VISUAL RELATIONSHIP DETECTION WITH OBJECT SPATIAL DISTRIBUTION

Yaohui Zhu, Shuqiang Jiang, Xiangyang Li

Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
{yaohui.zhu, xiangyang.li}@vipl.ict.ac.cn, sqjiang@ict.ac.cn

ABSTRACT

Recently, object recognition techniques have been rapidly developed. Most of existing object recognition focused on recognizing several independent concepts. The relationship of objects is also an important problem, which shows in-depth semantic information of images. In this work, toward general visual relationship detection, we propose a method to integrate spatial distribution of object to facilitate visual relation detection. Spatial distribution can not only reflect positional relation of object but also describe structural information between objects. Spatial distributions are described with different features such as positional relation, size relation, shape relation, and so on. By combing spatial distribution features with visual and concept features, we establish a modeling method to make these three aspects working together to facilitate visual relationship detection. To evaluate the proposed method, we conduct experiments on two datasets, which are the Stanford VRD dataset, and a newly proposed larger new dataset which contains 15k images. Experimental results demonstrate that our approach is effective.

Index Terms— visual relationship detection, object classification, region model

1. INTRODUCTION

Recently, fast progress has been made on image recognition, including both global image classification [1, 2, 3, 4] and local object detection [5, 6], with the assistance of deep learning techniques [7] and large scale training samples [8, 9, 10]. Most of existing object detection methods focused on recognizing several independent object concepts, and localizing their positions in images. However, the relations of object have not been much exploited, which goes one step further to not only discover concepts, but also extract triplet knowledge representations from images. Investigating on this problem can facilitate in-depth understanding of images, and should be a potentially important topic in multimedia. Furthermore, relationship is useful for improving image retrieval [11, 12], image description [13, 14], object detection [15].

There exist some previous work on object relations such as spatial relationships [16] like "below", "above", for object categorization, and human-object interactions [17, 18, 19] to discover activity information from images. These works mainly focus on learning a few specific relationships such as "people-playing-musical-instrument" [17] or learn visual phrases [15] to improve individual object detection (e.g. improve "person" and "horse" detection by detecting the visual phrase of "person ride horse"). However, these works neglect general object relationship detection, which is short of diverse relation representations and is limited for wide applications.

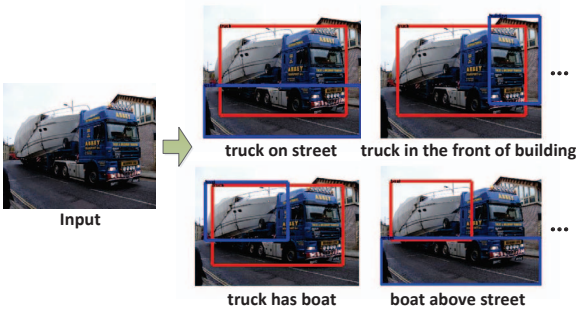
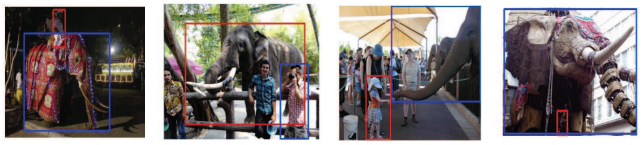


Fig. 1. Given an image, visual relationship detection is to acquire a series of relationships and bounding boxes of each two objects. e.g. in relationship of "truck has boat", red bounding boxe is "truck", and blue bounding boxe is "boat".

