Natural Language Guided Visual Relationship Detection

Wentong Liao[1] Bodo Rosenhahn[1] Ling Shuai[1] Michael Ying Yang[2]
[1]Institute für Informationsverarbeitung, Leibniz Universität Hannover, Germany
[2]Scene Understanding Group, University of Twente, Netherlands
{liao, shuai, rosenhahn}@tnt.uni-hanover.de michael.yang@utwente.nl

Abstract

*Reasoning about the relationships between object pairs in images is a crucial task for holistic scene understanding. Most of the existing works treat this task as a pure visual classification task: each type of relationship or phrase is classified as a relation category based on the extracted visual features. However, each kind of relationships has a wide variety of object combination and each pair of objects has diverse interactions. Obtaining sufficient training samples for all possible relationship categories is difficult and expensive. In this work, we propose a natural language guided framework to tackle this problem. We propose to use a generic bi-directional recurrent neural network to predict the semantic connection between the participating objects in the relationship from the aspect of natural language. The proposed simple method achieves the state-of-the-art on the Visual Relationship Detection (VRD) and Visual Genome datasets, especially when predicting unseen relationships (*e.g.*, recall improved from 76.42% to 89.79% on VRD zero-shot testing set).*

1. Introduction

Scene understanding is one of the most primal topics in the computer vision and machine learning communities. It ranges from the pure vision tasks, such as object classification/detection [18, 31], semantic segmentation [23, 45], to the comprehensive visual-language tasks, *e.g.*, image/region caption [16, 39], scene graph generation [40, 38, 21], and visual question-answering [2, 37], *etc.* Boosted by the impressive development of deep learning, the research of pure version tasks is becoming gradually mature. However, it is still challenging to let the machine understand the scene at a higher semantic level. Visual relation detection is a promising intermediate task to bridge the gap between the vision and the visual-language tasks and has caught increasing attention [24, 21, 41, 42, 15].

Visual relation detection targets on understanding the visually observable interactions between the detected objects in images. The relationships can be represented in a triplet form of ⟨*subject-predicate-object*⟩, *e.g.*, ⟨*kid-on-skate board*⟩, as shown in Figure 1. A natural approach for this task is to treat it as a classification problem: each kind of relationships/phrase is a relation category [32], as shown in Fig. 2. To train such reliable and robust model, sufficient training samples for each possible ⟨*subject-predicate-object*⟩ combination are essential. Consider the Visual Relationship Dataset (VRD) [24], with $N = 100$ object categories and $K = 70$ predicates, then there are $N^2K = 700k$ unique combination in total. However, it contains only $38k$ relationships, which means that each combination has less than 1 sample on average. The previous classification-based works can only detect the most common relationships, *e.g.*, [32] studied only 13 frequent relationships.

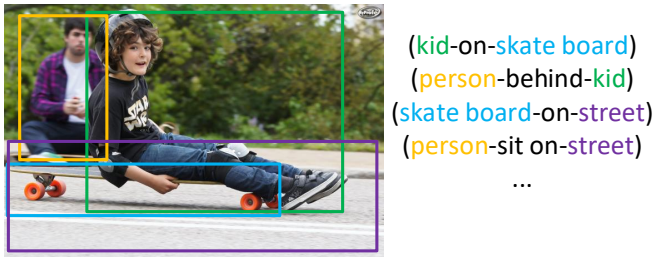
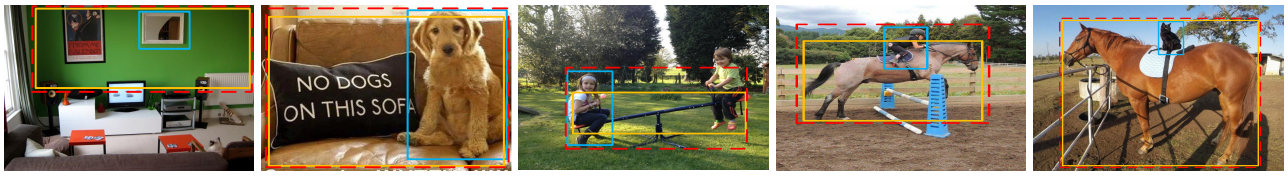


Figure 1: Visual relationships represent the interactions between observed objects. Each relationship has three elements: *subject*, *predicate* and *object*. Here is an example image from Visual Genome [17]. Our proposed method is able to effectively detect numerous kinds of different relationships from such image.

To handle above challenges, another approach is to predict the object categories and their predicate types independently. However, the semantic relationship between the objects and the predicates are ignored in this kind of method. Consequently, the phrase which has the same predicate but different agents is considered as the same type of relationship. For instance, the *"clock-on-wall"* (Fig. 2a) and *"dog-*



(a) picture-on-wall (b) dog-on-sofa (c) girl-playing-seesaw (d) kid-riding-horse (e) cat-siting on-horse

Figure 2: Examples of the wide variety of visual relationships, and its difference with the phrases. The solid bounding boxes indicate the individual objects and the dash red bounding boxes denote a phrase.

on-sofa" (Fig. 2b) belong to the same predicate type *"on"*, but they describe different semantic scenes. On the other hand, the type of relationship between two objects is not only determined by their relative spatial information but also their categories. For example, the relative position of between the kid and the horse (Fig. 2d) is very similar as the ones between the cat and the horse (Fig. 2e), but it is preferred to describe the relationship *"cat-sitting on-horse"* rather than *"cat-riding-horse"* in the natural language. It's also very rare to say *"person-sitting on-horse"*.

