Neural Message Passing for Visual Relationship Detection

Yue Hu [1] **Siheng Chen** [2] **Xu Chen** [1] **Ya Zhang** [1] **Xiao Gu** [1]

Abstract

Visual relationship detection aims to detect the interactions between objects in an image; however, this task suffers from combinatorial explosion due to the variety of objects and interactions. Since the interactions associated with the same object are dependent, we explore the dependency of interactions to reduce the search space. We explicitly model objects and interactions by an interaction graph and then propose a message-passing-style algorithm to propagate the contextual information. We thus call the proposed method neural message passing (NMP). We further integrate language priors and spatial cues to rule out unrealistic interactions and capture spatial interactions. Experimental results on two benchmark datasets demonstrate the superiority of our proposed method. Our code is available at https://github.com/PhyllisH/NMP.

1. Introduction

Visual relationship detection serves as a middle-level understanding task that bridges the gap between low-level image recognition, such as classification and object detection (Simonyan & Zisserman, 2014; Ren et al., 2015), and high-level image understanding tasks, such as image captioning (Vinyals et al., 2015), visual question answering (Antol et al., 2015). Visual relationship denotes the visually recognizable interaction between subject and object, which is defined as triplet (*subject-predicate-object*).

Assuming there are N object categories and K predicate categories, there will be N^2K relationship categories. The initial sequential mechanism treats each relationship triplet as a unique class and cannot apply to large dataset due to the explosive increase of the search space. (Lu et al., 2016) proposed a separation mechanism inferring objects and predicate separately, which reduces the complexity to $\mathcal{O}(N+K)$; however, this method leads to the missing context between objects and predicate. To address this, (Li et al., 2017b; Yin et al., 2018) proposes message passing within the relationship triplet to jointly extract features. Furthermore, (Cui et al., 2018) emphasizes global contexts by introducing the visual appearance of the surroundings; however, all the previous works ignore the relationship dependencies across relation triplets, *e.g.*, the interaction between 'bus' and 'road' is more likely to be 'park on' than 'drive on' given 'bus in the front of car' and 'car park on road'.

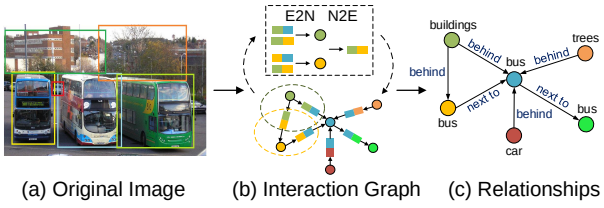

Figure 1. An interaction graph explicitly models objects and their interactions. We use message passing to propagate contextual information between objects and interactions to learn both node and edge embeddings. Pairwise relationships are detected based on edge embeddings.

To exploit this contextual information, we construct an interaction graph for each image whose nodes denote the objects and edges denote the interactions. Different from the visual graph in (Cui et al., 2018), which considers edges as an intermediate step to improve object embeddings, we explicitly use the edge embeddings to represent the relationship between objects. The edge embeddings are obtained through message passing over the interaction graph, which takes the high-order relationships into account. Considering visual appearance only may be difficult to capture all varieties of interactions. Semantic priors of objects and spatial locations are further introduced to rule out unreasonable interactions and capture spatial interactions. The main contributions of this paper are:

- We propose a novel graph-based method to explicitly model interactions between objects in an image and use a message-passing-style algorithm to capture high-order interactions;
- We introduce the word embedding of each object and the relative spatial location between pairwise objects as the

[1]Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai, China [2]Mitsubishi Electric Research Laboratories, Cambridge, Massachusetts, USA. Correspondence to: Ya Zhang <ya_zhang@sjtu.edu.cn>.

