Hindawi
Complexity
Volume 2020, Article ID 1543947, 11 pages
https://doi.org/10.1155/2020/1543947

Research Article

Image-Text Joint Learning for Social Images with Spatial Relation Model

Jiangfan Feng, Xuejun Fu, Yao Zhou, Yuling Zhu, and Xiaobo Luo

Chongqing University of Posts and Telecommunications, College of Computer Science and Technology,
Space Big Data Intelligent Technology Chongqing Engineering Research Center, Chongqing 400065, China

Correspondence should be addressed to Jiangfan Feng; fengjf@cqupt.edu.cn

Received 27 September 2019; Revised 30 December 2019; Accepted 11 February 2020; Published 28 March 2020

Academic Editor: Mahardhika Pratama

Copyright © 2020 Jiangfan Feng et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The rapid developments in sensor technology and mobile devices bring a flourish of social images, and large-scale social images have attracted increasing attention to researchers. Existing approaches generally rely on recognizing object instances individually with geo-tags, visual patterns, etc. However, the social image represents a web of interconnected relations; these relations between entities carry semantic meaning and help a viewer differentiate between instances of a substance. This article forms the perspective of the spatial relationship to exploring the joint learning of social images. Precisely, the model consists of three parts: (a) a module for deep semantic understanding of images based on residual network (ResNet); (b) a deep semantic analysis module of text beyond traditional word bag methods; (c) a joint reasoning module from which the text weights obtained using image features on self-attention and a novel tree-based clustering algorithm. The experimental results demonstrate the effectiveness of using Flickr30k and Microsoft COCO datasets. Meanwhile, our method considers spatial relations while matching.

1. Introduction

With the rise of cheap sensors, mobile terminals, and social networks, research on social images is making good progress, including image retrieval, object classification, and scene understanding. Compared with images in traditional applications, it is hard to understand social pictures using the low-level features. Meanwhile, most of the existing methods only capture the local patterns of images by utilizing low-level features (e.g., color and texture). Intuitively, knowing the spatial relation among local elements may help predict what objects and scenes are presented in the visual content. It has recently been widely adopted in the vision community that contextual information, i.e., the relation between objects, improves the accuracy of object recognition [1]. Therefore, the geometry relation of objects in social images is usually exploited to conduct annotation, which depends on the similarity measurement of visual objects.

Significant efforts have been taken to integrate visual and textual analyses [2–4]. For example, Wang et al. [5] present an algorithm to learn the relations between scenes, objects, and texts with the help of image-level labels. However, such a training process requires a large number of paired images and text data. Motivated by the success of the encoder-decoder network, studies have been proposed to apply it to generate text descriptions of the given images [6, 7]. Nevertheless, such impressive performance relies on the assumption that the training data and the test data should come from the same underlying data distribution. Some approaches [8, 9] exploit the spatial relations of objects indicated by the prepositions for image understanding. They suffer from the limitations that spatial relations have to be learned with the bounding boxes of objects and cannot be driven by the task goal.

Although there exist several successful image-text learning approaches or vision-based approaches to analyze social images, the following problems are still not addressed:

(1) Visual content and text are always separately learned, making the traditional methods hard to be trained end-to-end.

(2) Learning tasks converted to classification problems, empowered by large-scale annotated data with end-

to-end training using neural networks, which is not capable of describing concepts unseen in the training pairs.

(3) The spatial relations defined by prepositions have to be learned with the bounding boxes of objects, which are so immoderately challenging to obtain. Moreover, the spatial relationships from the textual descriptions are very scarce in reality.

Motivated by these observations, we aim at developing a method to learn the spatial relations across separate visual objects and texts for social image understanding. Therefore, this paper proposes a cross-modal framework, which builds a joint model of texts and images to extract features and combine the advantages of self-attention mechanism and deep learning models, generating interactive effects. In particular, we investigate (1) how to label social images with high-level features based on their political image position and (2) how to combine the text and visual content. The framework is established by taking spatial relationships as a basic unit of image-text joint learning. We use neural architectures to measure the semantic similarity between visual data, e.g., images or regions, and text data, e.g., sentences or phrases. Learning this similarity requires connecting low-level pixel values and high-level language descriptions and then, a joint latent space of standard dimensionality in which matching image and text features has high cosine similarity, to explore the semantics hidden inside a social image.

(1) We propose a framework that jointly trains two dual tasks: the spatial semantic of image and text-to-image synthesis, which improves the supervision and the generalization performance of social image caption.

(2) We extend the conventional model by adding the top-down attention mechanism. With a novel tree-based clustering method, it can demonstrate the effectiveness in learning the alignments of visual concepts of images and the semantics of texts.

2. Related Works

The related works generally fall into two categories: image tagging and relational inference methods.

2.1. Image Tagging. Image tagging has been widely studied for the past decade. Early image tagging models are built mainly from the view of statistics and probability. In practice, many image annotation works [10, 11] assign top-k class labels to each image, the quantities of class labels in different photos vary significantly, and the top-k annotations degrade the performance of image annotation. Besides, many works attempt to infer correlations or joint probability distributions between images and semantic concepts (or keywords). For example, Farhadi et al. [12] use detection methods to understand scene elements. Similarly, Li et al. [13] start with exposures and piece a final description using phrases containing detected objects and relationships. Further, powerful language models based on language parsing have been used as well [14]. Recently, deep learning has achieved great success in the field of image, text, and speech, for example, the m-RNN model [15] in which a multimodal component is introduced to explicitly connect the language model and the vision model by a one-layer representation.

Images can label at the pixel level, which has applications in intelligent video monitor, and self-driving cars, etc. More recently, the variable label number problem has been identified [16, 17]. These solutions treat the image annotation problem as an image-to-text translation problem and solve it using an encoder-decoder model. The multiscale approaches [18] propose a novel multiscale deep model for extracting rich and discriminative features capable of representing a wide range of visual concepts. Instead of CNN features, some works use more semantic information obtained from the image as the input to the decoder [19, 20]. In all, most methods still focus on recognizing objects separately. The spatial relationships between objects and scenes are always neglected.