Another important observation is that the relationships between the observed objects are naturally based on our language knowledge. For example, we would like to say the kind "sitting on" or "playing" the seesaw but not "riding" (Fig. 2c), even though it has the very similar pose as that the kind "riding" the horse in Fig. 2d. On the other hand, similar categories have a similar semantic connection, for example, *"person-ride-horse"* and *"person-ride-elephant"*, because "horse" and "elephant" belong to the same category of animal. It is an important cue for inferring the infrequent relationships from the frequent instances. Fortunately, this semantic connection has been well researched in the language model [26, 27]: an object class is closed to another one if they belong to the same object category and far from the one belonging to a different category in the word-encoded space. The vivid example given in [26] *king-man=queen-woman* reveals that the inherent semantic connection between "king" and "man" is the same as "queen" and "woman". Here, "king" and "queen" belong to the same category (ruler) while "man" and "woman" in the same category (person). Therefore, we resort the powerful semantic connection in the language to handle the challenging problems in the task of visual relationship detection.

In this work, we propose a new framework for visual relationship detection in large-scale datasets. The visual relationship detection task is roughly divided into two subtasks. The first task is to recognize and localize objects that are present in a given image. It provides the visual cues of *"what"* and *"where"* are the objects. The second task is to reason about the interaction between an arbitrary pair of the observed objects. It understands *"how"* they connect with each other in a semantic context. We show that our model is able to scale and detect thousands of relationship types by leveraging the semantic dependencies from language knowledge, especially to infer the infrequent relationships from the frequent ones.

The major **contributions** of this work are as follows:

1. We propose to use a generic bi-directional recurrent neural network (RNN) [33, 25] to predict the semantic connection, *e.g.*, predicate, between the participating objects in the relationship from the aspect of natural language knowledge.

2. The natural language knowledge can be learned from any public accessible raw text, *e.g.*, the image captions of a dataset.

3. The visual features of the union boxes of the two participating objects in the relationships are not required in our method. State-of-the-art methods [24, 20, 21, 43, 6] rely on such features. Furthermore, our method is able to infer infrequent relationships from the frequent relationship instances.

4. Our model is competitive with the state-of-the-art in visual relationship detection in the benchmark datasets of Visual Relationship Dataset [24] and Visual Genome [17], especially when predicting unseen relationships (*e.g.*, recall improved from 76.42% to 89.79% on VRD zero-shot testing set).

2. Related Work

As an intermediate task connecting vision and vision-language tasks, many works have attempted to explore the use of visual relationship for facilitating specific high-level tasks, such as image caption [4, 8], scene graph generation [40, 38], image retrieval [30], visual question and answering (VQA) [2, 1, 37], *etc*. Compared to these works which treat the visual relationship as an efficient tool for their specific tasks, our work dedicates to provide a robust framework for generic visual relationship detection.

Visual relationship detection is not a new concept in literature. [9, 11] attempted to learn four spatial relationships:

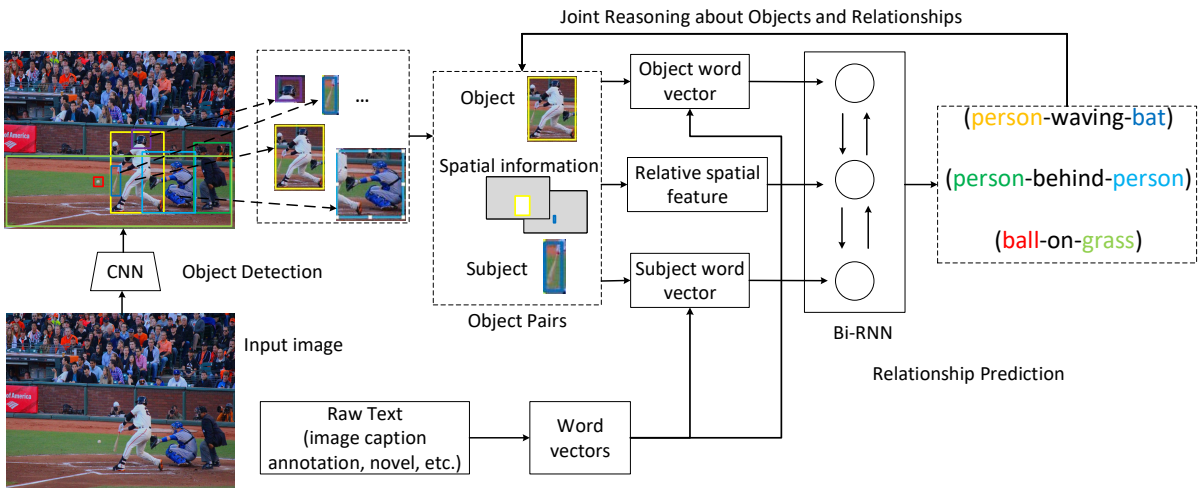


Figure 3: The proposed framework for visual relationship detection. First, Faster RCNN is utilized to localize objects and provide the classification probability of each detected object in the given image. Then, the possible meaningful object pairs are selected as the candidate relationships. Each object converted into the corresponding word vector.

