# Relationship Detection Based on Object Semantic Inference and Attention Mechanisms

Liang Zhang
Xidian University, Shanghai BNC
Xi'an, Shaanxi, P.R.China
liangzhang@xidian.edu.cn

Shuai Zhang
Xidian University
Xi'an, Shaanxi, P.R.China
szhang_7@stu.xidian.edu.cn

Peiyi Shen
Xidian University
Xi'an, Shaanxi, P.R.China
pyshen@xidian.edu.cn

Guangming Zhu
Xidian University
Xi'an, Shaanxi
gmzhu@mail.xidian.edu.cn

Syed Afaq Ali Shah
Murdoch University
Afaq.Shah@murdoch.edu.au

Mohammed Bennamoun
School of Computer Science and Software Engineering (CSSE), The University of Western Australia
mohammed.bennamoun@uwa.edu.au

## ABSTRACT

Detecting relations among objects is a crucial task for image understanding. However, each relationship involves different objects pair combinations, and different objects pair combinations express diverse interactions. This makes the relationships, based just on visual features, a challenging task. In this paper, we propose a simple yet effective relationship detection model, which is based on object semantic inference and attention mechanisms. Our model is trained to detect relation triples, such as *<man ride horse>, <horse, carry, bag>*. To overcome the high diversity of visual appearances, the semantic inference module and the visual features are combined to complement each others. We also introduce two different attention mechanisms for object feature refinement and phrase feature refinement. In order to derive a more detailed and comprehensive representation for each object, the object feature refinement module refines the representation of each object by querying over all the other objects in the image. The phrase feature refinement module is proposed in order to make the phrase feature more effective, and to automatically focus on relative parts, to improve the visual relationship detection task. We validate our model on Visual Genome Relationship dataset. Our proposed model achieves competitive results compared to the state-of-the-art method MOTIFNET.

## CCS CONCEPTS

• **Computing methodologies → Scene understanding**; *Semantic networks*; Image representations.

## KEYWORDS

Relationship detection; Semantic module; Attention mechanism; Feature refinement


*ICMR '19, June 10–13, 2019, Ottawa, ON, Canada*

ACM ISBN 978-1-4503-6765-3/19/06…$15.00
https://doi.org/10.1145/3323873.3325025

**ACM Reference Format:**
Liang Zhang, Shuai Zhang, Peiyi Shen, Guangming Zhu, Syed Afaq Ali Shah, and Mohammed Bennamoun. 2019. Relationship Detection Based on Object Semantic Inference and Attention Mechanisms. In *International Conference on Multimedia Retrieval (ICMR '19), June 10–13, 2019, Ottawa, ON, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3323873.3325025

## 1 INTRODUCTION

Identifying relationships between objects in an image is a fundamental but challenging task. The identification of relationships can be applied to high level visual tasks such as image retrieval [3, 12, 19] and image captioning [4, 23]. The image can then be interpreted with a set of relationship triples, and each relationship is made up of three elements. Subject and object are individual instances, *e.g., man, face, hand, shoe, sock etc.*, and a predicate is used to identify the pair-wise relationship between subject and object, *e.g., in, of, holding*.

Current state-of-the-art methods [2, 5, 10, 11, 13, 17, 18, 22, 25, 26] follow the pipeline of object detection [20]. A general approach uses a message passing structure [7–9, 22, 24], which **first** output a list of scored object instances and their corresponding feature maps, named object features. **Then**, based on the detected object pairs of sub-regions, sub-region features are extracted, which is also called phrase features. **Finally**, these two kinds of features communicate with each other through the message passing structure.

We propose a novel relationship detection model, which considers the semantics of the subject and object. The object and subject features, as well as the phrase feature are refined based on attention mechanisms. In the object feature refinement module, the global information, especially the features of the other objects, is used. In the phrase feature refinement module, an attention mechanism is also used to extract the regions of interested object and subject. This can remove the influence of the background in the sub rectangle region. Encoding the global context for each individual instance also provides more clues for object classification.