2.2. Relational Inference. The earliest reasoning form can date to a symbolic way, and the ties between symbols are established in logical and mathematical language, which is interpreted in terms of deduction, arithmetic, and algebra. As symbolic approaches suffer from the symbol grounding problem, they are not robust to small tasks and input variations [21, 22]. Many other methods, such as deep learning, in a traditional inference network, the inference part may be multilayer Perceptron (MLP), Long Short-Term Memory (LSTM), or LSTM with attention, which often runs into the problem of weak data [23, 24]. Santoro proposed a relation network to achieve the reasoning part. This structure clearly expresses two ideas: the final answer has to do with pairs of objects, and the problem itself will influence how to examine the objects. As an active research field, some recent works have also applied neural networks to structured graph data or tried to standardize network output through relations and knowledge bases. We believe that visual data reasoning should be both local and global: abandoning the two-dimensional image structure to involve the task is not only useful but also invalid. Beyond the object detection task that detects relationships for an image, the scene graph generation task [25, 26] endows the model with an entire structured representation capturing both objects and their semantic relationships, where the nodes are object instances in images, and the edges depict their pairwise correlations. The proposed methods usually require additional annotations of relations, while they demand only image-level annotations.

3. Cross-Modal Reasoning Framework

3.1. Overview of Cross-Modal Tasks and Pipeline. In this paper, we focus on two image-text tasks: spatial relation modeling and image-text matching. The former refers both to image-to-image and image-to-text, and definitions of the two scenarios are straightforward: given an input image (resp. sentence), the goal is to find the relationships between entities with semantic meaning. The second task refers to find the best matching sentences to the input images.

The architecture for the above tasks should consist of four-step pipeline, as summarized here: The First step is to design functional feature extractors, we use a refined TF-IDF method [27] to calculate the word frequency and then combine it with the embedding vector, which helps us to map it nonlinearly into another vector space to enhance semantic association in words and lower its dimension. The second step is to generate full image features and geographical features. We can deepen the network continuously, getting saturated then degrades rapidly [28]. The third step is to add a self-attention mechanism, using the image feature to get the weight of words. Then we combine the image and text features together to deduce the spatial relation of the picture. The final step performs instance recognition, which is a deep semantic understanding of social images.

At a conceptual level, there are two branches to achieve the goal. One is to train the network to map images and texts into joint embedding space. The second approach is to frame pictures and documents correspondence by cosine similarity, to obtain the probability that the two items match. Accordingly, we define a cross-modal reasoning framework to exploit image-text joint learning of social image retrieval with spatial relation model, which includes two variants of two-branch networks that follow these two strategies (Figure 1): the embedding network and the similarity network.

Two networks are needed in this framework, one for cross-modal topic detection and the other for semantic matching, which is trained one by one. The embedding network refers to cross-domain spatial relation modeling, as illustrated in Figure 1(a). The spatial relation includes both image-to-text and image-to-image, and the goal is to model spatial relationships in CNN based detection. The similarity network is illustrated in Figure 1(b), we first pretrain image data and text-image data, respectively, and then fine-tuned on the target domain training data via reinforcement learning. In detail, we use ResNet-50 as the CNN encoder of the framework to learn the social image annotations. By adding the cross-domain spatial relation model, the structure can attend the crucial parts of the image when decoding different words of the caption to generate interactive effects. Joint models with image-text similarity are given by cosine similarity [29].

3.2. Cross-Domain Spatial Relation Modeling.

The spatial relation between visual content and texts is presented as two levels. First, it represented by the matching probability of the purpose and second by the spatial relationships between the image objects represented by the interaction of the objects in different tasks.

3.2.1. Spatial Relation between Images and Texts.

Motivated by attention-based deep methods [30, 31], we first review a basic attention module called self-attention. As illustrated in Figure 2, the input consists of queries and keys of dimensions d_k and d_v. The dot product is performed between the question and all keys to obtain their similarity. A softmax function is applied to obtain the weights on the values. Given keys and values, the output value is the weighted average over input values:

$$\begin{gathered}\text{Attention}\,(Q, K, V) = \text{soft}\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V,\\ \text{Attention}\,(\text{Query}, \text{Source}) = \sum_{i=1}^{L_x} \text{Similarity}\,(\text{Query}, \text{Key}_i) * \text{Value}_i.\end{gathered} \tag{1}$$

In our case, we apply modifications to the output values. Specially, we define attention as follows: the source consists of a set of two-dimensional vectors < key, value > , given an element of a target named query, calculating the similarity or relevance of the question and each to get a weight coefficient of every key corresponding to the value. The weighted sum is performed by

$$\begin{aligned}\text{Attention}\,(\text{Query}, \text{Source}) = \textstyle\sum_{i=1}^{L_x} \text{Similarity}\,(\text{Query}, \text{Key}_\text{i})\\ \cdot\, \text{Value}_i\end{aligned}, \tag{2}$$

where x is the vector of an input image or word, L_x is the length of x, Query is the image vector, key$_i$is the word vector, and the dot product is performed to calculate similarity.

We apply the perceptron to calculate the weight of the word vector. In self-attention, each word can compute all terms with attention. The aim is to learn the internal word dependence and to capture the internal structure of the text. The characteristics of self-attention lie in ignoring the distance of sentence to an image, directly calculating its inner structure, to study the internal structure of a sentence. Further, the realization of parallel computing can also be relatively simple.

Inspired by the thought of learning to rank [32, 33], we measure the similarity between two samples using cosine distance.

$$D\left(f_{x_i}, f_{x_j}\right) = \frac{f_{x_i}}{\left\|f_{x_i}\right\|_2} \cdot \frac{f_{x_j}}{\left\|f_{x_j}\right\|_2}, \tag{3}$$

where$\|\cdot\|_2$means L2-norm and $(f_{x_i}, f_{x_j}) \in [-1, 1]$, for effectively taking two modalities into account, and the ranking loss can be written as

$$\begin{aligned}L_{\text{rank}} = \max\left[0, a - \left(D\left(f_{Ia}, f_{Ta}\right)\right)\right]_{\text{Ianchor}}\\ + \max\left[0, a - \left(D\left(f_{Ia}, f_{Ta}\right)\right)\right]_{\text{Tanchor}},\end{aligned} \tag{4}$$

where I denotes the visual input, T denotes the text input, and α is a margin. Then we use the idea of triplet loss [34], which is one of the very widely used similarity functions. Triplet loss function consists of (a) a sample called Anchor (write to x_a) that is chosen randomly from the training dataset,

