

Learning Social Image Embedding with Deep Multimodal Attention Networks

Feiran Huang
Beihang University
Beijing, China
huangfr@buaa.edu.cn

Xiaoming Zhang
Beihang University
Beijing, China
yolixs@buaa.edu.cn

Zhoujun Li
Beihang University
Beijing, China
lizj@buaa.edu.cn

Tao Mei
Microsoft Research
Beijing, China
tmei@microsoft.com

Yueying He
National Computer Network
Emergency Response Technical
Team/Coordination Center of China
Beijing, China
hyy@cert.org.cn

Zhonghua Zhao
National Computer Network
Emergency Response Technical
Team/Coordination Center of China
Beijing, China
zhaozh@cert.org.cn

ABSTRACT

Learning social media data embedding by deep models has attracted extensive research interest as well as boomed a lot of applications, such as link prediction, classification, and cross-modal search. However, for social images which contain both link information and multimodal contents (e.g., text description, and visual content), simply employing the embedding learnt from network structure or data content results in sub-optimal social image representation. In this paper, we propose a novel social image embedding approach called Deep Multimodal Attention Networks (DMAN), which employs a deep model to jointly embed multimodal contents and link information. Specifically, to effectively capture the correlations between multimodal contents, we propose a multimodal attention network to encode the fine-granularity relation between image regions and textual words. To leverage the network structure for embedding learning, a novel Siamese-Triplet neural network is proposed to model the links among images. With the joint deep model, the learnt embedding can capture both the multimodal contents and the non-linear network information. Extensive experiments are conducted to investigate the effectiveness of our approach in the applications of multi-label classification and cross-modal search. Compared to state-of-the-art image embeddings, our proposed DMAN achieves significant improvement in the tasks of multi-label classification and cross-modal search.

CCS CONCEPTS

- Information systems →Multimedia and multimodal retrieval;
- Computing methodologies →Image representations;

KEYWORDS

Social embedding, deep learning, attention model, Siamese-Triplet

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.

ThematicWorkshops'17, October 23–27, 2017, Mountain View, CA, USA

© 2017 ACM. ISBN 978-1-4503-5416-5/17/10...\$15.00

DOI: <http://dx.doi.org/10.1145/3126686.3126720>

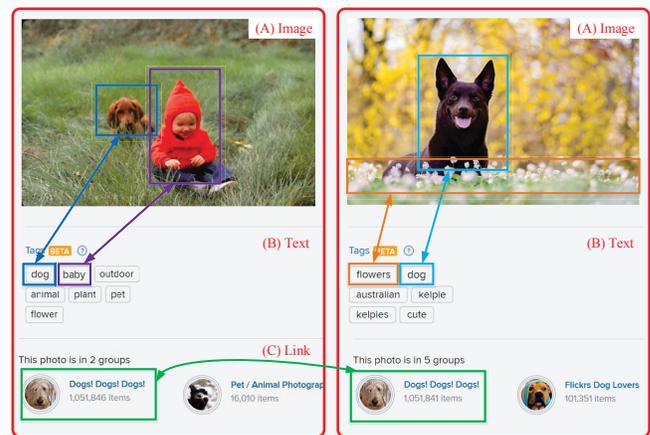


Figure 1: An example of social image: (A) a shared photo, (B) tags provided by the owner, (C) albums and galleries that contain many similar images; (A) and (B) constitute the multi-modal content of social image and (C) forms the link between images.

1 INTRODUCTION

With the rise of social network, the data like social images containing both content and link information has becoming more and more popular in miscellaneous social media (e.g., Facebook, Flickr, and Twitter), which requires an effective method to process and analyse them. Learning an efficient representation to capture the content and link information facilitates a good solution to it. The learnt social media data representations have gained great success in content-centric and network oriented applications, such as multilabel image classification, cross-modal image search, and link prediction. Therefore, how to represent the data to a vectorized space, also called social media data embedding, has been increasingly attracting attention in both academia and industry, which subsequently posits a significant challenge to social image embedding: can both the multimodal contents and links be combined for embedding learning?

It is of great challenge to deal with the social media data for embedding. First, the social images contain diverse patterns of manifestations such as image and text description. These data modalities are heterogeneous in the feature spaces. Second, there exists link relation among the data, which indicates that an efficient embedding should leverage both the nonlinear network information and data content to learn a unified representation. Third, the amount of social images on social networks has increased exponentially. Therefore, it needs an efficient method to effectively learn the embedding from the large amount of data. Figure 1 gives an example of social image.

Most of the existing social media data embedding methods can be categorized into two classes, i.e., network-based and content-based. The network-based embedding methods learn a representation for the nodes to capture the network structure, which includes the shallow model based methods, e.g., GraRep [2], Line [25] and PPNE [15], and the deep model based methods, e.g., SDNE [26]. These methods mainly use the proximity information in the network to learn the embedding, which ignores the content contained by each node. The content based methods mainly use a supervised or semi-supervised method to learn a joint representation for image and text [19, 23], which ignores the linkage among data and the fine-granularity relation between different data modalities. Though HNE [3] is proposed to combine both the network and content for embedding learning, it models different data modalities independently and the learning process is time-consuming.

Meanwhile, the multi-modal, heterogeneous and interconnected characteristics of social media data provide clues for social image embedding. First, social images does not exist in isolation, but are linked explicitly or implicitly. Though the interconnection violates the independently and identically distributed assumption in most statistical machine learning algorithms, both content and interconnection information can be exploited to complement each other for better solutions. Second, though different modalities of content are heterogeneous, there exists fine-granularity relation between them. For example, as shown in Figure 1, some words, such as “dog”, “baby”, and “flowers”, are related with the specific regions in the corresponding images. If the relation is parsed accurately, these words and visual regions can be modelled jointly in an intimate fashion and the salient features are allowed to come to the forefront as needed.