In the commonly used dataset, such as Visual Genome [6], we find that the categories of the subject and object are statistically related. This hints that the semantic relationship may supply enough information to exploit the relationship detection. Furthermore, the phrase region represents the area covering both subject and object, which contains redundant information, especially when the

subject and object are far away from each others, such as *<man, flying, kite>*, where the phrase region not only contains the useful information of man and kite, but also contains a lot of useless information.

In summary, our contributions are: **(1)** We propose a novel relationship detection approach, where the object feature is refined through an attention mechanism of the global image context, and the semantic inference is used to enhance the relationship detection accuracy. In addition, the phrase region feature is adaptively refined based on the phrase context. **(2)** In the object feature refinement module, according to the geometrical relationship of the detected objects in the image, each object's feature is refined adaptively based on the neighbouring objects. **(3)** In the phrase feature refinement module, by learning the attention map of the union region, this module focuses the attention on some specific parts to facilitate the predicate recognition, which can reduce the huge variability of the phrase region.

## 2 METHOD

An overview of our model is shown in Figure 1. For a given image, our main goal is to generate an accurate relationship triple, such as *<man, playing, skateboard>* and at the same time, localize the precise bounding box positions of the object and subject. In the pipeline of our proposed model, **first**, a set of candidates are generated by Faster-RCNN, which is a popular deep learning module to detect objects. **Second**, based on the bounding boxes and categories of the objects that are detected by Faster-RCNN, the object feature refinement module (the output of this module is called object refinement feature in the following) is used to enhance the object feature representation through the contextual information of all the other objects in the image. **Third**, the paired subject and object is selected, which contains the features of the subject refinement feature, object refinement feature, subject category, object category, and the phrase feature. **Fourth**, in order to form a comprehensive feature for the relationship detection, we concatenate the outputs of the semantic inference module, the phrase feature refinement module and the subject/object refinement features. In the semantic inference module, the categories of the subject and object are fused through a fully connected layer. In the phrase feature refinement module, an attention mechanism is used to adaptively refine the phrase feature according to its context. **Finally**, the concatenated features are used to produce a predicate classification score. In the following subsections, we discuss these modules in detail.

### 2.1 Object Feature Refinement Module

**Object Detection** pipelines serve as a basic block for visual relationship detection. In this work, we use Faster R-CNN [20] to locate a set of candidate objects. Each candidate object comes with a bounding box and an appearance feature, which are used for the Object Feature Refinement Module.

**Object Feature Refinement Through Attention Mechanism**. Specifically, we first use the output of conv5_3 of VGG16, and adopt RoI-align to generate the object region feature. Then this feature is input into two fully connected layers and it outputs a feature vector of 4096 dimensions $f_n$.

$$P_{mn} = \left( log\left(\frac{|x_m - x_n|}{w_m}\right), log\left(\frac{|y_m - y_n|}{h_m}\right), log\left(\frac{w_m}{w_n}\right), log\left(\frac{h_m}{h_n}\right)\right) \tag{1}$$

where $P_{mn}$ corresponds to the relative position representation of object $m$ to object $n$, where $x$, $y$, $w$ and $h$ denote the x, y-coordinates, width, and height of the object bounding box, respectively.

$$w_G^{mn} = ReLU\left(W_G \cdot Emb\left(P_{mn}\right)\right) \tag{2}$$

The relative position vector is first mapped to a high-dimension by adopting the method in [21], then it is transformed by a parameter matrix $W_G$ and activated by a ReLU unit to get a scalar weight.

$$w_A^{mn} = \frac{(W_1 f_m) \cdot (W_2 f_n)^T}{\sqrt{d_k}} \tag{3}$$

where $W_1$ and $W_2$ are transformation parameters.