"above", "below", "inside", and "around". [34, 40] detected the physical support relations between adjacent objects: support from "behind" and "below". In [32, 7], each possible combination of visual relationship is treated as a distinct visual phrase class, and the visual relationships detection is transformed to a classification task. Such methods suffer the long trail problem and can only detect a handful of the frequent visual relationships. Besides, all above works used the handcraft features.

In recent years, deep learning has shown its great power in learning visual features [18, 35, 13, 36, 23]. The most recent works [24, 38, 20, 21, 6, 44, 28, 29] use deep learning to learn powerful visual features for visual relationships detection. In [38], the visual relationships are treated as the directed edges to connect two object nodes in the scene graph. The relationships are inferred along the processing of constructing the scene graph in an iterative way. [20, 21] focused on extracting more representative visual features for visual relationships detection, object detection, and image caption [21]. [6, 44] reasoned about the visual relationships based on the probabilistic output of object detection. [44] attempted to project the observed objects into relation space and then predict the relationship between them with a learned relation translation vector. [6] proposed a particular form of RNN (DR-Net) to exploit the statistical relations between the relationship predicate and the object categories, and then refine the estimates of posterior probabilities iteratively. It achieves substantial improvement over the existing works. However, most of the existing works [20, 21, 6] require additional union bounding boxes which cover the object and subject together to learn the visual features for relationship prediction. Besides, their works are mainly designed based on visual aspect. In this paper, we analyze the visual relationships from the language aspect. The most related works are [24, 43, 29], which proposed to use linguistic cues for visual relationship detection. [24] attempted to find a relation projection function to transform the word vectors [26] of the participating objects into the relation vector space for relationship prediction. [43] exploited the role of both visual and linguistic representations and used internal and external linguistic knowledge to regularize the network's learning process. [29] proposed a framework for extracting visual cues from a given image and linguistic cues from the corresponding image caption comprehensively. In particular, for visual relationship detection in the VRD dataset, 6 Canonical Correlation Analysis [10] models are trained. Different from their works, our method uses a modified Bidirectional RNN (BRNN) to leverage the natural language knowledge, which is much simpler and outperforms [10] regarding visual relationships prediction.

3. Visual Relationship Prediction

The general expression of visual relationships is $\langle$*subject-predicate-object*$\rangle$. The component "predicate" can be an action (*e.g.* "wear"), or relative position (*e.g.* "behind"), *etc*. For convenience, we adopt the widely used convention [32, 24] to characterize each visual relationship in the triplet form as $\langle s\text{-}p\text{-}o \rangle$, such as $\langle$*person-wave-bat*$\rangle$. Here, s and o indicate the subject and object category respectively, while r denotes the relationship predicate. Concretely, the task is to detect and localize all objects presented in an image and predict all possible visual relationships between any two of the observed objects. Note that, "no rela-

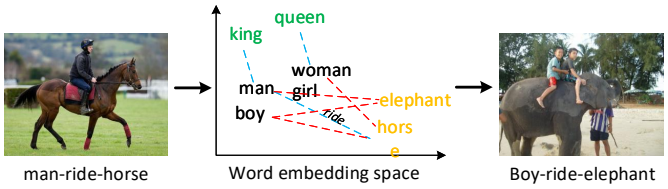
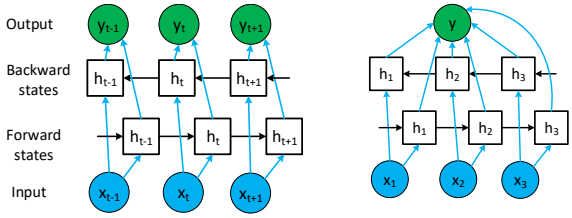


Figure 4: Example of inferring infrequent relationships (the left image) from frequent instances (the right image) guided by natural language knowledge. In the middle image, the blue dashed lines denote the distances between words. We assume this distance as the inherent semantic connection in natural language knowledge. The infrequent relationships, which is connected with red dashed lines, can be inferred from the frequent relationships.



(a) Standard BRNN (b) Our BRNN

Figure 5: The standard BRNN model [33] (a), and (b) our BRNN model used for predicate prediction. Our BRNN has three inputs in sequence(*subject*, *spatial information* and *object*) and one output (predicate prediction).

tion" is also a kind of visual relationship between two objects in this work. For instance, in Fig. 2e, there is no explicit visual relationship between the "cat" and the "tree". An overview of our proposed framework is shown in Fig. 3. It comprises multiple steps, as described as follows.

3.1. Object detection

Before reasoning about the visual relationships, objects present in the given images must be localized as a set of candidate components of the relationships. In this work, the Faster RCNN [31] is utilized for this task because of its high accuracy and efficiency. Each detected object comes with a bounding box to indicate its spatial information, and the predicted object class distribution $\mathbf{p_o} = \{p_1, \cdots, p_{N_o}\}$, N_o is the total number of object categories. And the location of each detected object is denoted as (x_s, y_s, w, h), where (x_s, y_s) is the normalized coordinate of the the bounding box center on the image plane, and (w, h) is the normalized *width* and *height* of the bounding box. The subscript 's' denotes the 'spatial' and prevents from confusion with following denotation.