To tackle the above challenges, we propose to take advantage of the link information and multimodal contents in social images for embedding. In particular, we investigate: (1) how to capture the fine-granularity relation between different data modalities in the learnt representation; (2) how to combine the link information for social image embedding. Our solutions to these questions result in a novel approach called Deep Multimodal Attention Networks (DMAN) for social image embedding. It aims at learning social image embedding that can encode both the multimodal contents and network structure based on a joint deep model. The framework is illustrated in Figure 2. A visual-text attention model is proposed to explore the fine-granularity relation between different data modalities for social image embedding, in which the alignment between image regions and textual words is leveraged to prevent the model from being dominated by single modality. To combine the link information for embedding, the Siamese-Triplet

neural network architecture built on Convolution Neural Network (CNN) is proposed to model the network structure. Then, a joint model is proposed to integrate the two components and embed the multimodal contents and link information into a unified vector space. To improve the efficiency of model inference, we apply the positive and negative sampling method on the Triplet network, which substantially reduce the time complexity of the optimization solution. The main contributions are summarized as follows:

- Different from traditional data embedding methods, we investigate the problem of learning linkage-embedded social image embedding, where the learned embedding can well capture both the multimodal contents and network structure. Our approach is unsupervised and task independent, which makes it suitable for many network orientated and multi-modal data based data mining applications.
- We propose a joint deep model (DMAN) to address the challenges of combing content and links for embedding learning, where two models are proposed to capture multimodal contents and network structure respectively, with a deep model to integrate them.
- We conduct extensive experiments to compare the proposed model with several state-of-the-art baselines on 3 real-world datasets. The experimental results demonstrate the superiority of the proposed model.

The rest of this paper is organized as follows. Section 2 summarizes the related work. Section 3 introduces our social image embedding model in details. Section 4 presents the experimental results. Finally we conclude in Section 5.

2 RELATED WORK

Learning a representation of social media data has attracted great research attention recently. Most of existing methods can be categorized into two classes, i.e., network-based and content-based.

The network-based methods embed network into a low dimensional space, i.e. learn a vector representation for each node, with the goal of reconstructing the network in the learned embedding space. Most of the network embedding methods adopt the shallow model, e.g., DeepWalk [20], and GraRep [2], etc. DeepWalk [20] learns the latent representations of the nodes of a social network from truncated random walks in the network. GraRep [2] further extends DeepWalk to utilize high-order proximities. These methods typically first construct the affinity graph based on the feature vectors and then solve the leading eigenvectors of the graph matrices to infer the node embedding. Despite these approaches have achieved certain performance for network embedding, they all adopt the shallow model which is difficult to effectively capture the non-linear structure of the underlying network. Recently, some deep model based methods have been proposed for network embedding. SDNE [26] is a semi-supervised deep model that exploits the first-order proximity and second-order proximity to characterize the local and global network structure. However, these network based models only consider the link information and cannot be directly applied to the social media data scene where nodes have abundant content and properties. To combine the content for embedding, HNE [3] is proposed to embed heterogeneous network using a deep model, in which each node is an image or a text document. However, in social

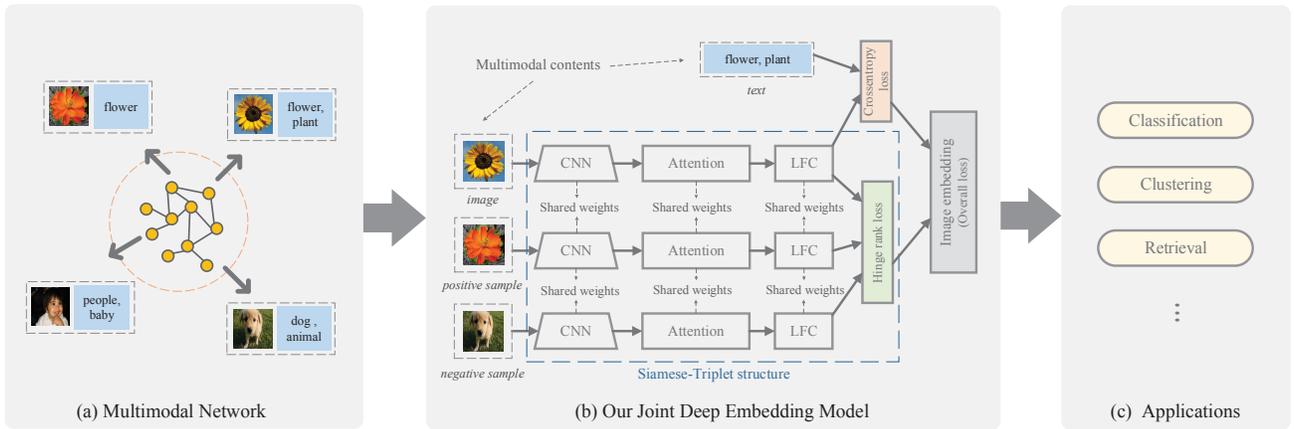


Figure 2: The framework of DMAN for social image embedding: (a) Multimodal network where each node contains two modalities of image and text. Each node is linked with others through some kind of relationship; (b) Our joint deep model to establish a joint representation of the multimodal network. The whole network is based on Siamese-Triplet structure which is composed of three identical subnetworks that share the same parameters. Specifically, in each subnetwork, we use CNN layers to extract image features and employ Attention layer and Locally Fully Connected (LFC) layers for visual-textual alignments. Hinge rank loss and Crossentropy loss are used to learn network information and multimodal contents respectively; (c) Various applications can be conducted on the learnt embedding.

media data, each data object usually contains multi-modal contents, and the internal relation between different data modalities makes it not suitable to consider them independently.