A more general visual relationship detection is to detect all possible relationships between objects. An object relationship in an image I can be represented by $R_{<obj_i, pre_k, obj_j>}$, where $obj_i = \{c_i, o_i\}, obj_j = \{c_j, o_j\}$, o_i, o_j are bounding boxes of obj_i, obj_j respectively, c_i, c_j belong to object category space, pre_k belongs to predicate category space. The goal of visual object detection is to extract a series of relationships $A = \{R_{<>}, R_{<>}, \cdots, R_{<>}\}$ from an image. As Fig.1 shows, the input is an image, and the output are a series of relationships such as "truck on street", "truck has boat" and bounding boxes of each two objects. It can be observed that the semantic space of possible relationship is big, and there exists unbalanced relationships. As shown in Fig.2, there exits the diverse relationships between the same two object concepts such as "person ride elephant", "elephant stand behind



(a) (b) (c) (d)

Fig. 2. (a) is "person ride elephant", (b) is "elephant stand behind person", (c) is "person feed elephant", (d) is "person inside elephant". The relationship between person and elephant is not only determined by image vision but also effected by positional relation between them, e.g. the positional relation of "person ride elephant" and "person inside elephant" are apparently different, where the position of person in "person ride elephant" is on the horse while the position is inverse in "person inside elephant".

person","person feed elephant", and rare relationships such as "person inside elephant".

To detect visual relationship needs to consider how to detect and represent objects and how to model the object relationships. Due to rapid development of object detection techniques such as R-CNN [20] and faster R-CNN [5], relatively reliable detection results can be obtained. To represent objects, visual feature such as CNN (fc7) [21] and word2vector [22, 23] features can be used, which have been validated to be effective in many visual tasks. Furthermore, the position and spatial distribution of two objects should also be included, which are important cues to better infer the object interactions. As shown in Fig.2, the positional relation of "person ride elephant" and "person inside elephant" are apparently different, where the position of person in "person ride elephant" is on the elephant while the position is inverse in "person inside elephant". In the previous work [12], to solve large relationship semantic space and unbalanced relationship, Cewu Lu et al. proposed a visual relationship detection method by learning visual appearance models with language priors. However, this approach does not consider the spatial distribution of objects.

The spatial distribution of objects can support visual relationship detection in two ways. First, different properties of object's spatial distributions may take different effects on various types of visual relationships. Second, as shown in Fig.2, spatial distributions are closely related and complementary to both visual and concept features of objects. In this paper, to facilitate visual relation detection, we propose a method with integrating object's spatial distribution. Spatial distributions are described with different features such as position relation, size relation, shape relation, and so on. By combing spatial features with visual and concept features, we establish a modeling method to make these three aspects working together. Experimentations are conducted on the Stanford VRD dataset [12] and a newly proposed larger new dataset which contains 15k images. The experimental results demonstrate that our proposed method is effective.

2. THE PROPOSED METHOD

In this section, we will frist introduce the baseline model [12], and then introduce how to represent spatial distribution of objects. Finally, the modeling method is explained in detail.

2.1. Baseline Model

We use visual relationship detection with language priors [12] as our baseline. The final relationship score is V()*L(), where V() is visual score calculated by Eq (1), and L() is language score calculated by Eq (2). $P_i(o_1)$ and $P_j(o_2)$ are the likelihood of categorizing bounding box o_1 as object category i and bounding box o_2 as category j, CNN(o_1, o_2) is the CNN features extracted from the united region of the o_1 and o_2, t_i and t_j are the two objects in a relationship, $g(x) = x$ (In [12], there is no $g()$, which is equivalent to $g(x) = x$). t_i and t_j are transformed into a word embedding space by a pre-trained word vectors model (word2vec) [23]. Convolutional neural network CNN (VGG net [1]) is trained to classify each object. Similarly, a second CNN (VGG net [1]) is trained to extract feature from the union bounding box of the two objects.

$$V(R_{<i,k,j>}|o_1, o_2) = P_i(o_1)g(W_k^v CNN(o_1, o_2) + b_k^v)P_j(o_2) \quad (1)$$

$$L(R_{<i,k,j>}) = g(W_k^l[w2vec(t_1), w2vec(t_2)] + b_k^l) \quad (2)$$

$$min\{C(W^v, W^l) + \lambda_1 L_{re}(W^l) + \lambda_2 K_{re}(W^l)\} \quad (3)$$

The Eq (3) is the objective function, where $C(W^v, W^l)$ is a rank loss function, $L_{re}(W^l)$ and $K_{re}(W^l)$ are regularization. $C(W^v, W^l)$ term enforces ground turth relationship to obtain the highest score (V()*L()). $L_{re}(W^l)$ term enforces higher likelihood occurring relationship to obtain a higher language score than lower likelihood occurring relationship, and $K_{re}(W^l)$ term enforces language semantics similar relationship to obtain the same language score. The details can be referred to [12].