3.2. Natural language guided relationship recognition

The word vectors embed the semantic context between different words in a semantic space [26, 27]. The words which have similar semantic meaning are close to each other in the space, for example as shown in the middle image of Fig. 4. On the other side, the distances between the words in a semantic group and the words in different semantic groups could be similar. Even though the distance between different words in the embedded word space is calculated as cosine distance [26], we assume that it is inherent semantic relationships connecting the two words rather than a mathematics distance in the embedding space. For example, the semantic connection between *"person"* and "horse" is normally *"ride"*. *"horse"* and *"elephant"* are in the same semantic group (*animal*). Therefore, *"ride"* is very likely the semantic connection between *"horse"* and *"elephant"*. This semantic property is important to lean the infrequent relationship (*e.g.*, "person ride elephant, camel, tiger, *etc.* ") from the very normal relationship ("person ride horse") in the real world. Fig. 4 illustrates a brief process of this inference.

Bi-directional RNNs (BRNNs) [33] have achieved great successes for natural language processing tasks [14, 3, 5, 12]. The standard BRNN structure is shown in Fig. 5a. The vector x_t is the input of a sequence at time point t and y_t is the corresponding output, while h_t is the hidden layer. A BRNN computes the hidden states twice: a *forward* sequence $\overrightarrow{\boldsymbol{h}}$ and a *backward* sequence $\overleftarrow{\boldsymbol{h}}$. Each component can be expressed as follows:

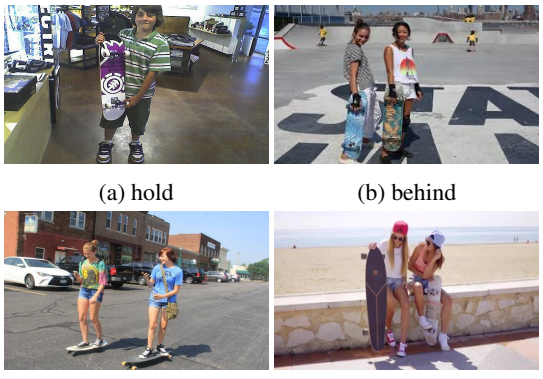
$$\overrightarrow{h_t} = \mathcal{H}(W_{x\overrightarrow{h}} x_t + W_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}), \quad (1)$$

$$\overleftarrow{h_t} = \mathcal{H}(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}), \quad (2)$$

$$y_t = W_{\overrightarrow{h}y} \overrightarrow{h_t} + W_{\overleftarrow{h}y} \overleftarrow{h_t} + b_y. \quad (3)$$

where $W_{x\overrightarrow{h}}$ denotes the input-hidden weight matrix in forward direction. $b_{\overrightarrow{h}}$ denotes the bias vector of the hidden layer in forward direction. $\mathcal{H}$ is the activation function of the hidden layers. We use the RELU function [19] in this work. The output sequence $\mathbf{y}$ is computed by iterating the considering both of the forward and backward input sequence $\mathbf{x}$. This process plays an important role in visual relationship detection. Since in a relationship expression $\langle$*subject-predicate-object*$\rangle$, the order of the two objects is decisive for the final prediction. *E.g.*, $\langle$*person-ride-horse*$\rangle$ is completely different from $\langle$*horse-ride-person*$\rangle$. A BRNN is able to learn such difference caused by the input sequence. This is the main reason why we used BRNN to learn the linguistic cues between the objects categories.

Besides the object categories, the relative position of the



(a) hold (b) behind

(c) on (d) next to

Figure 6: The relative position of the two objects is crucial for the relationship prediction.

participating objects is crucial for predicate prediction, as illustrated in Fig. 6. Even though the object categories are the same in all instances, the predicate between the *"person"* and the *"skateboard"* is different in each image. Therefore, we modify the standard BRNN with 3 sequential inputs $[x_1, x_2, x_3]$(Fig. 5b). x_1 and x_3 are the 300-dim word vectors of the participating objects, respectively. They are obtained by sum of the word vectors of each object category weighted by the predicted probability (namely *soft embedding*) formally defined as:

$$x_i = \mathbf{p}_{o_i} \mathbf{W}_{word2vec}, i \in 1, 3 \tag{4}$$

where $\mathbf{W}_{word2vec} \in R^{N \times 300}$ is the matrix of word vectors of N object categories, and $\mathbf{p}_{o_i}$ is the predicted distribution of object i. Here, the *Glove* algorithm [27] is used to learn the word vectors because of its high efficiency and robust performance. x_2 is the spatial configuration of the two objects and is obtained as fellows. firs, the relative spatial relationship of the two objects is represented as:

$$\begin{aligned} s = &[x_1^s, y_1^s, w_1, h_1, \frac{x_1^s - x_3^s}{W}, \frac{x_1^s - x_3^s}{H}, \\ &log\frac{w_1}{w_3}, log\frac{h_1}{h_3}, x_3^s, y_3^s, w_3, h_3], \end{aligned} \tag{5}$$

