Yueying He. 2018. Deep multi-view representation learning for social images. *Appl. Soft Comput.* 73 (2018), 106–118. <https://doi.org/10.1016/j.asoc.2018.08.010>
 - [22] Mark J. Huiskes and Michael S. Lew. 2008. The MIR flickr retrieval evaluation. In *Proceedings of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval (MIR’08)*, Michael S. Lew, Alberto Del Bimbo, and Erwin M. Bakker (Eds.). ACM, 39–43. <https://doi.org/10.1145/1460096.1460104>

- [23] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations (ICLR '17), Conference Track Proceedings*. OpenReview.net. <https://openreview.net/forum?id=SJU4ayYgl>.
- [24] Chaozhuo Li, Senzhang Wang, Dejian Yang, Zhoujun Li, Yang Yang, Xiaoming Zhang, and Jianshe Zhou. 2017. PPNE: Property preserving network embedding. In *Proceedings of the 22nd International Conference on Database Systems for Advanced Applications, (DASFAA '17), Part I*, Lecture Notes in Computer Science, Vol. 10177, K. Selçuk Candan, Lei Chen, Torben Bach Pedersen, Lijun Chang, and Wen Hua (Eds.). Springer, 163–179. https://doi.org/10.1007/978-3-319-55753-3_11
- [25] Chaozhuo Li, Lei Zheng, Senzhang Wang, Feiran Huang, Philip S. Yu, and Zhoujun Li. 2019. Multi-hot compact network embedding. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM'19)*, Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu (Eds.). ACM, 459–468. <https://doi.org/10.1145/3357384.3357903>
- [26] Zechao Li, Jinhui Tang, and Tao Mei. 2019. Deep collaborative embedding for social image understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 9 (2019), 2070–2083. <https://doi.org/10.1109/TPAMI.2018.2852750>
- [27] Lizi Liao, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2018. Attributed social network embedding. *IEEE Trans. Knowl. Data Eng.* 30, 12 (2018), 2257–2270. <https://doi.org/10.1109/TKDE.2018.2819980>
- [28] Shaowei Liu, Peng Cui, Wenwu Zhu, and Shiqiang Yang. 2015. Learning socially embedded visual representation from scratch. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference (MM '15)*, Xiaofang Zhou, Alan F. Smeaton, Qi Tian, Dick C. A. Bulterman, Heng Tao Shen, Ketan Mayer-Patel, and Shuicheng Yan (Eds.). ACM, 109–118. <https://doi.org/10.1145/2733373.2806247>
- [29] Yun Liu, Xiaoming Zhang, Feiran Huang, and Zhoujun Li. 2018. Adversarial learning of answer-related representation for visual question answering. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM'18)*, Alfredo Cuzzocrea, James Allan, Norman W. Paton, Divesh Srivastava, Rakesh Agrawal, Andrei Z. Broder, Mohammed J. Zaki, K. Selçuk Candan, Alexandros Labrinidis, Assaf Schuster, and Haixun Wang (Eds.). ACM, 1013–1022. <https://doi.org/10.1145/3269206.3271765>
- [30] Yun Liu, Xiaoming Zhang, Feiran Huang, Xianghong Tang, and Zhoujun Li. 2019. Visual question answering via attention-based syntactic structure tree-LSTM. *Appl. Soft Comput.* 82 (2019). <https://doi.org/10.1016/j.asoc.2019.105584>
- [31] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). 289–297. <http://papers.nips.cc/paper/6202-hierarchical-question-image-co-attention-for-visual-question-answering>.
- [32] Zhiwu Lu, Liwei Wang, and Ji-Rong Wen. 2014. Direct semantic analysis for social image classification. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, Carla E. Brodley and Peter Stone (Eds.). AAAI Press, 1258–1264. <http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8189>.
- [33] Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP'15)*, Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton (Eds.). The Association for Computational Linguistics, 1412–1421. <https://doi.org/10.18653/v1/d15-1166>
- [34] Julian J. McAuley and Jure Leskovec. 2012. Image labeling on a network: Using social-network metadata for image classification. In *Proceedings of the 12th European Conference on Computer Vision (ECCV'12), Part IV*, (Lecture Notes in Computer Science), Vol. 7575, Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid (Eds.). Springer, 828–841. https://doi.org/10.1007/978-3-642-33765-9_59
- [35] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML'11)*, Lise Getoor and Tobias Scheffer (Eds.). Omnipress, 689–696.
- [36] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14), A meeting of SIGDAT, a Special Interest Group of the ACL*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 1532–1543. <http://aclweb.org/anthology/D/D14/D14-1162.pdf>.
- [37] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 701–710.
- [38] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (Eds.). 91–99. <http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks>.