The content-based methods learn a joint representation by exploring the correlation between the multimodal contents for specific applications. Some of these methods leverage semantic information from unannotated text data to learn semantic relationships between labels, and explicitly map images into a rich semantic embedding space [7, 19]. [23] proposes a Gaussian visual-semantic embedding model for joint image-text representation, which leverages the visual information to model text concepts as Gaussian distributions in semantic space. To support user-centric applications, such as image recommendation and reranking, [16] proposes to learn image representation to capture both semantic labels and user intention labels. Meanwhile, there are also many works using the multi-modal learning method to learn the correlation between images and text to support various applications, such as image caption [13, 17, 30] and visual question answering [22, 32]. Canonical Correlation Analysis (CCA) [11] and its kernel version (KCCA) [11] are the popular methods used in many works [9, 10, 14], which finds a projection to maximize the correlation between the vectors projected from different views. Deep canonical correlation analysis (DCCA) [1, 31] is a deep model based CCA, which is applicable to high dimensional representation of image and text. However, these content based methods mainly learn the embedding to capture the correlation between different modalities of content, which cannot effectively explore the links among data to improve the embedding learning.

3 LEARNING SOCIAL IMAGE EMBEDDING

3.1 Problem Statement

Before the problem formulation, we define the notations used in the paper. In this paper, a set of social images is defined as a multi-modal network with each node containing multimodal contents and one or multiple types of links. As a mathematical abstraction, we define an undirected graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{\mathcal{V}_1, \dots, \mathcal{V}_N\}$ is a set of nodes and \mathcal{E} is a set of edges. An edge $\mathcal{E}_{ij}, \forall i, j \in \{1, 2, \dots, N\}$ belongs to the set \mathcal{E} if and only if an undirected link exists between \mathcal{V}_i and \mathcal{V}_j . For further simplification to comprehend, we assume two types of objects in the network: image(V) and text(T) and each node contains a pair of these two types of objects. Then $\forall i, \mathcal{V}_i = \{V_i, T_i\}$, where an image is presented as a squared tensor format as $V_i \in \mathbb{R}^{c \times h \times w}$ (c , h , and w denote the number of channels, the height, and the width of the image), and the text content is represented as $T_i \in \mathbb{R}^L$ (L denotes the size of the tag vocabulary).

Figure 2 illustrates the framework of our approach. In detail, to encode the linkage between social images, the Siamese-Triplet neural network is proposed to model the relationship among a triplet of images, i.e., a given image, a positive image which is a randomly sampled image linked to it, and a negative image which is a randomly sampled image with no link to it. The Siamese-Triplet neural network consists of three identical base networks which share the same parameters, with a hinge rank loss to learn the rank of the positive and negative images. To capture the fine-granularity relation between image regions and textual words, we propose a multimodal attention networks model, which assigns reasonable attention weights between the input words and the visual regions for a given social image. To combine the content and network for embedding, then a joint deep model is proposed to integrate the two components by simultaneously optimizing them. Since

the number of links in the network is exponential to the number of nodes, directly optimizing the object function by updating the whole network in each iteration will leads to an exponential complexity. Therefore, we propose a positive and negative image sampling method to decrease the complexity of training, which randomly sample a positive and K negative images for each image in the inference process, resulting in a linear complexity .

3.2 Siamese-Triplet Neural Network Model

Siamese-Triplet architecture is effective to model the network structure, which contains three identical subnetworks sharing the same configuration. Therefore, fewer parameters and less data are required to train it [8, 27, 28]. To capture the nonlinear structure of network, we propose a Siamese-Triplet neural network based on deep model for social image embedding as shown in Figure 3. First, we build a deep Convolution Neural Network (CNN) with an addition of several Full Connected (FC) layers as our base network to learn the features of every image. To encode the network information, we then build a network with Siamese-Triplet architecture over the base networks. Usually, a node is more similar to the linked nodes than to a random node. We use the Siamese-Triplet structure to capture the ranking information of the three nodes. Therefore, for a given image, we sample a positive image which has a link to it and a negative image which has no link to it. The three images compose the inputs of the Triplet network.

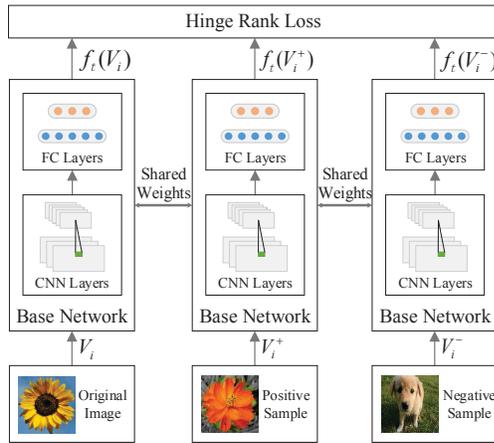


Figure 3: An illustration of the Siamese-Triplet network structure. The base networks consist of CNN and FC layers.

We use $f_t(\cdot)$ to denote the transformation of features. For each image V_i in the network, we sample a positive image V_i^+ to compose the positive pair, and we obtain its features from the final layer as $f_t(V_i)$. The similarity of the two images V_i and V_i^+ is defined as follows:

$$Sim(V_i, V_i^+; \theta_t) = \frac{f_t(V_i; \theta_t) \cdot f_t(V_i^+; \theta_t)}{\|f_t(V_i; \theta_t)\| + \|f_t(V_i^+; \theta_t)\|} \quad (1)$$

where θ_t is the parameters shared in the Siamese-Triplet network. Similarly, given the image V_i , we can sample a negative sample V_i^- , and the similarity $Sim(V_i, V_i^-; \theta_t)$ can be calculated like Eq. (1).

To encode the network structure information into feature representation $f_t(\cdot)$, it should be enforced that the similarity of a given image to the positive image is larger than to the negative image, i.e., $Sim(V_i, V_i^+) > Sim(V_i, V_i^-)$. The loss of the ranking is formulated by the hinge rank loss as follows:

$$\mathcal{L}_h(V_i, V_i^+, V_i^-; \theta_t) = \max[0, M - Sim(V_i, V_i^+; \theta_t) + Sim(V_i, V_i^-; \theta_t)] \quad (2)$$

where M denotes the gap parameter between two similarities. We empirically set $M = 0.3$ in the experiments. Then our objective function for training is formulated as follows:

$$\mathcal{L}_h(V; \theta_t) = \sum_{i=1}^N \max[0, M - Sim(V_i, V_i^+; \theta_t) + Sim(V_i, V_i^-; \theta_t)] \quad (3)$$

where N denotes the total number of the nodes. The L_2 normalization are replaced by dropout layers.