where $[x_i^s, y_i^s, w_i, h_i]$ is the predicted bounding box of object $i \in 1, 3$. $W and H$ is the width and height of the union bounding box of the two objects, respectively. $\frac{x_1^s - x_3^s}{W}, \frac{x_3^s - x_3^s}{H}, log\frac{w_3}{w_3}, log\frac{h_1}{h_3}$ encodes their relative spatial relationship which is important for relationship representation. s is then feed to a 2-layer MLP to achieve a 300-dim sparse spatial representation x_2 of visual relationship. Eq. (3) are redefined in our framework as:

$$\begin{aligned} y = &(W_{\overrightarrow{h_1}y} \overrightarrow{h_1} + W_{\overrightarrow{h_2}y} \overrightarrow{h_2} + W_{\overrightarrow{h_3}y} \overrightarrow{h_3}) + \\ &(W_{\overleftarrow{h_1}y} \overleftarrow{h_1} + W_{\overleftarrow{h_2}y} \overleftarrow{h_2} + W_{\overleftarrow{h_3}y} \overleftarrow{h_3}) + b_y \end{aligned} \tag{6}$$

.

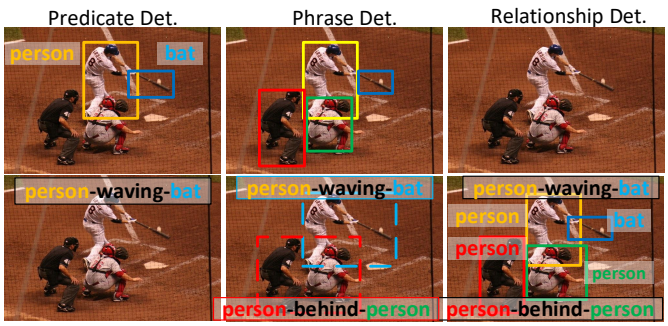


Figure 7: Illustration of different task settings. The first row depicts the inputs for different tasks and the second row is the corresponding outputs. The solid bounding boxes localize individual objects while the dashed bounding boxes localize the locations of phrases.

The first term is the information passed from the *forward* sequence and the second term is from the *backward* sequence. The output $y \in R^K$ is the predicted distribution of the K predicates. $\overrightarrow{h_0}$ is set as zero for *forward* pass, and so as for $\overleftarrow{h_4}$ for *backward* pass. Notes that, any public accessible corpus can be used to pretrain the word vectors, such as the image caption of the COCO dataset [22]. The pretrained word vectors can be further trained during the training of the models. In Sec. 4, we will study the influence of the word vectors learned from different corpus.

3.3. Joint recognition

At the test time, the categories of detected objects and the predicate types are jointly recognized. The joint probability of relationships and object detection can be written as:

$$p(O_s, r, O_o) = p(O_s)p(r|O_s, O_o)p(O_o). \tag{7}$$

where $p(O_s)$ and $p(O_o)$ are the probabilities of categories predicted by Faster RCNN of the *subject* and *object* respectively. $p(r|O_s, O_o)$ is the probabilities of predicted predicate given by the BRNN. On each test image, we find the optimal prediction using:

$$\langle O_s^*, r^*, O_o^* \rangle = \underset{O_s, r, O_o}{\arg\max}\, p(O_s, r, O_o) \tag{8}$$

4. Experiments

We evaluated our model on two datasets. (1) **VRD** [24]: the dataset contains $5,000$ images with 100 object categories and 70 predicates. There are $38k$ visual relationship instances that belong to $6,672$ relationship types. We follow the train/test split in [24]. Note that, $1,877$ relationships only present in the test set, which are used to evaluate the zero-shot relationship detention performance. (2)

Dataset	Comparison	Predicate Detection		Phrase Detection		Relationship Detection	
		Rec@50	Rec@100	Rec@50	Rec@100	Rec@50	Rec@100
VG	LP [24]	26.67	33.32	10.11	12.64	0.08	0.14
	SG [38]	58.17	62.74	18.77	20.23	7.09	9.91
	MSDN [21]	67.03	71.01	24.34	26.50	10.72	14.22
	DR-Net [6]	**88.26**	91.26	23.95	27.57	20.79	**23.76**
	Ours	85.02	91.77	**28.58**	**31.69**	**22.17**	23.62
	Ours+COCO [22]	83.87	**92.17**	27.26	30.87	20.54	23.05
	Ours*	84.44	89.47	27.97	30.09	22.01	23.53
	Ours†	12.79	16.07	5.85	7.58	0.12	0.53
VRD	LP [24]	47.87	47.87	17.03	16.17	14.70	13.86
	CCA [29]	-	-	16.89	20.70	15.08	18.37
	DR-Net [6]	80.78	81.90	19.93	23.45	17.73	20.88
	LK [43]	**85.64**	**94.65**	26.32	29.43	**22.68**	**31.89**
	Ours	84.39	92.73	**28.63**	**31.97**	20.63	21.97
	Ours+COCO [22]	82.04	90.15	25.37	31.83	21.02	23.30

Table 1: Experimental results of different methods in the VRD [24] and VG [17]. We compare our method with the existing works on the three tasks discussed in Sec. 4.1. "*" indicates that the pretrained word vectors are further jointly trained, and "†" denotes that the word vectors are randomly initialized and are jointly trained.

Visual Genome (VG) [17] dataset has $108K$ images and $998K$ relationships($74,361$ relationship types). We adopt the data split defined by Li *et al*. [21] which has $96k$ images among which $71k$ being used for training. There exist 150 object categories and 50 predicate types in this data setting.