Presented at the ICML 2019 Workshop on Learning and Reasoning with Graph-Structured Data

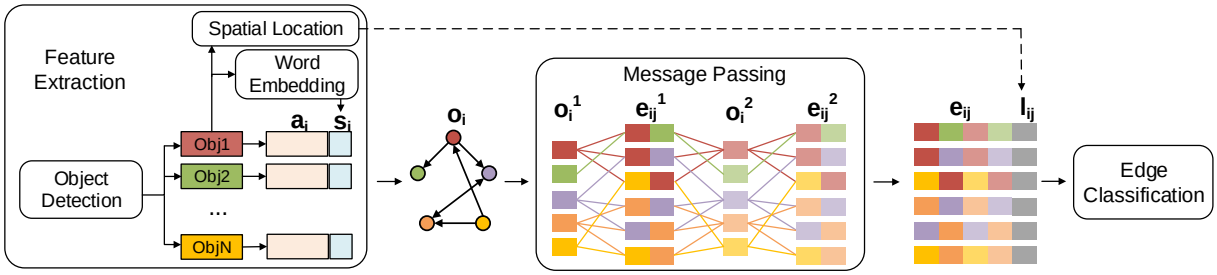

Figure 2. The overall framework of our proposed method, called neural message passing (NMP). Each detected object is represented by visual appearance and word embedding. A directed graph is built over these proposals, whose nodes denote the objects, edges denote the corresponding interactions. Message passing module is then applied to integrate contextual information. The concatenation of the enhanced interaction embeddings and the relative spatial locations are used for edge classification. Details can be found in Section 3.

complement to the visual appearance;
• The proposed method consistently outperforms the previous state-of-the-art methods on two widely used datasets.

2. Related Work

Visual relationship detection has been extensively studied in recent years. At the very beginning, (Galleguillos et al., 2008; Farhadi et al., 2010; Sadeghi & Farhadi, 2011; Ramanathan et al., 2015) assigned a unique class to each relationship triplet; however, with the increase of objects and predicates, the amount of relationship triplets is explosive. To reduce the complexity, (Lu et al., 2016) learned objects and predicates separately; however, the separate model results in the lack of context between the related components. To address this, (Yin et al., 2018) encouraged feature sharing by message passing between the three components. Furthermore, the global contexts are introduced by utilizing graph. (Liang et al., 2017) sequentially predicted interactions based on the semantic-action graph of the entire training set. (Cui et al., 2018) enhanced object embeddings by aggregating the visual appearance of the surroundings in the visual graph. However, the interaction embedding was ignored in the previous works. Instead, we explicitly model both interaction and object embeddings in the interaction graph. Language priors and spatial cues are further introduced to improve the performance in (Plummer et al., 2017; Liang et al., 2018). In this work, we integrate the word embeddings and spatial location to help estimate relationship.

Graph neural networks recently have got a lot of attention and achieved significant success in various fields (Wang et al., 2017; Gilmer et al., 2017; Li et al., 2017a; Battaglia et al., 2018; Yang et al., 2018a; Niu et al., 2018; Woo et al., 2018; Kipf et al., 2018; Zhang et al., 2018), especially in social networks (Hamilton et al., 2017), knowledge graphs (Kampffmeyer et al., 2018) and human object interaction (Qi et al., 2018). In this work, we apply graph neural networks to the application of visual relationship detection.

3. Methodology

Overview. Visual relationship detection has two settings: **predicate detection** and **relationship detection**. Predicate detection aims to predict the interactions between given pairs of objects. Relationship detection aims to simultaneously detect a set of objects and predict the interactions between pairs of objects.

We use multi-cues to better represent the objects and construct a graph to learn global contexts. The interaction graph organizes objects and interactions to structured data, such that we can jointly learn object and interaction embeddings. The main challenge is to explore the high-order interactions over structured data. To achieve this, we use a message-passing mechanism similar to (Kipf et al., 2018). The intuition is that each object is influenced by the related interactions and each interaction depends on the connected objects. The overview of the framework is in Fig. 2.

Feature Extraction. The functionality of this module is to get the visual and semantic cues of the objects and the relative spatial locations between pairwise objects. When only using visual appearance, the estimation of the predicate may be difficult due to the variety of relationships. While only using language priors, the relationship prediction is vague and not specified on the state of subject and object. Both visual and word embeddings represent each individual object, we further introduce relative spatial location to capture spatial interactions, such as 'near', 'under', 'on'.