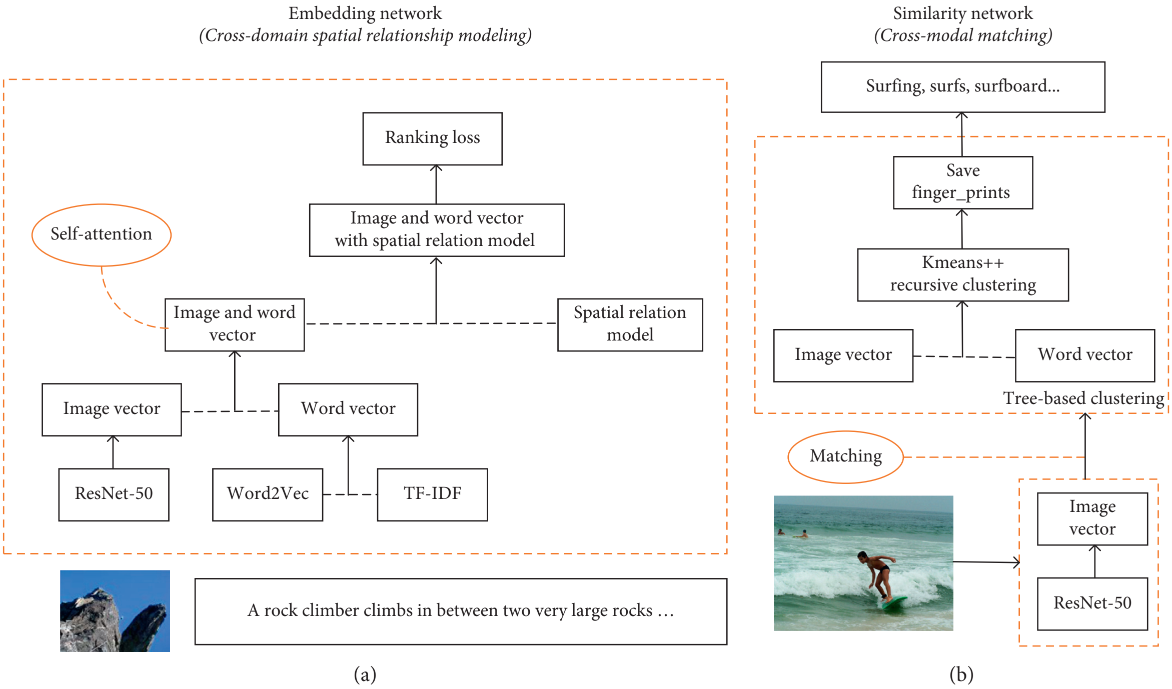


Figure 1: The cross-modal reasoning framework. (a) Embedding network. (b) Similarity network.

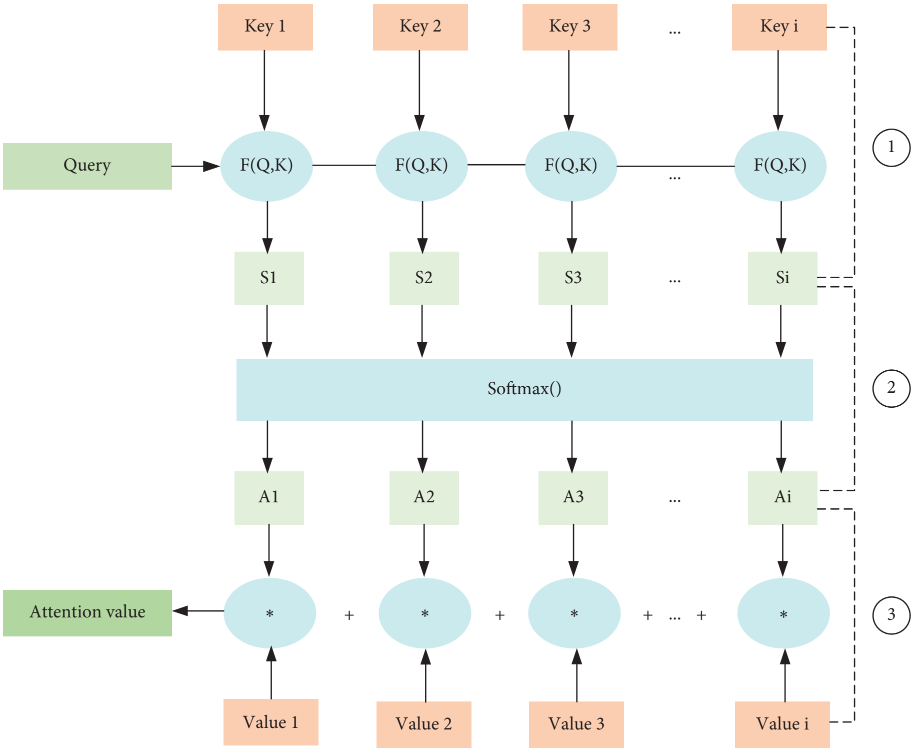


Figure 2: Example to illustrate the process of building an attention model.

(b) a same-class sample as Anchor called Positive (write to x_p), and (c) an example that class is different from Anchor called Negative (write to x_n). The purpose of triplet loss is to minimize the distance between x_a and x_p and maximize the distance between x_a and x_n. We call this kind of ranking loss the triplet ranking loss, and the formulation representation is

$$L_{\text{rank}}\left(x_a, x_p, x_n\right) = \max\left[0, \alpha + \left(D\left(x_a, x_p\right) - D\left(x_a, x_n\right)\right)\right]. \tag{5}$$

3.2.2. Spatial Relation in Image Objects. We now describe object relation computation. The basic idea of the spatial relationships between image objects have their roots in the crucial inference: a visual or hidden object in the image space tends to concentrate for relevant purposes, but distribute randomly for irrelevant ones. Let an object consists of its coordinate feature $\mathbf{f_C}$ and appearance feature $\mathbf{f_A}$. In this work, $\mathbf{f_C}$ is a 4-dimensional object bounding box with relative image coordinates, and $\mathbf{f_A}$ is related to the task.