It is non-trivial to select the negative samples for learning to rank. We use the mini-batch Stochastic Gradient Descent (SGD) method to train the model. For each pair of V_i and V_i^+ , we randomly sample K negative matches in the same batch B , which obtains K triplets of samples. For each triplet of samples, the gradients over the three samples are computed respectively and the parameters are updated using the back propagation method. To ensure that the pair of patches V_i and V_i^+ can look up different negative matches each time, all the images are shuffled randomly after each iteration of training. For the experiments, we set $K = 3$.

3.3 Visual-Textual Attention Model

Attention is a mechanism which allows for an alignment of the input and output sequence [30], by which the salient features are allowed to dynamically come to the forefront as needed. Recently, it has been proven to be beneficial for many vision related tasks, such as image captioning [30] and image question answering [32]. Different from these works, our attention model is formulated in the Siamese-Triplet architecture for multi-modal data. We use the attention model to capture the alignment between different data modalities, which can utilize the network information to semantically learn the mapping between words and image regions based on the deep model of Siamese-Triplet network.

Given an image-text pair, our goal is to automatically find the relation between the words and the image regions. Let $T_i = \{t_i^0, t_i^1, \dots, t_i^k, \dots, t_i^L\}$, $T_i \in \mathbb{R}^L$ denotes the text features of the i th pair, which is a one-hot words vector expression with length L , where k denotes for the index of the word. Let V_i denotes the raw image corresponded to T_i . We use the deep Convolution Neural Networks (CNN) to obtain the image region maps $R_i = \{r_{i,0}, r_{i,1}, \dots, r_{i,j}, \dots, r_{i,D}\} \in \mathbb{R}^{D \times M}$ for V_i as follows:

$$R_i = f_c(V_i; \theta_c), R_i \in \mathbb{R}^{D \times M} \quad (4)$$

where θ_c is the parameters of the CNN layers, j denotes the index of the region, D is the dimension of image region, and M is the dimension of the map.

In the attention model, a value between 0 and 1 is assigned to each image region $r_{i,j}$ based on its relevance to the word t_i^k . Formally, we aim to automatically generate the image attention

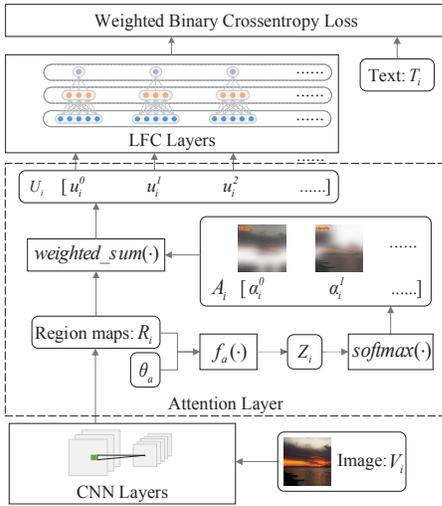


Figure 4: The architecture of the visual-textual attention model. A rectangle block stands for a process or a function, while the rounded box stands for inputs, parameters or tensors.

values for the words as follows:

$$Z_i = f_a(R_i; \theta_a), Z_i \in \mathbb{R}^{L \times D} \quad (5)$$

where Z_i denotes the unnormalized word attention values for the region maps R_i of the i th pair. Following [30, 33], Z_i is spatially normalized with the softmax function to obtain the final word attention maps A_i ,

$$\alpha_{i,j}^k = \frac{\exp(z_{i,j}^k)}{\sum_j \exp(z_{i,j}^k)}, A_i \in \mathbb{R}^{L \times D} \quad (6)$$

where $z_{i,j}^k$ and $\alpha_{i,j}^k$ represent the unnormalized and normalized attention values at region j in image i for word k respectively. If word k is assigned to the input image, higher attention values should be given to the image regions related to it. The attention estimator $f_a(\cdot)$ can be calculated in many ways such as CNN in [33]. In this paper, it is modelled as a sequentially distributed full connected layer and computed as follows:

$$Z_i = \tanh(wR_i^T + b), Z_i \in \mathbb{R}^{L \times D} \quad (7)$$

where $w \in \mathbb{R}^{L \times M}$ and $b \in \mathbb{R}^L$ that compose the parameter set θ_a of the attention model in Eq. (5) and will be updated through back propagation. The tanh activation are used to make the model nonlinear.

Let $r_{i,j} \in \mathbb{R}^M$ denotes the visual feature vector of region j in R_i . The normalized attention value is used as the weight to sum the features of the regions for each word k to obtain the output features as follows:

$$u_i^k = \sum_j \alpha_{i,j}^k r_{i,j}, U_i \in \mathbb{R}^{L \times M} \quad (8)$$

The architecture of the attention model is illustrated in Figure 4. The equation above behaves somewhat like a weighted average

pooling layer for each word. Compared with the original independent visual features shared by all words, the weighted visual feature mapping u_i^k is more effective to reflect the image regions related to word k . The dimensionality of U_i is $\mathbb{R}^{L \times M}$, while $T_i \in \mathbb{R}^L$. In order to make comparison between the visual output and text features, we stack several Locally Fully Connected (LFC) layers to obtain a L -dimensionality output of visual features. The LFC layers are locally fully connected for each word, and the parameter sets corresponding to different words are independent. That is, u_i^k is only related to the correspondingly attended t_i for each word k . Note that the last LFC layer has just one neuron for each word in the vocabulary, which sets the dimensionality of the final output equal to L . The activation sigmoid is used in the last LFC layer to normalise the feature representation for estimating words' confidence by comparing with the ground truth text vector. Let $Y_i \in \mathbb{R}^L$ be the final output of LFC layers:

$$Y_i = f_l(U_i; \theta_l), Y_i \in \mathbb{R}^L \quad (9)$$

We make a pipeline of the three functions mentioned above to obtain a whole process from image input to Y_i as:

$$Y_i = f_w(V_i; \theta_w), Y_i \in \mathbb{R}^L \quad (10)$$

where f_w is the pipeline of f_c , f_a and f_l , and θ_w is the set of θ_c , θ_a and θ_l .