The i-th object in the image is associated with a bounding box $b_i = \{x_i; y_i; w_i; s_i\}$ and a category c_i, which are given in the predicate detection task and obtained through object detection module in the relationship detection task. To extract deep visual features of an object, we adopt VGG16 (Simonyan & Zisserman, 2014). Firstly, we feed the original image into the network. When it comes to the last convolutional layer, we apply RoI align to crop out the bounding box, which is fed into the last fully connected layers afterward. The resulting features form the visual embedding $\mathbf{a_i}$. To complement the visual information, we use

the pre-trained word2vector (Mikolov et al., 2013) to map the object category c_i into word embedding $\mathbf{s_i}$. The object embedding of the i-th object $\mathbf{o}_i = [\mathbf{a}_i; \mathbf{s}_i]$ is the concatenation of the visual embedding $\mathbf{a}_i$ and the word embedding $\mathbf{s}_i$. As for the spatial information, we adopt the idea of box regression and use box delta to get the box differences. Furthermore, we use intersection over union (iou) and normalized distance between two objects. The union bounding box of b_i and b_j is denoted as b_{ij}. $\mathbf{\Delta}(b_i, b_j)$ (Zhang et al., 2017b) is the box delta that regresses the bounding box b_i to b_j. $\text{dis}(b_i, b_j)$ (Cui et al., 2018) and $\text{iou}(b_i, b_j)$ denote the normalized distance and iou between b_i and b_j. The spatial location between subject v_i and object v_j is $\mathbf{l_{ij}} = [\mathbf{\Delta}(b_i, b_j); \mathbf{\Delta}(b_i, b_{ij}); \mathbf{\Delta}(b_j, b_{ij}); \text{iou}(b_i, b_j); \text{dis}(b_i, b_j)]$.

Graph Construction. The interaction graph contains a node set $\mathbf{V}$ and an edge set $\mathbf{E}$. Each node $v_i \in \mathbf{V}$ represents an object, which is composed of a bounding box b_i, a corresponding object category c_i and the original embedding $\mathbf{o_i}$. Each edge $e_{ij} \in \mathbf{E}$ denotes the predicate between node v_i and v_j. The relationship triplet (v_i-e_{ij}-v_j) and (v_j-e_{ji}-v_i) represent two different instances. To distinguish e_{ij} from e_{ji}, we construct a directed graph.

For predicate detection, we construct a graph for each image based on the given object pairs. The edge exists when the connected objects are paired. We have observed that most of the interactions happen between close objects. For relationship detection, we assume that object interacts with the surroundings and assign the existence score of the edge based on $\text{dis}(b_i, b_j)$ and $\text{iou}(b_i, b_j)$ through the following formula, t_1 and t_2 are two thresholds.

$$\text{exist}(e_{ij}) = \begin{cases} 1, & \text{dis}(b_i,b_j) < t_1 \text{ or } \text{iou}(b_i,b_j) > t_2, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Neural Message Passing. The functionality of the message passing module is to improve interaction embeddings by aggregating global context cues. Instead of utilizing the common graph convolutional networks (Kipf & Welling, 2017) to strengthen the node embeddings, we leverage node-to-edge and edge-to-node message passing mechanism similar to (Gilmer et al., 2017; Kipf et al., 2018) to explicitly model the node and edge embeddings. In the node-to-edge phase, each edge receives messages from the connected nodes. In the edge-to-node phase, the node embedding is updated according to the linked edge embeddings. Mathematically, the overall message passing module works as

$$\mathbf{o_i^1} = f_{\text{emb}}(\mathbf{o_i}) \quad (2)$$

$$v \to e : \mathbf{e_{ij}^1} = f_e^1\left([\mathbf{o_i^1}; \mathbf{o_j^1}]\right) \quad (3)$$

$$e \to v : \mathbf{o_i^2} = f_v^1\left([\frac{1}{d_i^{in}} \sum_{e_{ji} \in \mathbf{E}} \mathbf{e_{ji}^1}; \frac{1}{d_i^{out}} \sum_{e_{ij} \in \mathbf{E}} \mathbf{e_{ij}^1}]\right) \quad (4)$$

$$v \to e : \mathbf{e_{ij}^2} = f_e^2\left([\mathbf{o_i^2}; \mathbf{o_j^2}]\right) \quad (5)$$

$$\mathbf{e_{ij}} = f_{\text{fusion}}\left([\mathbf{e_{ij}^1}; \mathbf{e_{ij}^2}]\right) \quad (6)$$