4.1. Experiment settings

training details. In the experiments, the Faster RCNN with VGG16 [35] is used as the underlying object detector and is pretrained on the ImageNet dataset. The BRNN model has two hidden layers, and each of them has 128 hidden states. Its parameters are initialized randomly. The word vectors (word2vec) are learned by *Glove* [27] from the caption annotation in VG [17]. The differences between fixing the pretrained word vectors and training them jointly are analyzed. Additionally, the word2vec are trained on the image caption from COCO dataset [22] to evaluate the generalization ability and the robustness of our method when the sources of language knowledge are different.

Task settings. Visual relationship detection involves localizing and classifying both the objects and predicting the predicate. We evaluate our model on three convention tasks [24]: (1) **Predicate detection**: In this task, the ground truth locations and labels of objects are given. This task aims at measuring the accuracy of predicate recognition without the effect of the object detection algorithms. (2) **Phrase detection**: The input is an image and the ground truth locations of objects, and the output is a set of relationships. When all the three entities are correctly predicted and the IoU between the predicted union boxes and the ground truth is above 0.5, this prediction is considered as correct.

predicate	[38]	Ours	predicate	[38]	Ours
on	**99.71**	99.39	under	56.93	**83.44**
has	96.47	**98.47**	sitting on	57.01	**91.07**
in	88.77	**93.87**	standing on	61.90	**78.06**
of	96.18	**97.80**	in front of	64.63	**75.67**
wearing	98.01	**99.59**	attached to	27.43	**70.00**
near	95.14	**99.57**	at	70.41	**86.33**
with	88.00	**93.69**	hanging from	0	**67.50**
above	70.94	**86.33**	over	0.69	**56.00**
holding	82.80	**96.18**	for	11.21	**57.22**
behind	84.12	**93.30**	riding	91.18	**95.08**

Table 2: The per-type predicate classification accuracy with metric Rec@5. These predicate types are the $Top-20$ most frequent cases in the dataset (sorted in descending order in the table).

This task evaluates the model ability of object classification and predicate prediction. (3) **Relationship detection**: Given an image, a set of relationships are predicted. Not only the two objects categories and their relation muss be correctly predicted, but also the IoUs between the predicted locations and their ground truth boxes of both *subject* and *object* are over 0.5 simultaneously. This task evaluates the model for both object and predicate detection. An illustration for different task settings is shown in Fig. 7. For evaluation, we follow the metrics for visual relationship detection [38] by using the *Top-K* recall(denoted as *Rec@K*), which is the fraction of ground truth instance which fall in

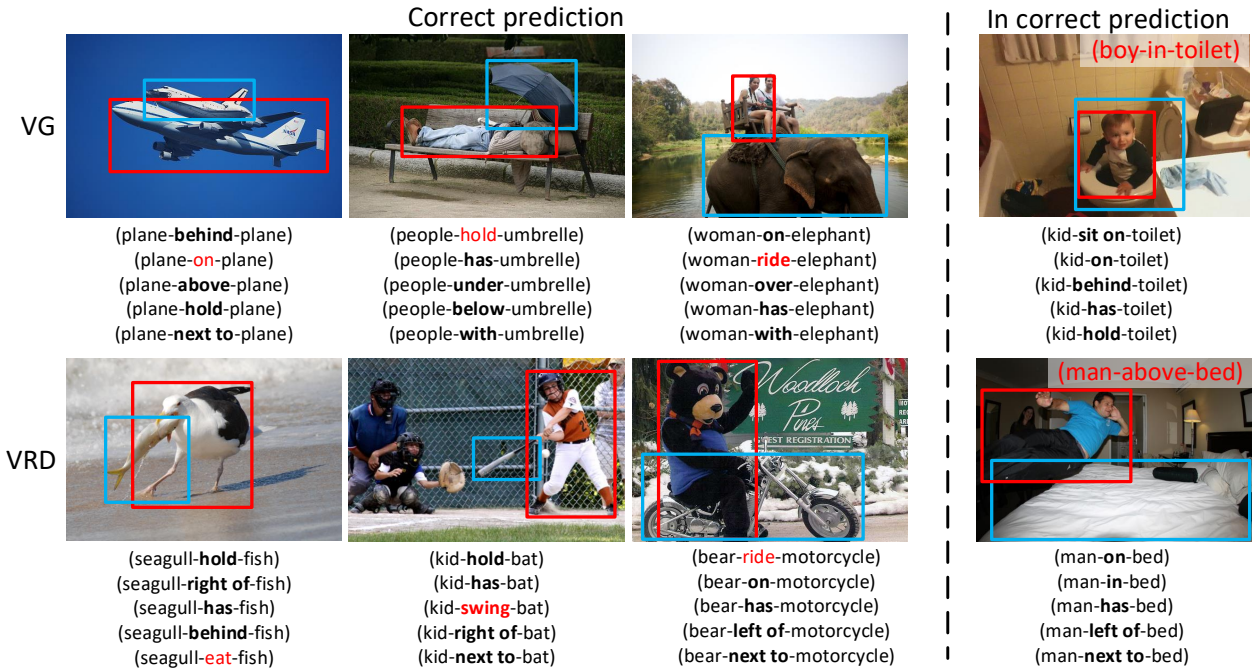


Figure 8: Qualitative examples of visual relationships detection on VG [17] (the first row) and VRD [24] (the second row) respectively. The first three columns illustrate the correctly recognized relationships in the images, while the last column is the failure example (the ground truth relationship phrase is denoted in red in the image). The red bounding boxes denote the subjects while the cyan boxes denote the objects. The relationships under the images are the $Top-5$ most probable relationships predicted by our method, in which the red denotes the ground truth.

the *Top-K* predictions.