The parameters θ_w are learned by minimizing the weighted binary crossentropy loss between Y_i and T_i ,

$$\begin{aligned} \mathcal{L}_c(V, T; \theta_w) &= \sum_{i=1}^N -(\lambda T_i \cdot \log(Y_i)) + (1 - T_i) \cdot \log(1 - T_i) \\ &= \sum_{i=1}^N -(\lambda T_i \cdot \log(f_w(V_i; \theta_w))) + (1 - T_i) \cdot \log(1 - f_w(V_i; \theta_w)) \end{aligned} \quad (11)$$

where N is the whole number of the pairs and λ is a balance parameter. Since the number of zero elements in T_i is much more than non-zero ones, it is reasonable to punish the false negatives more.

3.4 A Joint Deep Embedding Model

The Siamese-Triplet neural network learns the embedding by exploiting the network structure information, and the visual-textual attention model exploits the fine-granularity relation between data modalities for embedding learning. Intuitively, we propose a joint deep embedding model to combine the two components, which simultaneously optimizes them. Specifically, we change the FC layers in the base network of Siamese-Triplet model to an attention layer and several LFC layers. Then, we formulate the loss function as the summation of the hinge rank loss Eq. (3) and the weighted binary crossentropy loss Eq. (11) as follows:

$$\mathcal{L}(V, T; \theta_w) = \mathcal{L}_h(V; \theta_w) + \beta \mathcal{L}_c(V, T; \theta_w) \quad (12)$$

where β is a weight parameter. Since the Triplet network are added with the attention in the joint model, the parameter θ_t are replaced by θ_w which is shared in the whole model.

The computational complexity of the joint deep model is reduced greatly by using the positive sampling and negative smapling method. Assume the whole number of image-text pairs is N , the method in [3] learns the network representation by iterating the

whole network. It results in a computational complexity of $O(k(N \times N))$, where k is the number of iterations. Our methods only sample several nodes for each node for parameters updating in each iteration. Thus the computational complexity is reduced to $O(k(N))$.

4 EXPERIMENTS

In this section, we conduct a set of experiments to analyse the effectiveness of the embedding model DMAN by evaluating its performance in two tasks, i.e., multi-label image classification and cross-modal search.

4.1 Experimental Settings

The experiments are conducted on three popular datasets collected from flickr, which have the groundtruth labels provided by human annotators. Based on the study of [18] on these collections, we crawl the original images from the Flickr website. The details of these image collections are described below:

- The NUS Web Image Database (NUS-WIDE) dataset [4] is a web image dataset which contains 269,648 images. Among these images, 226,912 are available in the Flickr sources.
- The MIR Flickr Retrieval Evaluation (MIR) dataset contains one million images, but only 25,000 of them have been annotated [12]. 13,368 of the annotated images are available in the Flickr sources.
- The PASCAL Visual Object Classes Challenge (PASCAL) datasets [6] contains 9963 images. 9,474 of the annotated images are available in Flickr sources.

We preprocess these datasets as [3]. First, since there are many noisy images that do not belong to any of their groundtruth labels, we remove these samples. After that, we use the most frequent 1,000 tags as our text vocabulary and construct a 1000-D 0-1 vector for the text content. We further remove those image-text pairs that do not contain any words in the vocabulary. Finally, we randomly sample the image-text pairs for training and testing with the ratio of 4:1. We construct a network by treating each image-text pair as a node, and establish an edge between two nodes once they share at least one label. For each node, at most 50 links are random sampled to construct the sparse adjacency matrix. We evaluate our framework in an out-of-sample strategy. The final statistics of these datasets are shown in Table 1. Note that 90000 nodes of NUS-WIDE are randomly sampled with 53,844 for training and 36,352 for testing as [3] for fair comparison.

Table 1: Datasets statistics.

	NUS	MIR	PASCAL
#image	90,000	5,040	6,509
#tag	1,000	1,000	1,000
#tag per image	6.36	4.09	3.94
#label	81	14	20
#label per image	2.46	1.81	1.95
#node (training)	53,844	4,032	5,207
#edge (training)	2,682,005	216,318	284,961
#node (testing)	36,156	1,008	1,302

In the experiments, we take the image with size 224×224 and channel RGB as visual input, and CNNs is used for visual feature extraction. Specifically, our CNN layers employ the VGG16 network [24] pretrained on ImageNet 2012 classification challenge dataset [5] with Keras deep learning framework. Then we use the pool5's outputs as the visual features for image region maps, with size 49×512 . We stack three LFC layers to the attention layer, with the dimensions of 1000×128 , 1000×32 and 1000×1 respectively. As for the hyper-parameters λ and β , we set the values with 10 and 1 that can obtain a relatively good performance. In the training procedure, SGD is set with learning rate 0.01, momentum 0.9 and nesterov=True. All of the implementation are trained on $2 \times$ NVIDIA GTX 1080. All the source codes of our models will be released upon the publication of this work.

4.2 Baselines

We evaluate the performance of DMAN by comparing it with state-of-the-art approaches introduced below:

- **CCA** [11]: The Canonical Correlation Analysis embeds two types of input data into a common latent space by optimizing an objective function with respect to their correlations.
- **DT** [21]: A transfer learning method is proposed to bridge the semantic distances between image and text using latent embeddings.
- **LHNE** [3]: The linear version of Heterogeneous Network Embedding (HNE) [3].
- **HNE** [3]: Heterogeneous Network Embedding via Deep Architectures.
- **KCCA** [11]: The kernel version of the Canonical Correlation Analysis.
- **DCCA** [31]: An image-text matching approach based on deep canonical correlation analysis.
- **DMAN_{Triplet}**: Only the image is used to construct the triplet neural network and the representation is learned from the network directly. It is used to evaluate the effectiveness of the triplet network model for embedding
- **DMAN_{Triplet+Text}**: The text content is added by DMAN_{Triplet}, where the text content is combined using a full connected network rather than the attention network.