To be specific, given input set of N visual or hidden objects $\{(f_C^n, f_A^n)\}_{n=1}^N$, the relation feature $f_R(n)$ of the whole sets concerning the n^{th} object, is computed as

$$f_R(n) = \sum_m \omega^{mn} \cdot (W_V \cdot f_A^m). \tag{6}$$

The output is a weighted sum of appearance features from other purposes, linearly transformed by $\mathbf{W_V}$, which is corresponding to values V in equation (1). The relation weight ω^{mn} indicates the impact of other objects. $\mathbf{W_V}$ and ω^{mn} can be calculated by the object-relations model [35]. There are two steps. First, the coordinate features of the two objects are embedded in a high-dimensional representation by ResNet. Inspired by the widely used bounding box regression method DPM [36], the elements are transformed using $\log(\cdot)$ to calculate distant objects and close-by objects. Second, the embedded feature is turned into a scalar weight and trimmed. The trimming operation restricts relations only between objects of individual spatial relationships, which is related to the task and knowledge.

The cross-domain spatial relation model has the same dimension of images and texts at the output, which can be an essential building block in the framework.

3.3. Similarity Network

3.3.1. Image Representation. As the gradient vanishing problem prevents the deep network from being fully learned, we adopt the ResNet-50 to avoid loss of accuracy by learning residual functions to reformulate the layers concerning the inputs instead of learning unreferenced functions. The building block is defined as

$$y = f(x, \{w_i\}) + x, \tag{7}$$

where x is the input vector and y is the output vector of the layers considered. The function f is on behalf of the residual mapping to learn.

As a building block has two line shortcuts, we consider another option. Due to the shortcut connection and element-wise addition, it operates $F + x$. In equation (7), the numbers of channels of x and y are equal. If it is the opposite situation, we adopt the calculation:

$$y = f(x, \{w_i\}) + w_s x. \tag{8}$$

where w_s is the operation of convolution to adjust the channel dimension of x.

We learn the image representation vectors from the lower convolutional layer. In this manner, the decoder can attend to specific parts of an image by selecting a subset of the feature vectors. In this study, we employed 2048 feature maps of the 50-layer residual network.

3.3.2. Text Representation. Prevalent model architecture for Natural Language Processing (NLP) is one-hot coding. Unfortunately, when the output space grows, features cannot properly span the full feature space, consequently, one-hot encoding might result insufficient for fine-grained tasks, since the projection of the outputs into a higher-dimensional space dramatically increase the parameter space of computed models. Also, for datasets with a large number of words, the ratio of samples per word is typically reduced.

A straightforward way to solve the limitations mentioned above is to relax the problem into a real-valued linear programming problem and then threshold the resulting solution. We combine the vector of word embedding and frequency by TF-IDF [27], which depicts the occurrence frequency of a word in all texts, not just the number of occurrences. It allows each word to establish a global context and transforms the high-dimensional sparse vector into a low-dimensional dense vector. Specially, we use self-attention as it will enable each word to learn its relation to other words.

3.3.3. Cross-Modal Matching

Cosine Similarity. The angle between cosines can effectively avoid the differences of degrees in the same cognition of individuals and pay more attention to the differences between dimensions rather than the differences in numerical values. We use the extracted image feature vector to allocate the weight of the word vectors, rather than getting the weight of the word from the text. Then we use the resulting weight and the corresponding word vector to do dot multiplication with image vector to get the similarity between image and text. This allows the semantics of the image and text to interact.

Tree-based Clustering Vector Quantization Algorithm. Let $x = R^d$ be the d-dimensional instance and $y = \{1, 2, \ldots, q\}$ be the class space with a set of q possible classes. Note that two tasks are needed in this term: tree-based training and retrieval. After obtaining the features of the images and texts, the image and text are in the same vector space, and we can use a scalable K-means++ clustering algorithm [37] to both image and text vector.

We develop a tree-based algorithm for cross-modal learning, which presents a tree-based classification model for multiclass learning. A tree-based structure is constructed where the root node corresponds to the classes in the training dataset. Each node v contains a set of k-means++ classifiers. The top node contains all the training data. At the top level, the whole dataset is partitioned into five data subsets $\{A, B, C, D, E\}$. The instances are recursively partitioned into smaller subsets while moving down the tree. Each internal node contains the training instances of its child nodes. Especially, each node in the tree contains two components: cluster vectors and predictive class vectors.

The cluster vector is a vector with real values to measure clusters at a node, and we adopt the definition as

$$p_v(n) = \frac{\sum_{x_i \in D_V} C_{i,n}}{|D_v|}, \quad (9)$$

where $p_v(n)$ is the nth component of p_v, $|D_v|$ is the total number of instances in node v, $C_i = \{Y_{i,1}, Y_{i,2}, \ldots, Y_{i,q}\} \in \{0,1\}^q$. The predictive class vector is a vector with boolean values indicating the membership of predictive classes at a node. The value is 1 when $p_v(n)$ is larger than the threshold. It implies that the node is the proper class or subclass for instances of node v. Note that C_i and the threshold are obtained from the training process.

The algorithm uses three stopping criteria to determine when to stop growing the tree. (1) A node is a leaf node if identified as a predictive class. (2) A node is a leaf node if the data cannot be partitioned to more than three clusters using the classifiers. (3) The tree gets the max depth.

Figure 3 shows the flowchart of the training and retrieval steps. We save the tree model, we calculate the position of the real image or text in the leaf nodes, the path is the pictured fingerprint, we save all picture fingerprints, and then the tree model is built. When matching, it is also necessary to extract image features first. Starting from the root node, each clustering model recursively predicts the category of the vector at a different level. After the image has fallen to the leaf nodes, output the path of the leaf nodes as the fingerprint of the picture. Use cosine distance to find the same text with the fingerprint, and sort it to get the annotation. The process is shown in the Algorithms 1 and 2.

4. Experiments and Results

4.1. Datasets and Evaluation. We use the Flickr30k and Microsoft COCO to evaluate our proposed method, which is widely used in caption-image retrieval and image caption generation tasks. As a popular benchmark for annotation generation and retrieval tasks, Flickr30k contains 31,783 images focusing mainly on people and animals, and 158,915 English captions (five per image). We select 30783 images randomly for training and another 1000 for testing splits.

Microsoft COCO is the largest existing dataset with both captions, and region-level annotations are, which consists of 82783 training images and 40504 validation images, and five sentences accompany each image. We use a total of 82783 images for training and testing splits with 5000 images. The testing splits are images with more than three instances, which are selected from the validation dataset. For each testing image, we use the publicly available division, which is commonly used in the caption-image retrieval and caption generation tasks.

