



Scene Graph Prediction with Concept Knowledge Base

Runqing Miao^(✉) and Qingxuan Jia

Beijing University of Posts and Telecommunications, Beijing, China
tsingm@bupt.edu.cn

Abstract. Image understanding is an emerging research direction in computer vision, and scene graphs are the most mainstream form of understanding. A scene graph is a topological graph with objects in the scene as nodes and relationships as edges, used to describe the composition and semantic association of objects in an image scene. Scene graph prediction requires not only object detection, but also relationship prediction.

In this work, we propose a scene graph prediction method based on a conceptual knowledge base, which uses the condensed human understanding stored in the knowledge base to assist the generation of the scene graph. We designed a simple model to fuse image features, label features and knowledge features. Then the data filtered by the model is used as the input of the classic scene graph generation model, and better prediction results are obtained. Finally, we analyzed the reasons for the slight increase in the results, and summarized and prospected.

Keywords: Scene graph · Scene graph prediction · Knowledge graph

1 Introduction

Computer vision has always been an important branch of artificial intelligence research. Classic tasks like image recognition [6] and object detection [19] have a variety of models and methods. As far as we know, such tasks can be classified as intuitive regression problems, which is predicted directly by the model trained from labeled supervised data. Due to the subjectivity of human cognition, there are still many deficiencies in subjective image comprehension tasks such as image captioning [7, 28] and VQA (visual question answering) [20, 26].

The scene description corresponding to the image must be constructed accurately, not just the discrete marks of objects in the image in image understanding tasks. Therefore, scene graph prediction as an intermediate task has gradually become a new direction in the field of computer vision. If we define it mathematically, a node (G) in the scene graph stands for an object (O) in the visual

Supported by Major Project of the New Generation of Artificial Intelligence (No. 2018AAA0102900).

scene, and the node contains the object’s classification label (L) and the visual feature bounding box (B). While edges stand for the cognitive relationships (R) between objects, which is mainly divided into the following three types: spatial, relationships and logical relationships. At the same time, because the scene graph has the network structure of the topology graph, it is very similar to the knowledge graph [9], both are often closely related to graph neural network [24] when modeling.

While scene graph prediction [5, 10, 23, 25] have a number of methodological studies as a field, on the contrary almost no related datasets, only Visual Genome [11] has been widely recognized because of the hard work of annotation on relation between objects. Visual Genome has 1.3 million objects and 1.5 million relations in 108k images. Among them, 21 objects, 17 relations, 16 attributes are marked in every image. Although this dataset has a huge amount of data scale, there is a clear long-tail distribution in the labeling of relationships, the most common relationships, including “*on*”, “*has*”, “*in*” and “*wearing*” accounted for more than 80% of the total relationship notes. Another problem is that there are some unlabeled or duplicate relationships in the data set. Several state-of-the-art networks such as Neural Motifs [29] accurately match the Visual Genome from the perspective of data fitting. Although this scheme can obtain satisfactory recall rate, due to the limitation of bias of data set annotation itself, it is not accurate and consistent with human cognition, and the real relationship in the image can be incorporated into the scene graph.

In order to remedy the inaccurate scene graph prediction caused by Visual Genome dataset bias, some work has been pointed out and modification schemes have been proposed, which mainly include two parts. The first is to improve the prediction model to correct the interference caused by “blind guessing” [22], that is, to predict the relationship between objects directly from the labels of objects without the visual features of objects. By means of causal reasoning, the normal predicted relationship can be subtracted from the part directly predicted by the label. The second is the optimization of dataset [16], that is, filtering and modifying the annotated relations in Visual Genome to reduce bias. An attempt was made on VG150 [25], a simplified version of Visual Genome, to identify and eliminate part of the tag directly predictive relationship through a simple network.

The above two different methods are both scene graph prediction tasks from the perspective of pure vision and try to solve the problem of ignoring image features caused by data set BIAS with different ideas. However, there are still some problems in data set Bias that are not mentioned in the above methods, such as missing and duplication of relation annotations. At the same time, we have access not only to Visual Genome or similar scene graph datasets, but also to lots of wild-world knowledge bases [1, 4, 15, 21] in research fields such as knowledge engineering and recommendation system. These knowledge bases store common-sense concepts and relationships that conform to human cognition. If the knowledge base is introduced into the scene graph prediction, it will help us to predict more cognitive relationships for the scene graph [13].