Among, the first 4 methods are introduced in [3] and we will compare them with our model on the dataset NUS-WIDE for the two tasks as in their papers.

4.3 Multi-Label Classification

All the datasets are multi-labeled with unbalanced distribution over the classes. A comprehensive introduction of evaluation metrics for multi-label classification is presented in [29]. We employ macro/micro precision, macro/micro recall, macro/micro F1-measure, and Mean Average Precision (mAP) for performance evaluation. If the predicted label confidence for any label is greater than 0.5, the label is considered as positive. To ensure fair comparison, we use the neural network with 3 FC layers to learn a common classifier. After accomplishing the training process, we use the trained model to acquire the embeddings of the test set. Then we

Table 2: Multi-label classification results

Dataset	Model	Micro-P	Micro-R	Micro-F1	Macro-P	Macro-R	Macro-F1	mAP
NUS-WIDE	CCA [11]	-	-	-	-	-	-	52.54%
	DT [21]	-	-	-	-	-	-	53.22%
	LHNE [3]	-	-	-	-	-	-	53.32%
	HNE [3]	-	-	-	-	-	-	54.99%
	KCCA [11]	75.65%	52.96%	62.30%	59.12%	46.58%	52.11%	52.91%
	DCCA [31]	75.79%	54.48%	63.39%	61.21%	50.16%	55.14%	54.36%
	DMAN _{Triplet}	76.49%	54.30%	63.51%	59.69%	48.88%	53.75%	53.35%
	DMAN _{Triplet+Text}	77.00%	55.15%	64.26%	59.62%	51.40%	55.20%	54.84%
DMAN	77.25%	56.37%	65.18%	62.64%	52.62%	57.19%	57.52%	
MIR	KCCA [11]	73.69%	72.55%	73.12%	72.41%	70.86%	71.63%	69.53%
	DCCA [31]	77.21%	76.65%	76.93%	75.37%	70.11%	72.64%	73.13%
	DMAN _{Triplet}	76.51%	71.75%	74.58%	73.58%	72.67%	72.74%	72.05%
	DMAN _{Triplet+Text}	78.14%	77.66%	77.90%	73.14%	71.08%	72.09%	74.40%
	DMAN	79.14%	77.45%	78.28%	75.12%	72.11%	73.65%	75.19%
PASCAL	KCCA [11]	78.55%	31.89%	45.36%	49.25%	35.77%	41.44%	43.59%
	DCCA [31]	80.09%	35.27%	48.97%	53.91%	43.86%	48.37%	47.25%
	DMAN _{Triplet}	83.87%	30.32%	44.54%	56.16%	38.84%	45.92%	47.41%
	DMAN _{Triplet+Text}	79.19%	38.30%	51.63%	55.75%	41.12%	47.33%	49.54%
	DMAN	86.88%	44.14%	58.53%	62.23%	46.85%	53.45%	54.76%

Table 3: Cross modal retrieval results

Dataset	Model	Rank 1	Rank 5	Rank 10	Rank 20	Rank 50
NUS-WIDE	CCA [11]	21.05%	16.84%	18.95%	18.68%	-
	DT [21]	20.53%	25.26%	22.63%	22.37%	-
	LHNE [3]	26.32%	21.05%	21.02%	22.27%	-
	HNE [3]	36.84%	29.47%	27.89%	26.32%	-
	KCCA [11]	26.30%	27.69%	22.57%	20.36%	18.35%
	DCCA [31]	35.99%	31.12%	29.15%	24.48%	25.32%
	DMAN _{Triplet+Text}	36.16%	28.57%	28.57%	27.92%	23.11%
	DMAN	38.96%	40.77%	40.91%	38.96%	37.14%
MIR	KCCA [11]	50.83%	43.53%	37.60%	35.41%	30.50%
	DCCA [31]	59.67%	54.12%	51.81%	41.89%	38.65%
	DMAN _{Triplet+Text}	64.50%	53.47%	54.16%	52.18%	40.33%
	DMAN	66.67%	55.00%	53.50%	52.18%	43.67%
PASCAL	KCCA [11]	27.35%	28.06%	30.15%	25.12%	26.33%
	DCCA [31]	42.12%	45.14%	44.96%	42.87%	41.89%
	DMAN _{Triplet+Text}	41.18%	40.00%	41.76%	43.20%	38.35%
	DMAN	47.05%	55.29%	57.64%	56.17%	46.23%

use the FC classifier mentioned above to train and test on each dataset.

Experimental results on the 3 datasets are shown in Table 2. It shows that DMAN outperforms all of state-of-the-art models. First, from the results of NUS-WIDE, it can be concluded that the performance of DMAN_{Triplet} is better than that of CCA, DT and LHNE on the metric of mAP, which validate the effectiveness of embedding using the triplet network model. By combining the text content, DMAN_{Triplet+Text} almost reaches the score of HNE

on the metric of mAP and transcend DMAN_{Triplet} on all metrics, which confirms the importance of combining the multi-modal content for embedding learning. With the attention model, DMAN boosts mAP to 57.52% compared with 54.99% of HNE and makes an improvement on every metric comparing to DMAN_{Triplet+Text}. This is because that the attention model makes alignments between the multi-modal contents, which is useful to learn a more efficient representation of the multi-modal data. On the other side, HNE learns the features for image and text document independently,