4.2. Comparative results

Our method is compared with: **LP** [24] is the first work of visual relationship detection. It's deem as the baseline. **SG** [38] is the first work which detects visual relationships on the VG dataset. **MSDN** [21] and **DR-Net** [6] are the most recent work and report the state-of-the-art performance **CCA** [29] and **LK** [43] use the linguistic cues associating with the visual cues for visual relation detection.

Table 1 shows the results of different methods. Our method outperforms many of the existing works in the three task settings in both datasets, and competitive with [6, 43] on some tasks. From the table one can observe that, even though the improvements for predicate detection from SG [38] to DR-Net [6] are significant (28.52% for *Rec@100* on VG), the improvements for phrase detection are much less (7.34% for *Rec@100* on VG). Because [6, 38] only use visual features for visual relationship detection which cannot well represent the semantic connection between the dependencies of different object categories and their possible predicates. On the other side, our method achieves substantial improvements on the phrase detection task. These results show that our method can effectively pair the objects which have important relationships and precisely predict their predicate in the images.

LP [24] also used extracted word vectors for visual relationship detection. However, the linear projection function in their model, which transforms the objects categories into the relationship vector space, is inadequate for predicting numerous kinds of relationships. In contrast, our BRNN model includes multiple nonlinear activation function which can learn more representative features than LP [24]. Our method also outperforms CCA [29], which also used linguistic cues for visual relationship detection. While the performance of our method is inferior to LK [43] on the tasks of predicate detection (1.92% for *Rec@100*) and relationship detection (9.92% for *Rec@100*), our approach performs better on the task of phrase detection (2.54% for *Rec@100*).

Fig. 8 shows some qualitative examples of visual relationships detection in the two datasets. From the first three columns we can see that the $Top-5$ most probable predicted predicates between the objects are highly close to the ground truth, *e.g.*, *on* is very close to *ride* from the spatial aspect. The rare relationship of $\langle$*bear-ride-motorcycle*$\rangle$ is successfully predicted with the highest probability, which shows that our model can learn very rare relationships from

	Predicate Detection		Phrase Detection		Relationship Detection	
	Rec@50	Rec@100	Rec@50	Rec@100	Rec@50	Rec@100
LP [24]	8.45	8.45	3.75	3.36	3.52	3.13
CCA [29]	-	-	10.86	15.23	9.67	13.43
LK [43]	56.81	76.42	13.41	17.88	12.29	16.37
Ours	**80.75**	**90.52**	**21.10**	**22.35**	**19.61**	**22.03**

Table 3: Experimental results for zero-shot visual relationship detection on the VRD dataset [24].

the normal relationships guided by natural language knowledge. The last column gives a failure example of each dataset. ⟨*kid-in-toilet*⟩ and ⟨*man-above-bed*⟩ are both abnormal scenes in the real world. The natural language knowledge extracted by our model guides the prediction towards more probable results. Our current model fails to detect abnormal interactions in the natural scenes.

We also trained the word vector using the image caption in COCO dataset [22], as shown in the row of "Ours+COCO" in Table 1. We can see that the performance of using COCO corpus decreases a little. The main reason is that COCO dataset provides richer manual image caption annotation (5 captions per image) while the VG dataset provides region caption which is more like a phrase rather than a sentence. Therefore, the learned word vectors from VG are more closed to the relationship triplet representation. nevertheless, "Ours+COCO" also reports competitive results, which proofs that, first our model is robust to different corpus source, second natural language knowledge is useful for visual relationship detection. To further explore the effectiveness of natural language knowledge, one extensional experiment is conducted by using randomly initialized word vectors and jointly training them. The experimental results are listed in the row of "$Ours\dagger$" in Table 1. The performance deteriorates seriously: it reports the worst performance among the all methods. This phenomenon demonstrates that, the semantic connection cannot well explored directly from the visual appearance information. It also further proof that using natural language knowledge for visual relationship detection is an effective and robust scheme. There are many large-scale image datasets for object detection. But most of them don't have the annotation for visual relationship detection. It could be an effective solution to generate high-quality annotation of visual relationships using natural language knowledge automatically.

Table 2 shows the performance on predicting per-type predicate of SG [38] and our method. These results are calculated in the task of predicate detection. Our method reports much better results than SG [38] on each type predicate classification, in particular, 67.50% improvement on the type *hanging from*. This table shows that our model performs well in predicting frequent predicates.

4.3. Zero-shot learning

In the VRD dataset [24], there are $1,877$ relationships in the test set that have never occurred in the training set. Our trained model is utilized to detect these unseen relationships to evaluate its ability of inferring infrequent relationships based on frequent relationships that have ever seen, namely *zero-shot learning*. Table 3 shows the results from different works. Our method outperforms LP [24], CCA [29] and LK [43] by large margin, while only decreasing slightly compared with the results shown in Table 1. This table shows that our model has good generalization ability: it can detect thousands of relationships, even the instances that have never been seen.