Our method proposes a more comprehensive scenario relation annotation based on existing data, using common sense knowledge provided by ConceptNet to modify Visual Genome data sets to improve the reliability of scenario graph predictions. It mainly includes: using knowledge relation to filter out repeated or unreasonable relation annotations, strengthen the relationship annotations that conform to the description of the knowledge base and give this type of annotations a higher degree of confidence. Our experiments show that the new Visual Genome, which incorporates knowledge information from ConceptNet, performs better in scene prediction tasks than the original dataset.

The structure of this paper is as follows: The second section introduces relevant references. The third section introduces the method of dataset modification. The fourth section introduces the scene graph prediction experiment. The last section summarizes and prospects (Fig. 1).

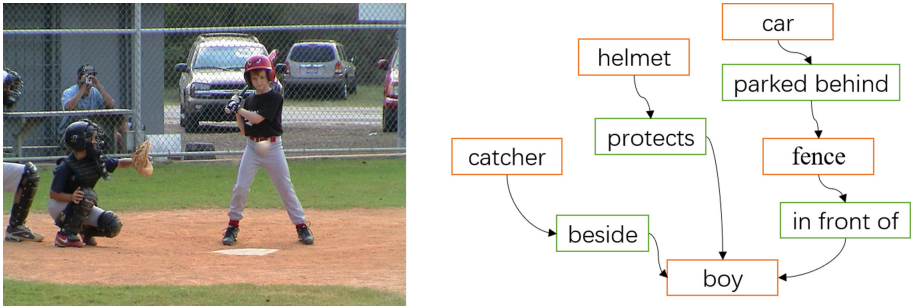


Fig. 1. Example of annotation from Visual Genome. A scene graph attempts to mark the objects in the image, including people and entities, as nodes of the graph (red), and then mark their relationships as edges between nodes (green). (Color figure online)

2 Related Works

2.1 Scene Graph Prediction

Scene graph [10, 12, 17, 25] is a concept from computer vision. It is a visual understanding form of graph structure that expresses the high level of current scene, which is primarily used to describe the objects in current scene. This is meaningful to visual comprehension tasks, like image retrieval [18], image question answering [8], image or video captioning [7, 28], and human-object relation detection [30] and image prediction [31].

In the early years, scene graph prediction [25] mainly used the message transmission of graph neural network, and iterated continuously to fuse the feature of adjacent nodes and edges. Yang et al. [27] purposed RPN and Attention Graph Convolution Network on basis. Zeller et al. [29] analysed hidden priori information in Action Genome. Then a method for deep matching of data sets is proposed, which has greatly improved the effect.

However, this goal-directed over-fitting of datasets does not necessarily predict a truly meaningful scene graph that conforms to human cognition. After Tang et al. [22] uses the model to predict, the part that directly maps from the object label to the relationship is reduced, so as to achieve the original intention of correcting the deviation of the dataset. Liang et al. [14] considered from Visual Genome, which used a simple network filtering and fitting relationship annotation to improve the quality of the dataset.

2.2 Knowledge Graph in Computer Vision

Knowledge Graph is a concept from natural language process, which is a kind of semantic network that composed of entities, relationships and attributes, usually used to represent the structure of facts [9]. Entities can be real-world abstract and figurative objects. Semantic descriptions of entities and their relationships include attributes and categories with cognitive meaning. A general knowledge graph consist of a reasoning engine and multi-layer database [9]. A knowledge base is a special kind of database, which has topological structure corresponding to human knowledge cognition system and contains different types of common sense such as statements, rules, facts and axioms. WordNet [15], Freebase [2] and DBpedia [3] are most popular knowledge bases these years. ConceptNet [21] is the knowledge graph version of the Open Mind strategy, a knowledge base of the most basic common sense possessed by humans. The original intention of the concept network was to construct a huge knowledge graph, using edges with labels (representing the type of edges) and weights (representing the credibility of the edges) to connect various words and phrases (Figs. 2 and 3).