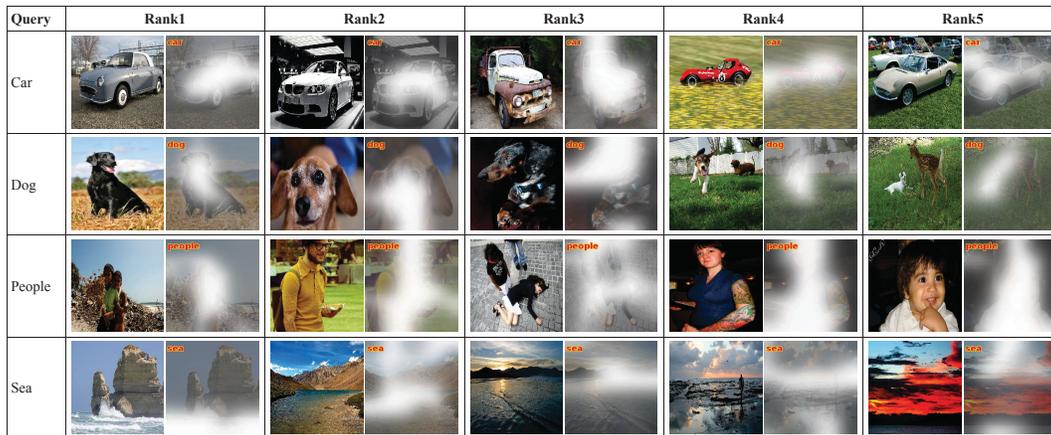


Figure 5: Some results of cross-modal search and the visualization of learned attentions

which is less effective to capture the correlation between different data modalities. Meanwhile, the other baselines also can not effectively exploiting the link information and the fine-granularity relations between different data modalities. The quantity of tags in PASCAL is less than those in other datasets, but the quality is better. Therefore, the improvement of $DMAN_{Triplet+Text}$ and $DMAN$ on PASCAL is greater than it on other datasets.

4.4 Cross-Modal Search

To further demonstrate the superiority of $DMAN$, we compare it with the baselines in the task of cross-modal search as [3]. In the datasets NUS-WIDE, MIR and PASCAL, there are 77/81, 12/14 and 17/20 of the groundtruth label words appearing in the text vector respectively. We manually construct 77, 12, and 17 query vectors with the dimensionality of 1000 for the three datasets respectively, by setting the corresponding label entry to one and the remaining entries to zero. Using the learned embedding function, we project the query vector to the latent space to retrieve all the image samples in the test set using the standard Euclidean distance. The average precision at rank k ($p@k$) over all queries is reported in Table 3. On the dataset of NUS-WIDE, $DMAN$ achieves about 10% higher AP compared to HNE, and far surpass to CCA, DT, and LHNE. On all the three datasets, $DMAN$ outperforms KCCA and DCCA greatly. It demonstrates the effectiveness of our model for cross-modal search. Meanwhile, the link information is also helpful to find the most similar images, which affects the performance of the methods that ignore the link information. For all of the datasets, we can find a substantial improvement of $DMAN$ compared to $DMAN_{Triplet+Text}$, which provides evidence that the attention model is effective for cross-modal search.

Figure 5 gives some samples of search results of MIR. For each query, we present the top-5 ranked images and their corresponding regions aligned by the attention model. For the query “Sea”, the aligned images have mismatched attention on white clouds and blue sky, which is because that the tags of “sea”, “cloud” and “sky” frequently co-occur in the same images. For the other queries, our model draws the right attentions on the images and hence improves the performance.

5 CONCLUSIONS

In this paper, we explore learning social image embedding to capture both multimodal contents and network information for social media data applications, such as multi-label classification and cross-modal search. A Deep Multimodal Attention Networks ($DMAN$) embedding model is proposed, in which a Siamese-Triplet neural network is designed to embed network information and the Visual-Textual Attention Model is proposed to capture the correlation between different data modalities. Then, a deep model is proposed to combine network information and content for embedding learning, in which the loss function is designed to simultaneously optimizing the Siamese-Triplet neural network and Visual-Textual Attention Model. The experiment results in multilabel classification and cross-modal search applications indicate that capturing the network structure and the correlation between different data modalities are helpful for social image embedding. In all of the experiments, our approach outperforms state-of-the-art baselines in various evaluations.

This work is an effort to combine network structure for multimodal data embedding. It is different from current networking embedding researches that mainly explore the link information and can not effectively exploit the multi-modal content within each node. It is also different from the current image embedding methods that ignore the link information and are mainly task dependent, such as image caption and classification. In the future, we will explore to better model the social link information and design a more reasonable deep model to make the learned embedding more effective. Furthermore, we can incorporate the strength of relationship between images to better model the network.

6 ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (Grand Nos. U1636211, 61672081, 61602237, 61370126), National High Technology Research and Development Program of China (No.2015AA016004) and Fund of the State Key Laboratory of Software Development Environment (No. SKLSDE-2017ZX-19).