2.2. Spatial distribution of objects

Our relationship detecting pipeline is illustrated in Fig.3, where a region model is designed to obtain the relationship score. As Fig.3 shows, relationship of two persons is influenced to a large extend by the positional relation of them. Therefore, the information of spatial distribution is an important issue to obtain more reliable relationships. In this paper, we describe spatial distribution of objects by using properties of regions, which contain positional relation, size relation, distance relation and shape relation.

Positional Relation (PR). The region of object is a rectangle, and can be indicated by two points. Let (x_1, y_1),

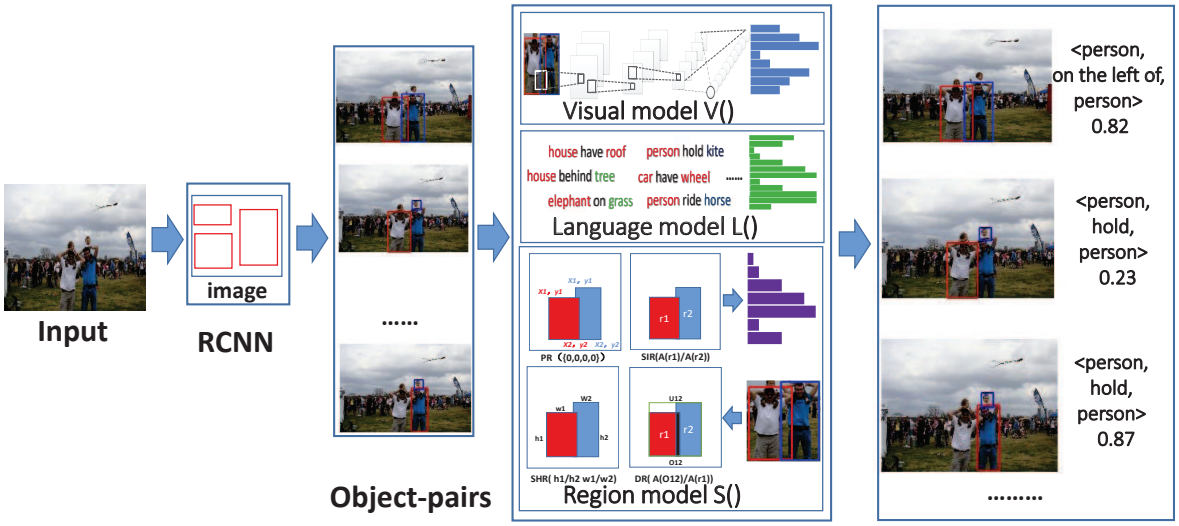


Fig. 3. An overview of our visual relationship detection pipeline. Given an image as input, RCNN [20] generates a set of object proposals. Each pair of object proposals is calculated by visual model, language model and region model. In this work, we establish a model with combing spatial distribution features with visual and concept features.

(x_2, y_2) denote the left top point and the bottom right one respectively, where $x_2 > x_1$ and $y_1 > y_2$. The position of object can be represented by four-tuples of (x_1, y_1, x_2, y_2). In this paper, we get the type of structured regions simply by comparing two four-tuples. For example, $(x_1^{r1}, y_1^{r1}, x_2^{r1}, y_2^{r1})$ and $(x_1^{r2}, y_1^{r2}, x_2^{r2}, y_2^{r2})$ are two four-tuples. The binary positional type code is $\{I(x_1^{r1} - x_1^{r2}), I(y_1^{r1} - y_1^{r2}), I(x_2^{r1} - x_2^{r2}), I(y_2^{r1} - y_2^{r2})\}$, where $I(x > 0)$ is indicator function. Consequently, we can obtain 16 types of positional relations from binary positional type code. For example, "above" mostly appears at positional type code of {0/1,1,0/1,1}. There exists a lot of predicates such as "below", "on the left of", "on the right of" and so on , which reflect positional relationships between the two objects. Therefore, the positional relation is important for some predicates of two objects.