4.2. Implementation Details. We use the ResNet-50 to exploit 2048 feature maps with a size of 1×1 in "conv5_3", which helps us to integrate the high-level features with the lower ones, and the visualization results are provided in Figure 4. By adding a full-connection layer of which dimension is 128, we name the output of graph embedding $v1$. In the text representation, the input layer is followed by two full-connection layers, with dimensions d_1 and d_2, respectively. The output of the last full-connection layer is the output of text module embedding, and we name it v_2. We add a self-attention mechanism to two embedding networks and calculate to get v_1', v_2'. Then, v_1' and v_2' are connected to a triplet ranking loss. Further, we use Adam optimizer to train the system.

After offline debugging, we finally set the parameter of the input layer length of the text module to be 100,000. After the statistics of the word occurrence number of all sentence segmentation words, we take the word occurrence number of the first 99999, and the rest words as a new word. In total, the number of neurons of the input layer is 100,000.

As for the text domain, the second neuron number is 512; the third is 128. Meanwhile, the length of the graph embedding is also 128. We set the margin (in equation (6)), and the parameters of Adam's algorithm are as follows: $\text{lr} = 0.001$, $\text{beta\_1} = 0.9$, $\text{beta\_2} = 0.999$, $\text{epsilon} = 1e{-}08$, $\text{decay} = 0.0$.

For speeding up the training, we conducted negative sampling. The sample sentence that matches the current image is positive. To thoroughly train the low-frequency sentences, we randomly select the part of the sentence outside the sample as negative samples.

4.3. Result of Experiments. Microsoft COCO consists of 123287 images and 616767 descriptions. Five text descriptions accompany each image. We randomly select the training with public 82783 images and remain 5000 images as test data. We can see that our loss function is slightly better than others in Figure 5 when the number of iterations was over 400. The Softmax loss optimizes the distance between classes being great, but it is weak when it comes to optimizing the range within categories. The triplet ranking loss addresses this problem to some extent.

As is shown in Figure 6, we can observe that the words such as "climbs," "rest at," "in the" which show the spatial relationships between objects get significantly higher scores than other words while things like "the" and "a" always have lower matching scores. However, some prepositions containing positional relations, such as "by" in the right-upper corner, have higher scores than other prepositions.

The use of attention causes features related to spatial relation traits to be weighted. As is shown in Figure 7, such as the image in the right-upper corner, we can infer from the coins in the cup next to the person lying down that he is begging.

4.4. Comparison with the State-of-the-Art

4.4.1. Qualitative Evaluation. To evaluate the effectiveness of the proposed method, we first compare its classification performance with state-of-the-art performance reported in the literature. We manually labelled the ground truth mapping from the nouns to the object classes. By varying the threshold, the precision-recall curve can be sketched to measure accuracy. We commonly used "R@1", "R@5", and "R@10", i.e., recall rates at the top 1, 5, and 10 results.

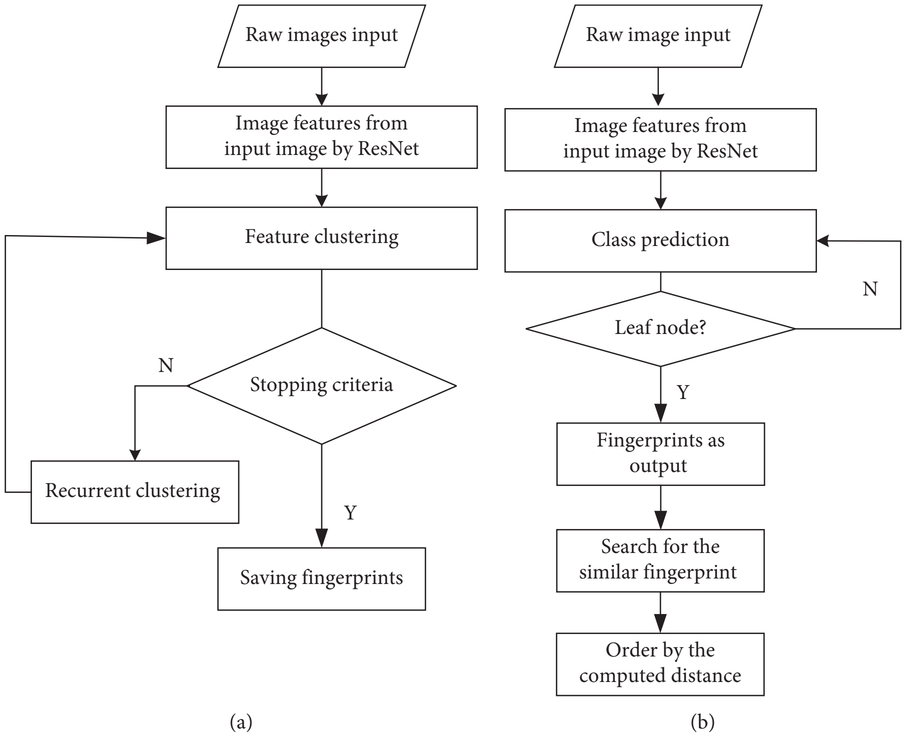


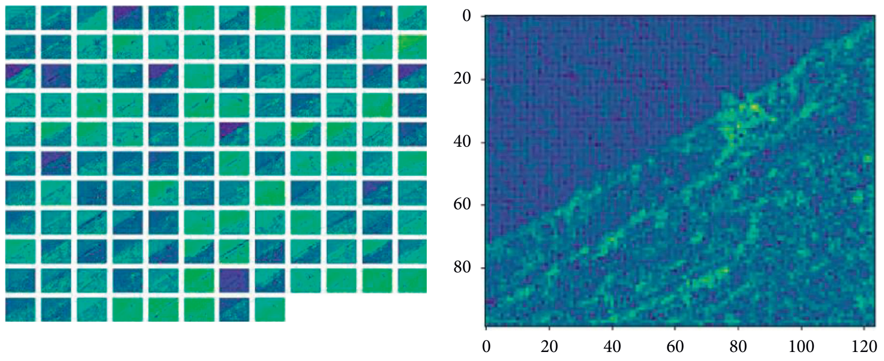
Figure 3: Flowchart of training and searching steps for tree-based algorithm. (a) Flowchart of training steps for tree-based algorithm. (b) Flowchart of searching steps for tree-based algorithm.

```
Input: M, an social image
    D, the current training instances
    N, cluster numbers
    R, a set of features to represent an image
    FP, fingerprints for the image
    T, cluster tree
Output: the fingerprints for M
  Initialize: N, R, T, FP, and resize M;
  compute R by ResNet to represent M;
  Repeat
    create a node v;
    update R according to N clusters to M with k-means++;
    let $D_j$ be the set of data in $D$ satisfying outcome j;
    attach the node returned by $T(D_j)$ to node v;
  until Stop(D, R) is true
  return FP as a finger for the image and attach leaf nodes to FP
```