where $\mathbf{o_i}$ is the object embedding of the i-th object and $\mathbf{e_{ij}}$ is the edge embedding of between the i-th and j-th objects. The function f_{emb} maps the original node embedding into the hidden space. Then we use the object embedding $\mathbf{o_i^1}$ to obtain edge embedding $\mathbf{e_{ij}^1}$. $[;]$ denotes concatenation. We use the concatenation rather than the sum or mean in order to distinguish the direction of the edges. d_i^{in} is the amount of edges pointing to v_i, while d_i^{out} is the amount of edges v_i pointing out; both of which are used to normalize the edge embeddings. $\mathbf{e_{ij}^1}$ only depends on two node embeddings $\mathbf{o_i}$ and $\mathbf{o_j}$, while $\mathbf{e_{ij}^2}$ leverages more global information. Afterward, the final edge embedding $\mathbf{e_{ij}}$ in Eq. (6) is a fusion of the local embedding $\mathbf{e_{ij}^1}$ and the global embedding $\mathbf{e_{ij}^2}$.

The functions f_e^1, f_v^1, f_e^2 are neural networks used for mapping between node and edge embeddings. In our experiments, we adopt two-layers fully-connected networks (MLPs) with Elu activation function which introduces non-linearity to enhance feature expression.

Edge Classification. The functionality of this module is to classify the interactions between objects. The interaction embedding is the concatenation of the final edge embedding and spatial location; that is, $\mathbf{x_{i,j}} = [\mathbf{e_{ij}}; \mathbf{l_{ij}}]$. Then, the confidence of the predicate category between the i-th and the j-th objects is $y_{ij} = \text{softmax}\left(\mathbf{W}\mathbf{x_{i,j}}\right)$, where $\mathbf{W}$ is the embedding matrix that maps interaction embeddings to match predicate categories. In our experiment, we use multi-class cross entropy loss for classification.

4. Experiments

Datasets. Visual Relationship Detection (VRD) (Lu et al., 2016) contains 5,000 images with 100 object categories and 70 predicate categories. There are 1,877 relationship triplets only exist in the test set, which is used for zero-shot evaluation. Visual Gnome (VG) (Krishna et al., 2017; Zhang et al., 2017a) contains 99,658 images with 200 object categories and 100 predicates.

Evaluation Metrics. Following (Lu et al., 2016), we use Recall@50 (R@50) and Recall@100 (R@100) as the evaluation metrics. R@n computes the fraction of true positive predicted relationships over the total annotated relationships among the top n confident predictions. Let k be the number of predicates associated with each object. Similarly to (Yu

Table 1. Predicate and relationship detection results (%) in VRD dataset. "-" denotes the results are not reported in the original paper. k denotes the number of predicates associated with each object. The total predicate category of VRD dataset is 70.

k	Methods	Predicate Det. R@50	Predicate Det. R@100	Relationship Det. R@50	Relationship Det. R@100
$k=1$	LP	47.87	47.87	13.86	14.70
	VTE	44.76	44.76	14.07	15.20
	STA	48.03	48.03	-	-
	CAI	53.79	53.79	15.63	17.39
	ViP-CNN	-	-	17.32	20.01
	Zoom-Net	50.69	50.69	18.92	21.41
	VRL	-	-	18.19	20.79
	NMP	**57.69**	**57.69**	**20.19**	**23.98**
$k=70$	DR-Net	80.78	81.90	17.73	20.88
	Zoom-Net	84.25	90.59	21.37	27.30
	DSR	86.01	93.18	19.03	23.29
	CDDN	87.57	93.76	21.46	26.14
	NMP	**90.61**	**96.61**	**21.50**	**27.50**

et al., 2017), we report R@n under various k values.

Compared with State-of-the-art Methods. We compare our proposed model **NMP** against several previous state-of-the-art methods in Table 1: **LP**(Lu et al., 2016), **VTE**(Zhang et al., 2017a), **STA**(Yang et al., 2018b), **CAI**(Zhuang et al., 2017), **DR-Net**(Dai et al., 2017), **ViP-CNN**(Li et al., 2017b), **Zoom-Net**(Yin et al., 2018), **VRL**(Liang et al., 2017), **DSR**(Liang et al., 2018), **CDDN**(Cui et al., 2018). We can see that (i) **NMP** consistently outperforms state-of-the-art methods under all settings; (ii) **NMP** outperforms **CDDN** by about 3% according to Recall@100 on predicate detection task, which shows the effectiveness of our message passing algorithm over the directed interaction graph; (iii) we improve the state-of-the-art to 96.61% and 27.50% on predicate and relationship detection tasks.