- [39] Abhishek Sharma and David W. Jacobs. 2011. Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. In *Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*. IEEE Computer Society, 593–600. <https://doi.org/10.1109/CVPR.2011.5995350>
- [40] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR* abs/1409.1556 (2014). <http://arxiv.org/abs/1409.1556>.
- [41] Nitish Srivastava and Ruslan Salakhutdinov. 2012. Multimodal learning with deep Boltzmann machines. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012*, Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger (Eds.). 2231–2239. <http://papers.nips.cc/paper/4683-multimodal-learning-with-deep-boltzmann-machines>.
- [42] Joshua B. Tenenbaum, Vin De Silva, and John C. Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 5500 (2000), 2319–2323.
- [43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. <http://papers.nips.cc/paper/7181-attention-is-all-you-need>.
- [44] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations (ICLR'18), Conference Track Proceedings*. OpenReview.net. <https://openreview.net/forum?id=rJXMpikCZ>.
- [45] Petar Velickovic, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. 2019. Deep graph infomax. In *Proceedings of the 7th International Conference on Learning Representations (ICLR'19)*. OpenReview.net. <https://openreview.net/forum?id=rklz9iAcKQ>.
- [46] Vedran Vukotic, Christian Raymond, and Guillaume Gravier. 2016. Bidirectional joint representation learning with symmetrical deep neural networks for multimodal and crossmodal applications. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval (ICMR'16)*, John R. Kender, John R. Smith, Jiebo Luo, Susanne Boll, and Winston H. Hsu (Eds.). ACM, 343–346. <https://doi.org/10.1145/2911996.2912064>
- [47] Zhitao Wang, Chengyao Chen, and Wenjie Li. 2017. Predictive network representation learning for link prediction. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White (Eds.). ACM, 969–972. <https://doi.org/10.1145/3077136.3080692>
- [48] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning (ICML'15) (JMLR Workshop and Conference Proceedings)*, Francis R. Bach and David M. Blei (Eds.), Vol. 37. JMLR.org, 2048–2057. <http://jmlr.org/proceedings/papers/v37/xuc15.html>.
- [49] Xing Xu, Fumin Shen, Yang Yang, Heng Tao Shen, and Xuelong Li. 2017. Learning discriminative binary codes for large-scale cross-modal retrieval. *IEEE Trans. Image Processing* 26, 5 (2017), 2494–2507. <https://doi.org/10.1109/TIP.2017.2676345>
- [50] Fei Yan and Krystian Mikołajczyk. 2015. Deep correlation for matching images and text. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. IEEE Computer Society, 3441–3450. <https://doi.org/10.1109/CVPR.2015.7298966>
- [51] Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Y. Chang. 2015. Network representation learning with rich text information. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI'15)*, Qiang Yang and Michael Wooldridge (Eds.). AAAI Press, 2111–2117. <http://ijcai.org/Abstract/15/299>.
- [52] Zhenguo Yang, Qing Li, Zheng Lu, Yun Ma, Zhiguo Gong, and Wenying Liu. 2017. Dual structure constrained multimodal feature coding for social event detection from flickr data. *ACM Trans. Internet Techn.* 17, 2 (2017), 19:1–19:20. <https://doi.org/10.1145/3015463>
- [53] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In *IEEE International Conference on Computer Vision (ICCV'17)*. IEEE Computer Society, 4904–4912. <https://doi.org/10.1109/ICCV.2017.524>
- [54] Yi Zhuang, Nan Jiang, Qing Li, Lei Chen, and Chunhua Ju. 2015. Progressive batch medical image retrieval processing in mobile wireless networks. *ACM Trans. Internet Techn.* 15, 3 (2015), 9:1–9:27. <https://doi.org/10.1145/2783437>

Received May 2020; revised July 2020; accepted August 2020