Size Relation (SIR). Some relationships between objects can be acquired by the area ratio of two regions. Let r_1, r_2 denote two regions, and $A(x)$ denotes the area of x. The ratio value $\frac{A(r_1)}{A(r_2)}$ is quantized into an 8 dimensional discreted feature. As Fig.4(a) shows, the relationship of $< obj_1, wear, obj_2 >$ mostly appears when the size of obj_1 is much larger than obj_2, and the relationship of $< obj_1, on, obj_2 >$ mostly appears when the size of obj_1 much is smaller than obj_2.

Shape Relation (SHR). Shape of regions is measured by calculating two regions ratio of height and width. Let w_1, w_2, h_1, h_2 denote height and width of two regions respectively. The two ratio values are $\frac{w_1}{w_2}$ and $\frac{h_1}{h_2}$, which can be further quantized into a 4 dimensional discreted feature respectively. As Fig.4(b) shows, the relationship of $< obj_1, on\,the\,top\,of, obj_2 >$ mostly appears when the height of obj_1 is much less than obj_2, and the relationship of $< obj_1, stand\,on, obj_2 >$ mostly appears when the height of obj_1 is much larger than obj_2.

Distance Relation (DR). Distance between regions is measured by calculating overlap ratio in two regions and union ratio. Let o_{12} denotes intersection of r_1 and r_2, u_{12} denotes union of r_1 and r_2, U_{12} denote the minimum rectangular region containing r_1 and r_2. Overlap ratio contains two ratio value $V_{r1} = \frac{A(o_{12})}{A(r_1)}$ and $V_{r2} = \frac{A(o_{12})}{A(r_2)}$. Union ratio is $V_u = \frac{A(u_{12})}{A(U_{12})}$. Finally, V_{r1} and V_{r2} quantized in to a 11 dimensional discreted feature respectively, and V_u is quantized into a 10 dimensional discreted feature. As Fig.4(c) shows, the relationship of $< obj_1, above, obj_2 >$ mostly appears when obj_1 is far away from obj_2, and the relationship of $< obj_1, hold, obj_2 >$ mostly appears when obj_1 touch obj_2 and the area of contact is smaller compared with obj_1.

Region model. This model is to acquire a score of predicate when the type of positional relation (PR) is obtained. Let $T_s(o_1, o_2)$ denotes the type of PR between o_1 and o_2. The region score is calculated by Eq (4), where $s_i = T_s(o_1, o_2) \epsilon \{0, 1, ..., 15\}$, $S_{vec}(< o_1, o_2 >)$ is region feature vector (48 dimensions) containing SIR, SHR and DR, W^s_{k,s_i} and b^s_{k,s_i} are the learned parameters, s_i is obtained from o_1 and o_2.

$$S(k|s_i = T_s(o_1, o_2)) = g(W^s_{k,s_i} S_{vec}(< o_1, o_2 >) + b^s_{k,s_i}) \quad (4)$$

2.3. Objective function

We combine visual model, language model and region model using the rank loss function in Eq (5). To make our model more probable to choose the correct relationship, our fi-