The final weight is obtained by combining $w_G^{mn}$ and $w_A^{mn}$ together, as follows.

$$w_{mn} = \frac{w_G^{mn} \cdot exp\left(w_A^{mn}\right)}{\sum_{k=1}^{N} w_G^{kn} \cdot exp\left(w_G^{kn}\right)} \tag{4}$$

In Eq.4, we compute all the weights to object $n$, and then get the final refinement features $c_n$ for object $n$.

$$c_n = f_n + \sum_{m=1}^{N} \left(w^{mn} \cdot f_m\right) \tag{5}$$

### 2.2 Semantic Inference Module

We use a simple and effective method . The word vectors can denote the embedded semantic context between different words in a semantic space [14, 16]. Prior-work [14] calculates the cosine distance to determine the similarity between different words in the embedded word space. But since there can be multiple relationships between a pair of objects, the cosine distance therefore cannot adequately express the variety of relationships. In our proposed work, the semantic spatial distance is also used to express the correlation between subject and object. In addition, we use instead the Hadamard Product to express the correlation in the embedded word space.

**first**, we take the categories of subject and object to generate feature embedded vectors. Specifically, we split each embedded word vector into two parts, the first represents the semantic vector of the subject and the second corresponds to the representation of the object. **Second**, according to the candidate pair, the correlation between the subject embedded vector and the object embedded vector is computed by the Hadamard Product operation. **Finally**, the correlation distance is transformed by a Fully-Connected layer to produce a semantic feature vector. We individually experiment with this semantic prediction module and the results are shown in Table 3. We use this model as a baseline to further explore the efficiency of our proposed method.

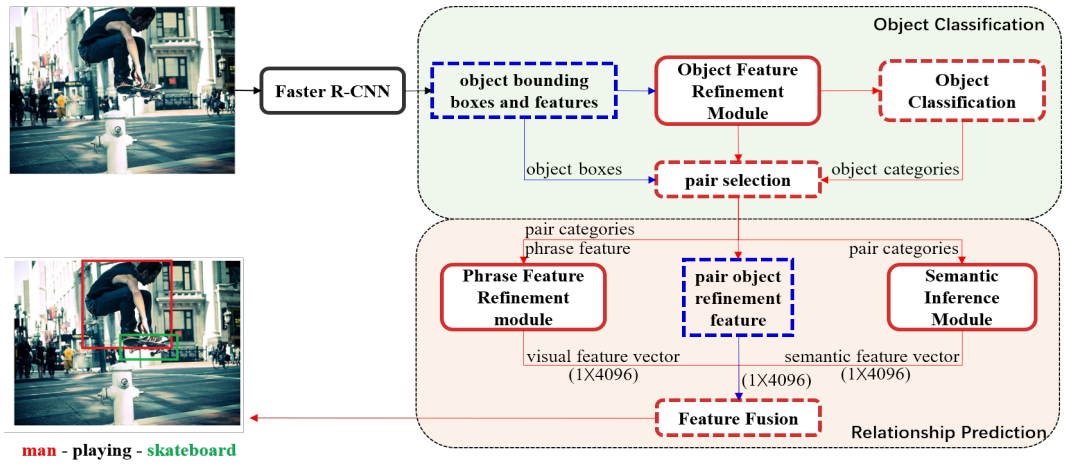


**Figure 1: Overview of the proposed object relationship detection framework**

## 2.3 Phrase Feature Refinement Module

We also extract the effective visual appearance information to promote the final learning task. Specifically, phrase proposals are constructed to express a relationship triple as in [2, 9, 22]. Each phrase proposal is a box region which covers both subject and object. Due to the complexity of the scene, the meanings are totally different when they produce different relationships. For example, *<man riding horse>* and *<man standing beside horse>* are two different triples because of the different surrounding appearance information. When the surrounding context is considered, more noisy information is prone to be added. We design an attention mechanism to automatically focus on the important parts of the union area. Our approach is shown in Figure 2.

**First**, for each candidate pair of objects, the union area of subject and object is fed into a CNN to extract an $L \times L \times C$ dimensional appearance feature map $X$, which is used to represent the predicate, where $L$ is the spatial size of the feature map while $C$ is the number of feature channels. **Second**, in order to enrich the local feature and produce an enhanced phrase feature, we use the low rank attentional pooling operation to approximate the second-order pooling [1] on union region feature map. Specifically each phrase feature map $X$ is used to produce a bottom-up attention score with $h = ReLU(Xb) \in R^{L \times L}$. We then use the score $h$ to compute a weight-average feature $X^{'}$ with $X^{'} = Xh \in R^{L \times L \times C}$. **Third**, we query the weight-average feature $X^{'}$ to decide if each $L \times L$ image region belongs to the subject or object or none of them. It is computed as follows:

$$X_s = ReLU(X^{'} \cdot Emb(S)) \tag{6}$$

$$X_o = ReLU(X^{'} \cdot Emb(O)) \tag{7}$$

where $S$ and $O$ are the subject and object category vectors, respectively. By embedding the category vector into $C$ dimension, the class-specific attention feature map can be generated by the dot product. $X_s$ and $X_o$ denote the attention over the subject and object. **Finally**, we use element-wise multiplication on the subject attention map, object attention map and the union feature map to produce the refined phrase feature, and two fully-connected layers transform it to a 4096 dimensional vector.

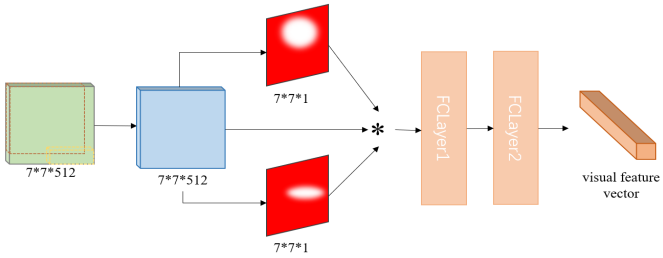


**Figure 2: Our proposed union feature refinement module.**

## 3 EXPERIMENTS

We evaluate our method on the Visual Genome (VG) dataset [6].

### 3.1 Task Definition

Similar to [13], we make three task settings for evaluation: **(1) predicate classification** (PredCls): gives the labels and locations of both the subject and object, the model only focuses on predicate classification. **(2) scene graph classification** (SGCls): gives the locations of both the subject and object, the model needs to perform object classification for the located objects, and then classify each pairwise relationship. **(3) scene graph detection** (SGDet): outputs a set of relation triples *<subject, predicate, object>*, which requires that the IoU of detected subject with its ground truth box, and the IoU of detected object with its ground truth box are bother bigger than 0.5. Since the annotated relationships are incomplete in the dataset, **recall@K** is adopted as evaluation metrics.

### 3.2 Comparison with state-of-the-art

We compare with the following models which use the same dataset partition criteria [22] on VG dataset. The models that we use for

**Table 1: Experimental results of different methods on VG[6]. The results of other methods and our proposed model are included. We report per type predicate classification accuracy with recall rates.**

| Model | Predicate Classification | | Scene Graph Classification | | Scene Graph Detection | |
|---|---|---|---|---|---|---|
| | Recall@50 | Recall@100 | Recall@50 | Recall@100 | Recall@50 | Recall@100 |
| VRD[13] | 27.9 | 35.0 | 11.8 | 14.1 | 0.3 | 0.5 |
| MESSAGE PASSING[22] | 44.8 | 53.0 | 21.7 | 24.4 | 3.4 | 4.2 |
| ASSOC EMBED[15] | 54.1 | 55.4 | 21.8 | 22.6 | 8.1 | 8.2 |
| VRL[10] | - | - | - | - | 12.5 | 13.4 |
| MOTIFNET[25] | **65.2** | **67.1** | 35.8 | 36.5 | **27.2** | **30.3** |
| Attention | 65.0 | **67.1** | **36.3** | **37.1** | 26.6 | 29.5 |

**Table 2: Ablation analysis of our proposed model. B is Baseline Model, V is visual module. We record each model's relative improvement to the Baseline model.**

| Model | Predicate Classification | | Scene Graph Classification | | Scene Graph Detection | | |
|---|---|---|---|---|---|---|---|
| | Recall@50 | Recall@100 | Recall@50 | Recall@100 | Recall@50 | Recall@100 | mean |
| $B$ | 58.2 | 62.3 | 31.6 | 33.3 | 24.9 | 28.3 | 39.76 |
| $B+V$ | +5.6 | +2.4 | +0.5 | +0.4 | +0.8 | +1.0 | +1.78 |
| $B+V_{hard}$ | +6.0 | +3.9 | +2.0 | +1.4 | +1.2 | **+1.4** | +2.65 |
| $B+V_{soft}$ | +6.5 | +4.5 | +4.5 | +3.7 | +1.4 | +0.9 | +3.58 |
| $Attention$ | **+6.8** | **+4.8** | **+4.7** | **+3.8** | **+1.7** | +1.2 | **+3.83** |