REFERENCES

- [1] Galen Andrew, Raman Arora, Jeff A. Bilmes, and Karen Livescu. 2013. Deep Canonical Correlation Analysis. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013 (JMLR Workshop and Conference Proceedings)*, Vol. 28. JMLR.org, 1247–1255. <http://jmlr.org/proceedings/papers/v28/andrew13.html>
- [2] Shaosheng Cao, Wei Lu, and Qionghai Xu. 2015. Grarep: Learning graph representations with global structural information. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*. ACM, 891–900.
- [3] Shiyu Chang, Wei Han, Jiliang Tang, Guo-Jun Qi, Charu C Aggarwal, and Thomas S Huang. 2015. Heterogeneous network embedding via deep architectures. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 119–128.
- [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 2009, Santorini Island, Greece, July 8-10, 2009*, Stéphane Marchand-Maillet and Yiannis Kompatsiaris (Eds.). ACM. DOI: <https://doi.org/10.1145/1646396.1646452>
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, 248–255. DOI: <https://doi.org/10.1109/CVPRW.2009.5206848>
- [6] Mark Everingham, Luc J. Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* 88, 2 (2010), 303–338. DOI: <https://doi.org/10.1007/s11263-009-0275-4>
- [7] 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. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, 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>
- [8] Vijay Kumar B. G, Gustavo Carneiro, and Ian D. Reid. 2016. Learning Local Image Descriptors with Deep Siamese and Triplet Convolutional Networks by Minimizing Global Loss Functions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 5385–5394. DOI: <https://doi.org/10.1109/CVPR.2016.581>
- [9] Yunhao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. 2014. A Multi-View Embedding Space for Modeling Internet Images, Tags, and Their Semantics. *International Journal of Computer Vision* 106, 2 (2014), 210–233. DOI: <https://doi.org/10.1007/s11263-013-0658-4>
- [10] Yunhao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014. Improving Image-Sentence Embeddings Using Large Weakly Annotated Photo Collections. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV (Lecture Notes in Computer Science)*, David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.), Vol. 8692. Springer, 529–545. DOI: <https://doi.org/10.1007/978-3-319-10593-2.35>
- [11] David R. Hardoon, Sándor Szedmak, and John Shawe-Taylor. 2004. Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Computation* 16, 12 (2004), 2639–2664. DOI: <https://doi.org/10.1162/0899766042321814>
- [12] Mark J Huiskes and Michael S Lew. 2008. The MIR flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. ACM, 39–43.
- [13] Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 3128–3137. DOI: <https://doi.org/10.1109/CVPR.2015.7298932>
- [14] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2014. Fisher Vectors Derived from Hybrid Gaussian-Laplacian Mixture Models for Image Annotation. *CoRR abs/1411.7399* (2014). <http://arxiv.org/abs/1411.7399>
- [15] Chaozhuo Li, Senzhang Wang, Dejian Yang, Zhoujun Li, Yang Yang, Xiaoming Zhang, and Jianshe Zhou. 2017. PPNE: Property Preserving Network Embedding. In *Database Systems for Advanced Applications - 22nd International Conference, DASFAA 2017, Suzhou, China, March 27-30, 2017, Proceedings, Part I (Lecture Notes in Computer Science)*, K. Selçuk Candan, Lei Chen, Torben Bach Pedersen, Lijun Chang, and Wen Hua (Eds.), Vol. 10177. Springer, 163–179. DOI: <https://doi.org/10.1007/978-3-319-55753-3.11>
- [16] 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, Brisbane, Australia, October 26 - 30, 2015*, Xiaofang Zhou, Alan F. Smeaton, Qi Tian, Dick C. A. Bulterman, Heng Tao Shen, Ketan Mayer-Patel, and Shuicheng Yan (Eds.). ACM, 109–118. DOI: <https://doi.org/10.1145/2733373.2806247>
- [17] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. 2014. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). *CoRR abs/1412.6632* (2014). <http://arxiv.org/abs/1412.6632>
- [18] Julian J. McAuley and Jure Leskovec. 2012. Image Labeling on a Network: Using Social-Network Metadata for Image Classification. In *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV (Lecture Notes in Computer Science)*, Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid (Eds.), Vol. 7575. Springer, 828–841. DOI: <https://doi.org/10.1007/978-3-642-33765-9.59>
- [19] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg Corrado, and Jeffrey Dean. 2013. Zero-Shot Learning by Convex Combination of Semantic Embeddings. *CoRR abs/1312.5650* (2013). <http://arxiv.org/abs/1312.5650>
- [20] 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.
- [21] Guo-Jun Qi, Charu C. Aggarwal, and Thomas S. Huang. 2012. Transfer Learning of Distance Metrics by Cross-Domain Metric Sampling across Heterogeneous Spaces. In *Proceedings of the Twelfth SIAM International Conference on Data Mining, Anaheim, California, USA, April 26-28, 2012*. SIAM / Omnipress, 528–539. DOI: <https://doi.org/10.1137/1.9781611972825.46>
- [22] Mengye Ren, Ryan Kiros, and Richard S. Zemel. 2015. Exploring Models and Data for Image Question Answering. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (Eds.), 2953–2961. <http://papers.nips.cc/paper/5640-exploring-models-and-data-for-image-question-answering>
- [23] Zhou Ren, Hailin Jin, Zhe L. Lin, Chen Fang, and Alan L. Yuille. 2016. Joint Image-Text Representation by Gaussian Visual-Semantic Embedding. In *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*, Alan Hanjalic, Cees Snoek, Marcel Worring, Dick C. A. Bulterman, Benoit Huet, Aisling Kelliher, Yiannis Kompatsiaris, and Jin Li (Eds.). ACM, 207–211. DOI: <https://doi.org/10.1145/2964284.2967212>
- [24] 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>
- [25] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-scale Information Network Embedding. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, Aldo Gangemi, Stefano Leonardi, and Alessandro Panconesi (Eds.). ACM, 1067–1077. DOI: <https://doi.org/10.1145/2736277.2741093>
- [26] Daixin Wang, Peng Cui, and Wenwu Zhu. 2016. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1225–1234.
- [27] Xiaolong Wang and Abhinav Gupta. 2015. Unsupervised Learning of Visual Representations Using Videos. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2794–2802. DOI: <https://doi.org/10.1109/ICCV.2015.320>
- [28] Paul Wohlhart and Vincent Lepetit. 2015. Learning descriptors for object recognition and 3D pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 3109–3118. DOI: <https://doi.org/10.1109/CVPR.2015.7298930>
- [29] Xi-Zhu Wu and Zhi-Hua Zhou. 2016. A Unified View of Multi-Label Performance Measures. *CoRR abs/1609.00288* (2016). <http://arxiv.org/abs/1609.00288>
- [30] 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 2015, Lille, France, 6-11 July 2015 (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>
- [31] Fei Yan and Krystian Mikolajczyk. 2015. Deep correlation for matching images and text. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 3441–3450. DOI: <https://doi.org/10.1109/CVPR.2015.7298966>
- [32] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2016. Stacked Attention Networks for Image Question Answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 21–29. DOI: <https://doi.org/10.1109/CVPR.2016.10>
- [33] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. 2017. Learning Spatial Regularization with Image-level Supervisions for Multi-label Image Classification. *CoRR abs/1702.05891* (2017). <http://arxiv.org/abs/1702.05891>