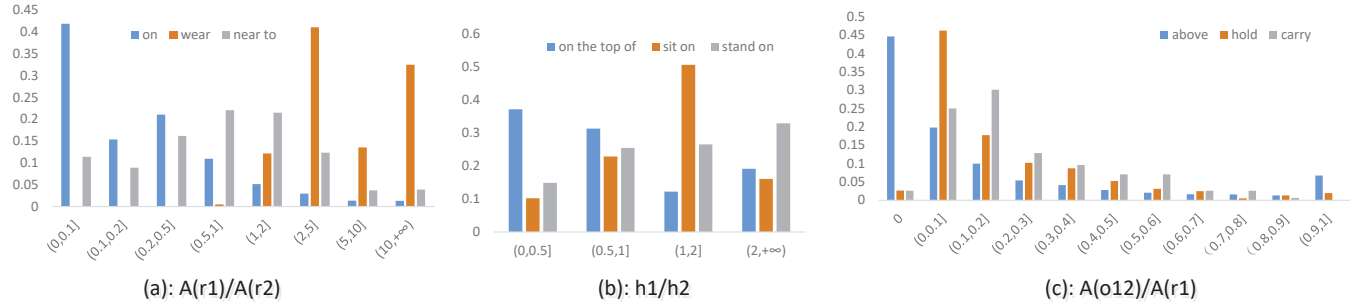


Fig. 4. The spatial distribution statistical histograms for some predicates in training data. (a) is the histogram of size-relation, (b) is the histogram of shape-relation, (c) is the histogram of distance-relation. Different relationships have different statistical distributions .

nal objective function is Eq (6) adding Eq (7) (W^s is the parameter set of $\{W^s_{k,0}, b^s_{k,0}, ..., W^s_{k,15}, b^s_{k,15}\}$). Eq (7) enforces higher likelihood occurring predicate (k) to have a higher region score than lower likelihood occurring predicate ($k^{'}$) in the same type of positional relation.

$$C(W^v,W^s,W^l) = \sum_{<o_1,o_2>} max(m_c - V(R|o_1,o_2)S(k|s_i = T_s(o_1,o_2))L(R) + \max_{R' \neq R} V(R^{'}|o_1,o_2)S(k|s_i = T_s(o_1,o_2))L(R^{'}), 0) \quad (5)$$

$$min\{C(W^v,W^s,W^l) + \lambda_1 L_{re}(W^l) + \lambda_2 S_{re}(W^s) + \lambda_3 K_{re}(W^l)\} \quad (6)$$

$$S_{re}(W^s) = \sum_{s_i} \sum_{k,k'} max(S(k^{'}|s_i = T_s(o_1,o_2)) - S(k|s_i = T_s(o_1,o_2)) + m_s, 0) \quad (7)$$

3. EXPERIMENTS

3.1. Dataset and Evaluation Metric

We use two datasets for experimental evaluations: Stanford VRD dataset [12] and ViGen15K. The Stanford VRD dataset [12] contains 5000 images with 100 object categories and 70 predicates, which has 37,993 relationships with 7,701 types (The total types of relationship are 100*100*70). We use 4000 images to the train model which contain 30,355 relationships with 6,672 types, and test on the remaining 1000 images which contain 7,638 relationships with 2,747 types. In the test dataset, 1,169 relationships with 1,029 types occur, but they never occur in the training set. The ViGen15K dataset is a new dataset, which is a subset of visual genome [9] containing 15,000 images with 100 object categories and 80 predicates, and has 90,748 relationships with 8,584 types (The total types of relationship are 100*100*80). We use 12,000 images to train model which contain 72,688 relationships with 7,622 types, and test on the remaining images which contain 18,060 relationships with 3,314 relationship types. In the test dataset, 1,386 relationships with 962 types occur, but they never occur in the training set.

Table 1. Results for visual relationship detection in Stanford VRD dataset [12].

	Phrase Det		Relationship Det		Predicate Det	
	R@100	R@50	R@100	R@50	R@100	R@50
VRD [12]	17.03	16.17	14.70	13.86	47.87	47.87
Cues[24]	20.70	16.89	18.37	15.08	-	-
Our VRD	17.44	15.42	14.61	12.92	48.47	48.47
Our-R	18.13	15.84	15.17	13.40	51.16	51.16
Our-S	17.09	15.12	14.52	13.01	47.33	47.33
Our-S-R	18.89	16.94	15.77	14.31	51.50	51.50