Algorithm 1: Build fingerprints with a cluster tree.

```
Input: M, an social image
    MD, image dataset for search
    N: number of the image dataset
    FP, fingerprint for the image
Output: S, top-k similar images for image M
Initialize: M, MD, FP, S, and resize images;
M and MD is represented as R by ResNet;
Checking procedure to find the top-k similar images to M in MD.
For each image from MD, get a cosine similarity by FP of the image pair;
return top-k similar images order by cosine similarity
```

Algorithm 2: Find top-k similar images by fingerprints.



(a)

(b)

Figure 4: The visualization of ResNet-50. (a) After 50 layers of convolution. (b) Fusion by 1 : 1 ratio.

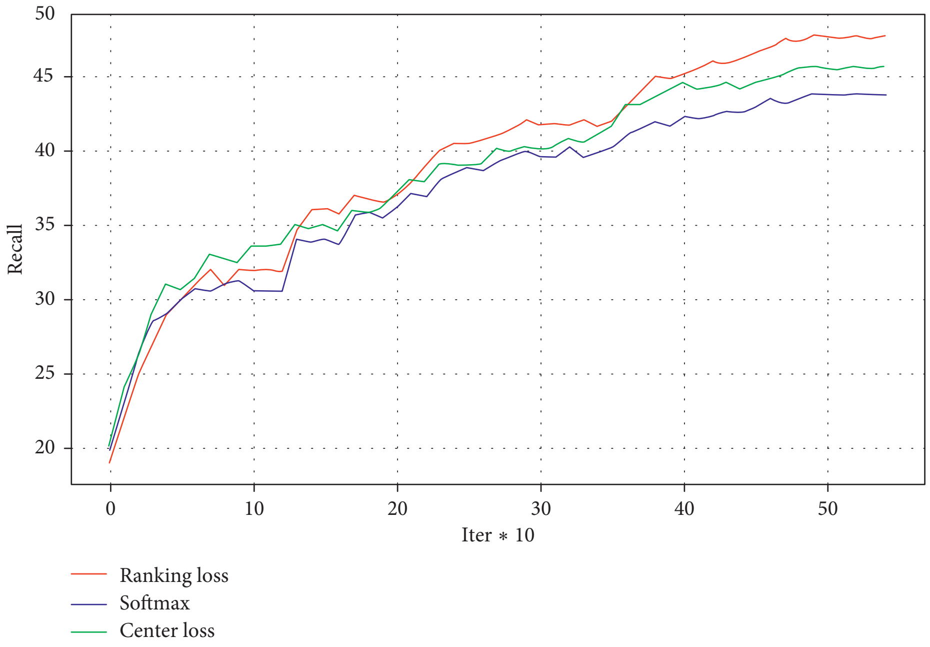
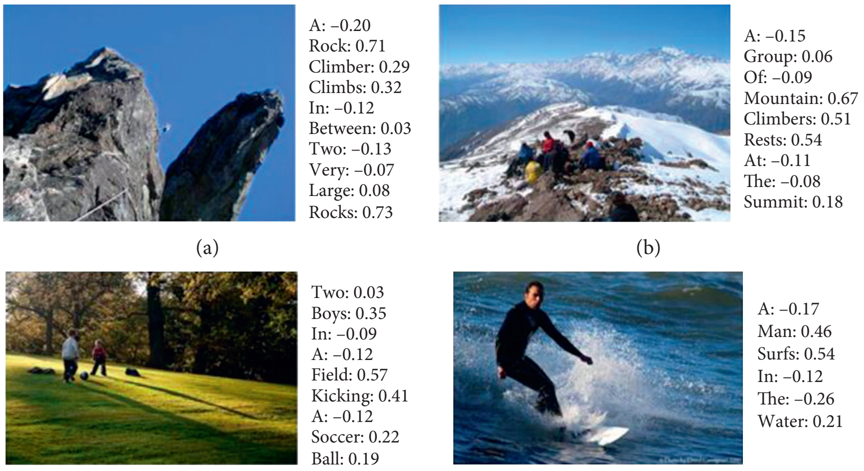


Figure 5: Recall curves when training on Microsoft COCO.



(a)

(b)

(c)

(d)

Figure 6: Similarity matching between image and text.

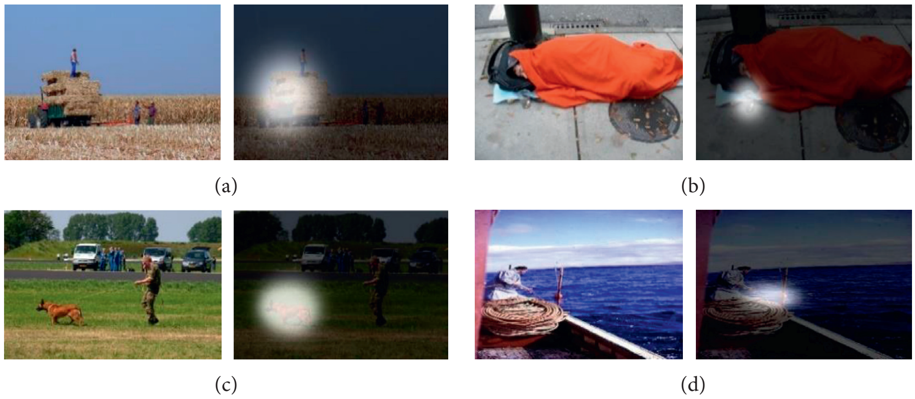


Figure 7: Examples of attending to the correct object (*white* indicates the attended regions, words below the figure indicate the labelling of spatial relation features). (a) Havest. (b) Begging. (c) Search. (d) Salvage.