5. Conclusion

This paper presents a natural language knowledge guided method for detecting visual relationships in images. The semantic connection between the object categories and predicate are embedded in the word vector learned by natural language processing. We designed a BRNN model to predict the predicate between two observed objects based on this natural language knowledge and their spatial information. In particular, our method is able to infer infrequent relationships from the frequent relationship instances, which is important to deal with the long tail problem. Experiments on the Visual Genome and Visual Relationship Datasets show substantial improvements compared with most existing works for visual relationship detection in terms of accuracy and generalization ability. In the zero shot learning task, the proposed method shows the potential for detection thousands of relationships. In the future work, we would like to extend our current model to an end-to-end framework which can learn better features from image and language knowledge from raw text simultaneously.

Acknowledgment

The work is funded by DFG (German Research Foundation) YA 351/2-1 and RO 4804/2-1 within SPP 1894. The authors gratefully acknowledge the support. The authors also acknowledge NVIDIA Corporation for the donated GPUs.

References

[1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Deep compositional question answering with neural module networks. In *CVPR*, 2016.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *ICCV*, pages 2425–2433, 2015.

[3] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio. End-to-end attention-based large vocabulary speech recognition. In *Acoustics, Speech and Signal Processing*, pages 4945–4949, 2016.

[4] A. C. Berg, T. L. Berg, H. Daume, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, et al. Understanding and predicting importance in images. In *CVPR*, pages 3562–3569, 2012.

[5] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Intell. Res.(JAIR)*, 55:409–442, 2016.

[6] B. Dai, Y. Zhang, and D. Lin. Detecting visual relationships with deep relational networks. In *CVPR*, pages 3076–3086, 2017.

[7] S. K. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*, pages 3270–3277, 2014.

[8] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From captions to visual concepts and back. In *CVPR*, pages 1473–1482, 2015.

[9] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *CVPR*, pages 1–8, 2008.

[10] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 106(2):210–233, 2014.

[11] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-class segmentation with relative location prior. *IJCV*, 80(3):300–316, 2008.

[12] A. Graves, N. Jaitly, and A.-r. Mohamed. Hybrid speech recognition with deep bidirectional lstm. In *Automatic Speech Recognition and Understanding*, pages 273–278, 2013.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[14] M. Honkala, L. M. J. Kärkkäinen, A. Vetek, and M. Berglund. Generating using a bidirectional rnn variations to music, 2016. US Patent App. 15/081,654.

[15] S. Jae Hwang, S. N. Ravi, Z. Tao, H. J. Kim, M. D. Collins, and V. Singh. Tensorize, factorize and regularize: Robust visual relationship learning. In *CVPR*, pages 1014–1023, 2018.

[16] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.

[17] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[19] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[20] Y. Li, W. Ouyang, and X. Wang. Vip-cnn: Visual phrase guided convolutional neural network. In *CVPR*, pages 1347–1356, 2017.

[21] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang. Scene graph generation from objects, phrases and region captions. In *ICCV*, pages 1261–1270, 2017.

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[23] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[24] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *ECCV*, pages 852–869. Springer, 2016.

[25] G. Mesnil, X. He, L. Deng, and Y. Bengio. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *Interspeech*, pages 3771–3775, 2013.

[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv*, 2013.

[27] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[28] J. Peyre, I. Laptev, C. Schmid, and J. Sivic. Weakly-supervised learning of visual relations. In *ICCV*, pages 5179–5188, 2017.

[29] B. A. Plummer, A. Mallya, C. M. Cervantes, J. Hockenmaier, and S. Lazebnik. Phrase localization and visual relationship detection with comprehensive image-language cues. In *ICCV*, pages 1928–1937, 2017.

[30] N. Prabhu and R. Venkatesh Babu. Attribute-graph: A graph based approach to image ranking. In *ICCV*, pages 1071–1079, 2015.

[31] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[32] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, pages 1745–1752, 2011.

[33] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Tran. Signal Proc.*, 45(11):2673–2681, 1997.

[34] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760. Springer, 2012.

[35] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[36] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, pages 4278–4284, 2017.

[37] P. Wang, Q. Wu, C. Shen, and A. v. d. Hengel. The vqa-machine: Learning how to use existing vision algorithms to answer new questions. In *CVPR*, pages 1173–1182, 2017.

[38] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, pages 5410–5419, 2017.

[39] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015.

[40] M. Y. Yang, W. Liao, H. Ackermann, and B. Rosenhahn. On support relations and semantic scene graphs. *ISPRS Journal of Photogrammetry and Remote Sensing*, 131:15–25, 2017.

[41] X. Yang, H. Zhang, and J. Cai. Shuffle-then-assemble: Learning object-agnostic visual relationship features. In *ECCV*, pages 36–52, 2018.

[42] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, J. Shao, and C. Change Loy. Zoom-net: Mining deep feature interactions for visual relationship recognition. In *ECCV*, pages 322–338, 2018.

[43] R. Yu, A. Li, V. Morariu, and L. Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *ICCV*, 2017.

[44] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua. Visual translation embedding network for visual relation detection. In *CVPR*, pages 5532–5540, 2017.

[45] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, 2017.