Table 3 shows the comparison of the zero-shot predicate detection task. We exclude the performance on zero-shot relationship detection task since it is very sensitive to the number of detected bounding boxes. We can see that the performance is improved by about 5% and 4.5% on R@50 and R@100 respectively, which proves the promising generalization ability of our algorithm. To further prove the ability of our algorithm, we conduct experiments on the larger dataset: VG. Similarly to the previous works, we report results on the predicate detection task in Table 4. Our algorithm achieves considerably superior performance than the previous works.

Ablation Study. We introduce message passing, visual embedding, word embedding and spatial location in our proposed network. Table 2 shows the influence of each factor to the performance. We test on both predicate and relationship detection tasks on VRD dataset. We see that (i) the spatial location and the word embedding improve the performance by around 1% and 3% respectively on predicate detection. Because the spatial information is not easy to learn from the visual appearance, and language priors can help rule out some obviously unreasonable compositions; (ii) message passing (**GRAPH**) improves the recall stably by around 3%, 2% on predicate and relationship detection, respectively. The gain in relationship detection is smaller than that of predicate detection; this may be caused by the wrongly detected objects and incomplete annotation.

5. Conclusions

In this paper, we address the lack of context between interactions in previous works. We construct an interaction graph with a message-passing mechanism to explore high-order interactions. Besides, we use visual appearance, language priors, and spatial cues to complement each other. Experimental results show that our proposed method outperforms the state-of-the-art methods on two benchmark datasets.

References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V. F., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., aglar Gülehre, Song, F., Ballard, A. J., Gilmer, J., Dahl, G. E., Vaswani, A., Allen, K. R., Nash, C., Langston, V., Dyer, C., Heess, N., Wierstra, D., Kohli, P., Botvinick, M., Vinyals, O., Li, Y., and Pascanu, R. Relational inductive biases, deep learning, and graph networks. *CoRR*, abs/1806.01261, 2018.

Cui, Z., Xu, C., Zheng, W., and Yang, J. Context-dependent diffusion network for visual relationship detection. In *2018 ACM Multimedia Conference on Multimedia Conference*, pp. 1475–1482. ACM, 2018.

Dai, B., Zhang, Y., and Lin, D. Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3076–3086, 2017.

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., and Forsyth, D. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pp. 15–29. Springer, 2010.

Galleguillos, C., Rabinovich, A., and Belongie, S. Object categorization using co-occurrence, location and appearance. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. IEEE, 2008.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chem-

Table 2. Ablation Study (%) on VRD dataset. **A** denotes the visual appearance of the object bounding box. **S** denotes the word embedding of the object category. **L** denotes the spatial location. **GRAPH** denotes contextual information through message passing.

FEATURE	PREDICATE DET.			RELATIONSHIP DET.			
	$k=1$	$k=70$		$k=1$		$k=70$	
	R@50/100	R@50	R@100	R@50	R@100	R@50	R@100
A	48.98	86.24	94.34	17.92	21.59	19.29	25.20
A+L	50.93	87.13	94.89	18.54	22.03	19.89	25.37
A+L+S	53.60	89.57	96.19	18.49	22.27	20.30	26.14
GRAPH+A	52.88	87.12	95.03	19.65	23.19	20.75	26.43
GRAPH+A+L	54.13	88.71	95.64	19.99	23.51	21.48	26.90
GRAPH+A+L+S	**57.69**	**90.61**	**96.61**	**20.19**	**23.98**	**21.50**	**27.50**

Table 3. Zero-shot predicate detection results (%) in VRD dataset. Those methods without reporting the results on zero-shot setting are excluded from comparison.

k	METHODS	PREDICATE DET.	
		R@50	R@100
$k=1$	LP	8.45	8.45
	NMP	**27.50**	**27.50**
$k=70$	DSR	60.90	79.81
	CDDN	67.66	84.00
	NMP	**72.95**	**88.44**

Table 4. Predicate detection results (%) in VG dataset. The total predicate category of VG dataset is 100.

k	METHODS	PREDICATE DET.	
		R@50	R@100
$k=1$	VTE	62.63	62.87
	STA	62.71	62.94
	NMP	**67.03**	**67.29**
$k=100$	DSR	69.06	74.37
	CDDN	70.42	74.92
	DR-NET	88.26	91.26
	NMP	**89.69**	**95.54**

istry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1263–1272. JMLR. org, 2017.