The task of visual relationship detection is to extract a set of visual of relationships and the bounding boxes of each object in an image. Since the ground truth annotations can not be exhaustively annotated with all possible relationships in an image, we use the evaluation metrics recall @100 and recall @50 which is reported in [12]. Recall @x computes the fraction of times the correct relationship is predicted in the top x confident relationship predictions. The measurement of visual relationship prediction has the following conditions: predicate detection, phrase detection, relationship detection. In predicate detection, the input is an image and set of localized objects, the task is to predict a set of possible predicates between pairs of objects without the limitations of object detection [20]. In phrase detection, the input is an image and the task is to output a set of relationship labels and localize the entire relationship as one bounding box having at least 0.5 overlap with the ground truth box. In relationship detection, the input is an image, the task is to output a set of relationship labels and localize both obj_1 and obj_2 in the image having at least 0.5 overlap with their ground truth boxes simultaneously.

3.2. Implementation Details

Our objective function Eq (7) is a quadratic closed form. We perform stochastic gradient descent iteratively. Since Eq (7) has four components which are difficult to optimize all com-

Table 2. Results for zero-shot visual relationship detection in Stanford VRD dataset [12].

	Phrase Det		Relationship Det		Predicate Det	
	R@100	R@50	R@100	R@50	R@100	R@50
VRD [12]	3.75	3.36	3.52	3.13	8.45	8.45
Cues[24]	15.23	10.86	13.43	9.67	-	-
Our VRD	1.85	1.62	1.45	1.29	10.23	10.23
Our-R	2.57	2.24	2.07	1.79	12.98	12.98
Our-S	2.40	2.29	1.96	1.85	10.35	10.35
Our-S-R	3.52	3.24	3.08	2.85	14.60	14.60

ponent in one step, we optimize each component one by one in a batch data. In $L_{re}(W^l)$, the higher likelihood occurring relationship ($R_{<i,k,j>}$) is existed in language data, and we select three types of lower likelihood occurring relationship ($R^{'}_{<i',k,j>}(i^{'} \neq i)$, $R^{'}_{<i,k',j>}(k^{'} \neq k)$,$R^{'}_{<i,k,j'>}(j^{'} \neq j)$) which do not exist in language data. In $S_{re}(W^s)$, higher likelihood occurring predicate (k) here means the appearing frequency more than 4 and the low likelihood occurring predicate ($k^{'}$) means the appearing frequency no more than 2 in the same type of PR. In $K_{re}(W^l)$, we randomly sample 500k pairs of relationships. At the test time, we use RCNN [20] to produce a set of candidate object proposals for every test image, and get a maximum score of relationship by Eq (8) for every pair of RCNN object proposals $< o_1, o_2 >$. We obtain the best performance by hill-climbing method. Finally, we set $\lambda_1 = 0.01$, $\lambda_2 = 0.01$, $\lambda_3 = 0.01$, $m_c = 0.03$, $m_l = 0.03$, $m_s = 0.03$.

$$R^* = arg \max_R V(R|o_1, o_2)S(k|s_i = T_s(o_1, o_2))L(R) \quad (8)$$

3.3. Experimental results

We compare our method with some state-of-that-art approaches. Different from [24], this work does not need to select Cues. In the process of experiment, when $g(x) = x$ in Eq (1) or Eq (2). As there is no limitation in language score (L()) or visual score (V()) in the process of solving Eq (3) or Eq (6), we get negative score in some times. In addition, the variational range of the language score (L()) or visual score (V()) is large. In order to solve this problem, we set up the function $g(x) = \frac{1}{1+e^{-x}}$ (sigmoid function). As the performance of objects detection will affect the performance of relationship detection, and we can obtain 71.9% accuracy of object classification [1]. We also implement the method in [12]. Furthermore, we compare the influence of adding region model and score scaling to experimental performance.