Table 1: Results of caption-image retrieval on the Flickr30K dataset.

	R@1	R@5	R@10
RVP (*T* + *I*) [38]	11.9	27.7	47.7
CDCCA [39]	16.8	39.3	53.0
MNLM [40]	23.0	50.7	62.9
BRNN [41]	22.2	48.2	61.4
DSPE + FV* [42]	40.3	68.9	79.9
m-RNN# [15]	33.9	65.1	76.3
m-CNN# [43]	34.6	64.7	76.8
Softmax-VGG	21.8	50.7	61.7
Softmax-ResNet	22.3	52.9	63.2
Softmax-ResNet-Attention	35.6	65.4	75.6
Ranking-ResNet-Attention	41.3	69.1	81.3

*Using external text corpora. #The ensemble or multimodel methods.

Table 2: Results of caption-image retrieval on the Microsoft COCO dataset.

	R@1	R@5	R@10
DSPE + FV*	49.4	76.1	86.4
m-RNN#	41.7	72.7	83.5
Sg-LSTM [44]	42.8	73.9	84.7
MNLM	43.9	75.7	85.8
DVSA [45]	39.2	69.9	80.5
m-CNN#	43.1	73.9	84.1
Softmax-VGG	47.5	74.9	84.3
Softmax-ResNet	49.1	76.1	85.2
Softmax-ResNet-Attention	50.3	77.5	87.5
Ranking-ResNet-Attention	52.7	79.3	88.7

*Using external text corpora. #The ensemble or multimodel methods.

Table 3: The results of different caption methods on the Microsoft COCO.

	Bleu-1	Bleu-2	Bleu-3	Bleu-4
SCA-CNN [19]	0.714	0.543	0.409	0.309
Semantic ATT [30]	0.699	0.415	0.306	0.232
SCN-LSTM [20]	0.726	0.581	0.421	0.318
Softmax-ResNet-Attention	0.736	0.587	0.423	0.325
Ranking-ResNet-Attention	0.741	0.598	0.438	0.319
Softmax-ResNet-Attention*	0.703	0.524	0.413	0.311
Ranking-ResNet-Attention*	0.705	0.540	0.416	0.313

*Discarding spatial relation model.

Table 4: The results of different caption methods on the Flickr30K.

	Bleu-1	Bleu-2	Bleu-3	Bleu-4
SCA-CNN [19]	0.705	0.515	0.402	0.304
Semantic ATT [30]	0.694	0.427	0.301	0.224
SCN-LSTM [20]	0.718	0.569	0.413	0.308
Softmax-ResNet-Attention	0.715	0.573	0.429	0.325
Ranking-ResNet-Attention	0.729	0.581	0.416	0.319
Softmax-ResNet-Attention*	0.706	0.533	0.406	0.301
Ranking-ResNet-Attention*	0.712	0.542	0.418	0.309

*Discarding spatial relation model.

As shown in Tables 1 and 2, we compare our proposed method with some of the latest and most advanced processes on the Flickr30k and Microsoft COCO datasets. We can find that our approach works better than the compared methods. Different from DSPE + FV* that uses external text corpora to learn discriminative sentence features, our model learns them directly from scratch in an end-to-end manner. When comparing among our methods, we can conclude as follows: (a) our attention scheme is sufficient since the model with attention consistently outperforms those without notice on both datasets; (b) using ResNet as the underlying network to understand the deep semantics of images to get the spatial relation features with the help of text context relations.

It takes 235 seconds to use a brute force algorithm and 18 seconds to mark with a tree-based clustering algorithm on Microsoft COCO datasets.

4.4.2. Quantitative Evaluation. We also report quantitative evaluation results with the frequently used BLEU metric [46] for the proposed datasets. The results for Microsoft COCO and Flickr30K datasets are listed in Tables 3 and 4. The image encoders of all methods listed here are either VGG-Net or ResNet, which are prevalent in this field. We also report the ablation test in terms of discarding the spatial relation model.

The results demonstrate that our method has a better performance than discarding the spatial relation model.

Besides, our approach is slightly better than the proposed state-of-the-art methods, which verifies the efficiency of topic-condition in image-captioning. Note that our model uses ResNet-50 as the encoder, which is a simple attention model. Thus, our approach is competitive with those models.

5. Discussion and Conclusion

This paper proposes an integrated model to recognize instances and objects jointly by leveraging the associated textual descriptions and presents a learning algorithm to estimate the model efficiently. The learning process requires only separate images and texts without high-level captions. We use the residual network to deepen the learning of image semantic and combine with the text to obtain some hidden relation features contained in the picture. By constructing a joint inference module with self-attention, we make a fusion of local and global elements. We also show that integrating images and text for deep semantic understanding to label the spatial relation features. Furthermore, the use of a tree clustering algorithm accelerates the matching process. Experiments verify that the proposed method achieves competitive results on two generic annotation datasets Flickr30k and Microsoft COCO.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

The project was sponsored by the National Nature Science Foundation of China (41571401 and 41871226) and Natural Science Foundation of Chongqing, China (cstc2019jcyj-msxmX0131).

References

[1] J. Zhu, L. Fang, and P. Ghamisi, “Deformable convolutional neural networks for hyperspectral image classification,” *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 8, pp. 1254–1258, 2018.

[2] W. Lu, J. Li, T. Li, W. Guo, H. Zhang, and J. Guo, “Web multimedia object classification using cross-domain correlation knowledge,” *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1920–1929, 2013.

[3] M. Katsurai, T. Ogawa, and M. Haseyama, “A cross-modal approach for extracting semantic relationships between concepts using tagged images,” *IEEE Transactions on Multimedia*, vol. 16, no. 4, pp. 1059–1074, 2014.

[4] R. Krishna, I. Chami, M. Bernstein, and L. Fei-Fei, “Referring relationships,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6867–6876, Salt Lake City, UT, USA, June 2018.

[5] B. Wang, D. Lin, H. Xiong, and Y. F. Zheng, “Joint inference of objects and scenes with efficient learning of text-object-scene relations,” *IEEE Transactions on Multimedia*, vol. 18, no. 3, pp. 507–520, 2016.

[6] K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang, “Aligning where to see and what to tell: image captioning with region-based attention and scene-specific contexts,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2321–2334, 2017.