Hamilton, W., Ying, Z., and Leskovec, J. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pp. 1024–1034, 2017.

Kampffmeyer, M., Chen, Y., Liang, X., Wang, H., Zhang, Y., and Xing, E. P. Rethinking knowledge graph propagation for zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Kipf, T., Fetaya, E., Wang, K.-C., Welling, M., and Zemel, R. Neural relational inference for interacting systems. In *International Conference on Machine Learning*, pp. 2693–2702, 2018.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Machine Learning*, 2017.

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

Li, R., Tapaswi, M., Liao, R., Jia, J., Urtasun, R., and Fidler, S. Situation recognition with graph neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4173–4182, 2017a.

Li, Y., Ouyang, W., Wang, X., and Tang, X. Vip-cnn: Visual phrase guided convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1347–1356, 2017b.

Liang, K., Guo, Y., Chang, H., and Chen, X. Visual relationship detection with deep structural ranking. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Liang, X., Lee, L., and Xing, E. P. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 848–857, 2017.

Lu, C., Krishna, R., Bernstein, M., and Fei-Fei, L. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pp. 852–869. Springer, 2016.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Niu, S., Chen, S., Guo, H., Targonski, C., Smith, M. C., and Kovačević, J. Generalized value iteration networks: Life beyond lattices. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Plummer, B. A., Mallya, A., Cervantes, C. M., Hockenmaier, J., and Lazebnik, S. Phrase localization and visual relationship detection with comprehensive image-language cues. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1928–1937, 2017.

Qi, S., Wang, W., Jia, B., Shen, J., and Zhu, S.-C. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 401–417, 2018.

Ramanathan, V., Li, C., Deng, J., Han, W., Li, Z., Gu, K., Song, Y., Bengio, S., Rosenberg, C., and Fei-Fei, L. Learning semantic relationships for better action retrieval in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1100–1109, 2015.

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99, 2015.

Sadeghi, M. A. and Farhadi, A. Recognition using visual phrases. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1745–1752. IEEE, 2011.

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

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

Wang, H., Shi, X., and Yeung, D.-Y. Relational deep learning: A deep latent variable model for link prediction. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Woo, S., Kim, D., Cho, D., and Kweon, I. S. Linknet: Relational embedding for scene graph. In *Advances in Neural Information Processing Systems*, pp. 558–568, 2018.

Yang, J., Lu, J., Lee, S., Batra, D., and Parikh, D. Graph r-cnn for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 670–685, 2018a.

Yang, X., Zhang, H., and Cai, J. Shuffle-then-assemble: learning object-agnostic visual relationship features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 36–52, 2018b.

Yin, G., Sheng, L., Liu, B., Yu, N., Wang, X., Shao, J., and Change Loy, C. Zoom-net: Mining deep feature interactions for visual relationship recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 322–338, 2018.

Yu, R., Li, A., Morariu, V. I., and Davis, L. S. Visual relationship detection with internal and external linguistic knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1974–1982, 2017.

Zhang, C., Ren, M., and Urtasun, R. Graph hypernetworks for neural architecture search. *CoRR*, abs/1810.05749, 2018. URL http://arxiv.org/abs/1810.05749.

Zhang, H., Kyaw, Z., Chang, S.-F., and Chua, T.-S. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5532–5540, 2017a.

Zhang, J., Elhoseiny, M., Cohen, S., Chang, W., and Elgammal, A. Relationship proposal networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5678–5686, 2017b.

Zhuang, B., Liu, L., Shen, C., and Reid, I. Towards context-aware interaction recognition for visual relationship detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 589–598, 2017.