Our VRD is the method implemented as VRD [12] by us, Our-R is the method proposed by this paper with $g(x) = x$, Our-S is similar to Our VRD but $g(x)$ is the sigmoid function, Our-S-R is similar to Our-R but $g(x)$ is the sigmoid function. In Table 1 and Table 2, due to the performance of object classification, we can not achieve the performance of relationship detection reported in [12] especially for the performance of zeros-shot learning. Compared with Our VRD model, the performance of Our-R model achieves improvement with an increase of 2.69%, 2.75% at R@50 of predicate detection and zeros-shot learning respectively. This illustrates that spatial distribution of objects is useful to visual relationship detection. The performance of Our-S model is not improved compared to Our VRD model, while the performance of Our-S-R model is improved compared to Our-R model, which is mainly due to optimizing three components (L()*S()*V()) more difficult than two components (L()*V()) without scaling. We can alleviate this problem with using sigmoid function, because it can prevent from obtaining negative score. In relationship detection evaluation and phrase detection evaluation, the proposed method achieves slightly improvement mainly due to the influence of some errors in object detection and object classification. The experimental results of ViGen15K are shown in Table 3, by using spatial distribution of objects, the performance is also improved.

Table 3. Results for visual relationship detection in ViGen15K. 'z' note zero-shot learning task.

	Phrase Det		Relationship Det		Predicate Det	
	R@100	R@50	R@100	R@50	R@100	R@50
Our VRD	9.28	9.26	7.97	7.97	57.66	57.66
Our-S-R	9.99	9.98	8.28	8.28	58.72	58.72
Our VRD^z	0.33	0.33	0.33	0.33	7.82	7.82
Our-S-R^z	0.62	0.62	0.58	0.58	8.55	8.55

4. CONCLUSION

This paper proposes a method with the spatial distribution of object to visual relationship detection. This method is established to make spatial distribution of objects, image vision and linguistic concept work together. We conduct experiment in Stanford VRD dataset [12] and ViGen15K with 90,748 relationships. Experimental results demonstrate that our approach is effective. This work detects each relationship only using two objects, which can be regarded as local information. In further work, we will learn to detect each relationship using context information in an image.

Acknowledgement: This work was supported in part by the National Natural Science Foundation of China under Grant 61532018 and 61322212, in part by the Beijing Municipal Commission of Science and Technology under Grant D161100001816001, in part by the Lenovo Outstanding Young Scientists Program, in part by National Program for Special Support of Eminent Professionals and National Program for Support of Top-notch Young Professionals.

5. REFERENCES

[1] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[2] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

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

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[6] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, and Scott Reed, "Ssd: Single shot multibox detector," *arXiv preprint arXiv:1512.02325*, 2015.

[7] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. IEEE, 2009, pp. 248–255.

[9] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *arXiv preprint arXiv:1602.07332*, 2016.

[10] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *CVPR*. IEEE, 2010, pp. 3485–3492.

[11] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A Shamma, Michael S Bernstein, and Li Fei-Fei, "Image retrieval using scene graphs," in *CVPR*. IEEE, 2015, pp. 3668–3678.

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

[13] Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele, "Translating video content to natural language descriptions," in *ICCV*, 2013, pp. 433–440.

[14] Desmond Elliott and Frank Keller, "Image description using visual dependency representations.," in *EMNLP*, 2013, vol. 13, pp. 1292–1302.

[15] Mohammad Amin Sadeghi and Ali Farhadi, "Recognition using visual phrases," in *CVPR*. IEEE, 2011, pp. 1745–1752.

[16] Carolina Galleguillos, Andrew Rabinovich, and Serge Belongie, "Object categorization using co-occurrence, location and appearance," in *CVPR*. IEEE, 2008, pp. 1–8.

[17] Bangpeng Yao and Li Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," in *CVPR*. IEEE, 2010, pp. 9–16.

[18] Bangpeng Yao and Li Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *CVPR*. IEEE, 2010, pp. 17–24.

[19] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1775–1789, 2009.

[20] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[21] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *ECCV*. Springer, 2014, pp. 818–833.

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[23] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[24] Bryan A Plummer, Arun Mallya, Christopher M Cervantes, Julia Hockenmaier, and Svetlana Lazebnik, "Phrase localization and visual relationship detection with comprehensive linguistic cues," *arXiv preprint arXiv:1611.06641*, 2016.