**Table 3: Ablation studies of object feature refinement type.**

| Exp | Weight Type | | Scene Graph Classification | |
|---|---|---|---|---|
| | Visual | Geometry | Recall@50 | Recall@100 |
| 1 | | | 34.7 | 35.6 |
| 2 | | √ | 35.9 | 36.8 |
| 3 | √ | √ | **36.3** | **37.1** |

comparison include **VRD** [13], **MESSAGE PASSING** [22], **ASSOC EMBED** [15], **VRL** [10] and recently introduced the state-of-the-art **MOTIFNET** [25]. The results are listed in Table 1. The **Attention** model is our final model, which combines the object feature refinement module, the semantic inference module and the phrase feature refinement module together. From Table 1 one can note that: **(1)** The results of **VRD** [13] and **MESSAGE PASSING** [22] are quite poor. This is due to the fact that a great number of relation types and the imbalanced examples distribution make it difficult for their techniques to identify the predicates using either the linguistic or the visual cues. **(2) MOTIFNET** [25] uses bidirectional LSTM to encoding context information for both object and relation, in the mean time, they explore statistics in VG dataset, which is benfical to the specific dataset. **(3)** Our final model achieves a relative gain on Scene Graph Classification task, which indicates that our object feature refinement within images is benefited for the classification task.

### 3.3 Ablation Studies

In our ablation studies, we give a detailed comparison of combinations of different modules which construct our proposed model. The models are explained as follows.

**Baseline Model (*B*)**, which directly our semantic inference module to predict the relationships.

**Visual Model ($B+V$)** additionally uses the union area visual appearance without refinement.

**Hard Attention Visual Model ($B+V_{hard}$)** constructs two binary masks to represent the relative position of the subject and the object. Then by adding the spatial mask feature, the regions of subject and object are enriched.

**Soft Attention Visual Model ($B+V_{soft}$)** uses our phrase feature refinement module, which learns to focus on the specific parts of the union region feature. This model can explore more specific expressions at the instance-level.

**Attention Model (*Attention*)** is our complete model. We combine the object feature refinement module, the semantic inference module and the phrase feature refinement to jointly predict relationships.

We also perform an ablation study to validate the effectiveness of the object feature refinement module and the different refining strategy for object feature refinement. Our results are listed in Table 3. The first experiment (Exp.1) removes the object feature refinement operation after Faster R-CNN, and directly uses the object detection feature. In Exp.2 we only use the geometrical relationship of the pair boxes to refine each object feature. Exp.3 shows the joint use of the geometrical and visual information to determine the fusion weight for object feature refinement.

## ACKNOWLEDGMENTS

The work partitally was supported by Shaanxi province key research and development program (No. 2018ZDXM-GY-036 ) and Shanghai Science and Technology Committee (No 17511104202).

## REFERENCES


[1] Joao Carreira, Rui Caseiro, Jorge Batista, and Cristian Sminchisescu. 2012. Semantic segmentation with second-order pooling. In *European Conference on Computer Vision*. Springer, 430–443.

[2] Bo Dai, Yuqi Zhang, and Dahua Lin. 2017. Detecting visual relationships with deep relational networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* 2017-Janua (2017), 3298–3308. https://doi.org/10.1109/CVPR.2017.352 arXiv:1704.03114

[3] J. Johnson, R. Krishna, M. Stark, L. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. 2015. Image retrieval using scene graphs. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 00. 3668–3678. https://doi.org/10.1109/CVPR.2015.7298990

[4] Andrej Karpathy and Li Fei-Fei. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 3128–3137.

[5] Alexander Kolesnikov, Christoph H. Lampert, and Vittorio Ferrari. 2018. Detecting Visual Relationships Using Box Attention. *CoRR* abs/1807.02136 (2018).