[7] Q. Wu, C. Shen, L. Liu, A. Dick, and A. van den Hengel, “What value do explicit high level concepts have in vision to language problems?” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 203–212, Las Vegas, NV, USA, June 2016.

[8] M. J. Choi, A. Torralba, and A. S. Willsky, “A tree-based context model for object recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 240–252, 2012.

[9] R. Mottaghi, X. Chen, X. Liu et al., “The role of context for object detection and semantic segmentation in the wild,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014.

[10] H. Hu, G.-T. Zhou, Z. Deng, Z. Liao, and G. Mori, “Learning structured inference neural networks with label relations,” in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 2960–2968, Las Vegas, NV, USA, June 2016.

[11] J. Johnson, L. Ballan, and L. Fei-Fei, “Love thy neighbors: image annotation by exploiting image metadata,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4624–4632, Tampa, FL, USA, December 2015.

[12] A. Farhadi, M. Hejrati, M. A. Sadeghi et al., “Every picture tells a story: generating sentences from images,” in *Proceedings of the European conference on computer vision*, Springer, Berlin, Heidelberg, 2010.

[13] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, “Composing simple image descriptions using web-scale n-grams,” in *Proceedings of the Conference on Computational Natural Language Learning*, Portland, OR, USA, June 2011.

[14] M. Mitchell, X. Han, J. Dodge et al., “Generating image descriptions from computer vision detections,” in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, 2012.

[15] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, “Deep captioning with multimodal recurrent neural networks (m-RNN),” in *Proceedings of the ICLR*, pp. 1–17, San Diego, CA, USA, May 2015.

[16] H. Hu, G.-T. Zhou, Z. Deng, Z. Liao, and G. Mori, “Learning structured inference neural networks with label relations,” in *Proceedings of the CVPR*, pp. 2960–2968, Las Vegas, NV, USA, June 2016.

[17] F. Liu, T. Xiang, T. M. Hospedales, W. Yang, and C. Sun, “Semantic regularisation for recurrent image annotation,” 2016, https://arxiv.org/abs/1611.05490.

[18] Y. Niu, Z. Lu, J.-R. Wen, T. Xiang, and S.-F. Chang, “Multimodal multi-scale deep learning for large-scale image annotation,” *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1720–1731, 2019.

[19] L. Chen, H. Zhang, J. Xiao et al., “Spatial and channel-wise attention in convolutional networks for image captioning,” in *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*, pp. 6298–6306, Honolulu, Hawaii, July 2017.

[20] Z. Gan, “Semantic compositional networks for visual captioning,” in *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*, pp. 1141–1150, Honolulu, HI, USA, July 2017.

[21] A. Santoro, D. Raposo, D. G. Barrett et al., A Simple Neural Network Module for Relational Reasoning, pp. 4967-4976.

[22] S. Harnad, “The symbol grounding problem,” *Physica D: Nonlinear Phenomena*, vol. 42, no. 1–3, pp. 335–346, 1990.

[23] M. Garnelo, K. Arulkumaran, and M. Shanahan, “Towards deep symbolic reinforcement learning,” 2016, http://arxiv.org/abs/1609.05518.

[24] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, “Building machines that learn and think like people,” *Behavioral and Brain Sciences*, vol. 40, 2017.

[25] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, “Scene graph generation by iterative message passing,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3097–3106, Honolulu, Hawaii, July 2017.

[26] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, “Scene graph generation from objects phrases and region captions,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1270–1279, Venice, Italy, October 2017.

[27] I. Yahav, O. Shehory, and D. Schwartz, “Comments mining with TF-IDF: the inherent bias and its removal,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 3, pp. 437–450, 2019.

[28] K. He, X. Zhang, S. Ren et al., “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NA, USA, July 2016.

[29] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: a unified embedding for face recognition and clustering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, Boston, MA, USA, 2015.

[30] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, “Image captioning with semantic attention,” in *Proceedings of IEEE the Conference Computer Vision and Pattern Recognition*, pp. 4651–4659, Las Vegas, NV, USA, June 2016.

[31] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine translation by jointly learning to align and translate,” 2014, http://arxiv.org/abs/1409.0473.

[32] A. Karpathy, A. Joulin, and L. F. Fei-Fei, “Deep fragment embeddings for bidirectional image sentence mapping,” in *Proceedings of the International Conference on Neural Information Processing Systems*, pp. 1889–1897, Montreal, Canada, 2014.

[33] H. Nam, J. Ha, and J. Kim, “Dual attention networks for multimodal reasoning and matching,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2156–2164, Honolulu, HI, USA, July 2017.

[34] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions,” *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[35] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, “Relation networks for object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3588–3597, Salt Lake City, UT, USA, June 2018.

[36] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part based models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, 2010.

[37] B. Bahmani, “Scalable K-Means++,” in *Proceedings of the Vldb Endowment, 5(7)*, pp. 622–633, Istanbul, Turkey, August 2012.

[38] R. Kiros, R. Salakhutdinov, and R. S. Zemel, “Unifying visual-semantic embeddings with multimodal neural language models,” 2014, http://arxiv.org/abs/1411.2539.

[39] Y. Yu, S. Tang, K. Aizawa, and A. Aizawa, “Category-based deep CCA for fine-grained venue discovery from multimodal data,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 4, pp. 1250–1258, 2019.

[40] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” *Computer Vision-ECCV 2016*, pp. 694–711, 2016.

[41] M. Sinecen, “Comparison of genomic best linear unbiased prediction and bayesian regularization neural networks for genomic selection,” *IEEE Access*, vol. 7, pp. 79199–79210, 2019.

[42] L. Wang, Y. Li, and S. Lazebnik, “Learning deep structure-preserving image-text embeddings,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, NA, USA, June 2016.

[43] Y. Y. ZhangD. S. Zhou et al., “Single-image crowd counting via multi-column convolutional neural network,” in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 589–597, Las Vegas, NV, USA, June 2016.

[44] Y. Xian and Y. Tian, “Self-guiding multimodal LSTM-when we do not have a perfect training dataset for image captioning,” *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5241–5252, 2019.

[45] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 664–676, 2017.

[46] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU:A method for automatic evaluation of machine translation,” in *Proceedings of the Conference Association for Computational Linguistics*, Philadelphia, PA, USA, 2002.