[6] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.

[7] Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiao'ou Tang. 2017. ViP-CNN: Visual phrase guided convolutional neural network. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* 2017-Janua (2017), 7244–7253. https://doi.org/10.1109/CVPR.2017.766 arXiv:1702.07191

[8] Yikang Li, Wanli Ouyang, Bolei Zhou, Yawen Cui, Jianping Shi, and Xiaogang Wang. 2018. Factorizable Net: An Efficient Subgraph-based Framework for Scene Graph Generation. *CoRR* abs/1806.11538 (2018). arXiv:1806.11538 http://arxiv.org/abs/1806.11538

[9] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. 2017. Scene Graph Generation from Objects, Phrases and Region Captions. *Proceedings of the IEEE International Conference on Computer Vision* 2017-Octob (2017), 1270–1279. https://doi.org/10.1109/ICCV.2017.142 arXiv:1707.09700

[10] Xiaodan Liang, Lisa Lee, and Eric P. Xing. 2017. Deep variation-structured reinforcement learning for visual relationship and attribute detection. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* 2017-Janua, 1 (2017), 4408–4417. https://doi.org/10.1109/CVPR.2017.469 arXiv:1703.03054

[11] Wentong Liao, Shuai Lin, Bodo Rosenhahn, and Michael Ying Yang. 2017. Natural Language Guided Visual Relationship Detection. *CoRR* abs/1711.06032 (2017).

[12] Dahua Lin, Sanja Fidler, Chen Kong, and Raquel Urtasun. 2014. Visual semantic search: Retrieving videos via complex textual queries. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2014), 2657–2664. https://doi.org/10.1109/CVPR.2014.340

[13] Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Fei-Fei Li. 2016. Visual Relationship Detection with Language Priors. *CoRR* abs/1608.00187 (2016), 852–869. arXiv:1608.00187 http://arxiv.org/abs/1608.00187

[14] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013). arXiv:1301.3781 http://arxiv.org/abs/1301.3781

[15] Alejandro Newell and Jia Deng. 2017. Pixels to graphs by associative embedding. In *Advances in neural information processing systems*. 2171–2180.

[16] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[17] Julia Peyre, Ivan Laptev, and Cordelia Schmid. 2018. Detecting rare visual relations using analogies. *arXiv 2018* (2018). arXiv:arXiv:1812.05736v1

[18] François Plesse, Alexandru Ginsca, Bertrand Delezoide, and Françoise Prêteux. 2018. Visual Relationship Detection Based on Guided Proposals and Semantic Knowledge Distillation. (2018). arXiv:1805.10802 http://arxiv.org/abs/1805.10802

[19] Vignesh Ramanathan, Congcong Li, Jia Deng, Wei Han, Zhen Li, Kunlong Gu, Yang Song, Samy Bengio, Chuck Rosenberg, and Li Fei-Fei. 2015. Learning semantic relationships for better action retrieval in images. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 1100–1109.

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (2017), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031 arXiv:1506.01497

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. Nips (2017). https://doi.org/10.1017/S0140525X16001837 arXiv:1706.03762

[22] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* 2017-Janua (2017), 3097–3106. https://doi.org/10.1109/CVPR.2017.330 arXiv:1701.02426

[23] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring Visual Relationship for Image Captioning. *CoRR* abs/1809.07041 (2018). arXiv:1809.07041 http://arxiv.org/abs/1809.07041

[24] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao, and Chen Change Loy. 2018. Zoom-Net: Mining Deep Feature Interactions for Visual Relationship Recognition. 1 (2018). arXiv:1807.04979 http://arxiv.org/abs/1807.04979

[25] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2017. Neural Motifs: Scene Graph Parsing with Global Context. *CoRR* abs/1711.06640 (2017). arXiv:1711.06640 http://arxiv.org/abs/1711.06640

[26] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. 2017. Visual Translation Embedding Network for Visual Relation Detection. *CoRR* abs/1702.08319 (2017). arXiv:1702.08319 http://arxiv.org/abs/1702.08319