3 Methodology

3.1 Problem Definition

Given an image I , the first step is using Faster R-CNN, which is a widely used target detection algorithm with excellent detection speed. In fact, Faster RCNN can be divided into four main contents: Conv layers, Region Proposal Networks, Roi Pooling and Classification. Faster RCNN can directly output bounding-box and tag of objects in the image:

$$[O_1, O_2, \dots, O_N] = f_{fasterrcnn}(I) \quad (1)$$

where $f_{fasterrcnn}$ corresponds to the Faster R-CNN module, and O_1, O_2, \dots, O_N are N object proposals in the image, which have a bounding box and a label. Formally, we can write:

$$O = (B, L) = [[x, y, w, h], l] \quad (2)$$

where x and y is the top-left coordinates of the bounding-box, w and h stand for width and height of the bounding-box. The next step is to apply (B_m, L_m)

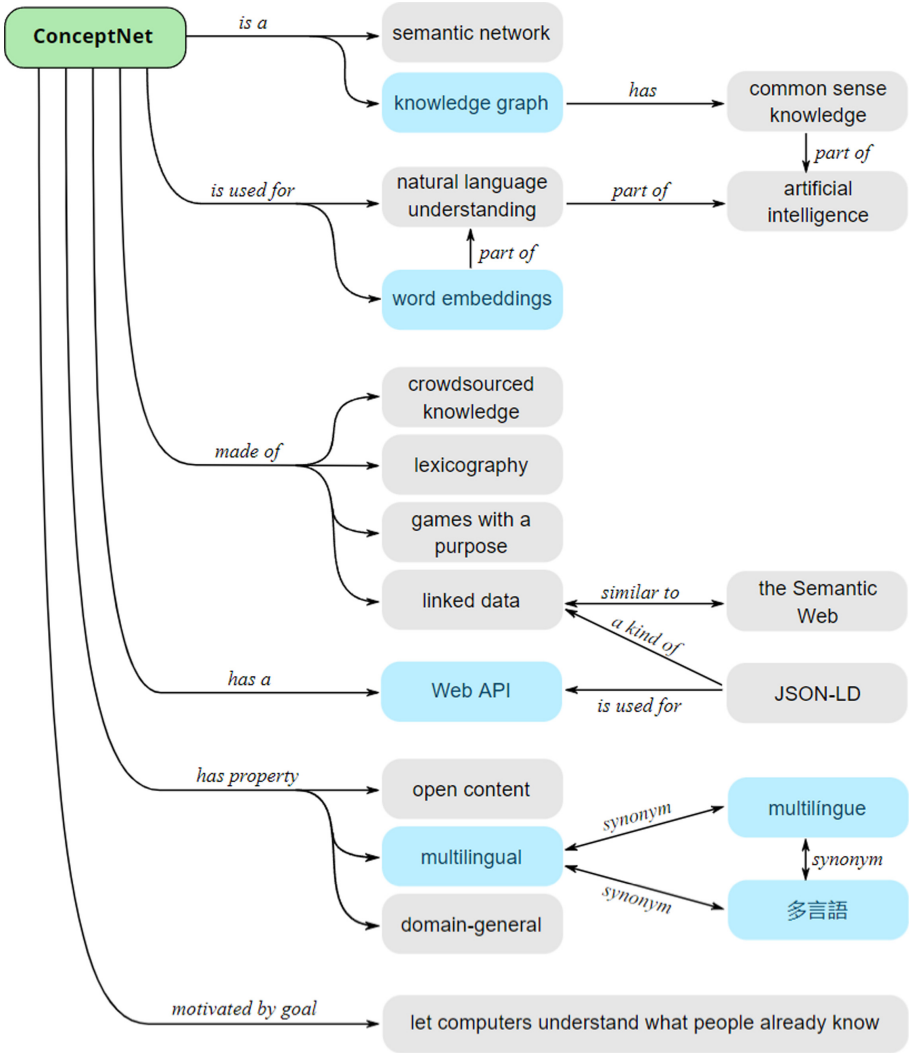


Fig. 2. Structure of ConceptNet.

and (B_n, L_n) to predict the relation $R_{m,n}$ between object O_m and O_n . To sum up, we could write this step as:

$$R_{m,n} = f((B_m, L_m), (B_n, L_n)) \tag{3}$$

In knowledge graph such as ConceptNet, the triples can be summarized as:

$$E_{tail} - E_{head} = R_{head \rightarrow tail} \tag{4}$$

where E_{head} and E_{tail} sign for the entity of head and tail. Vector differential between the two is R . Rather than high dimensional vector space or structured

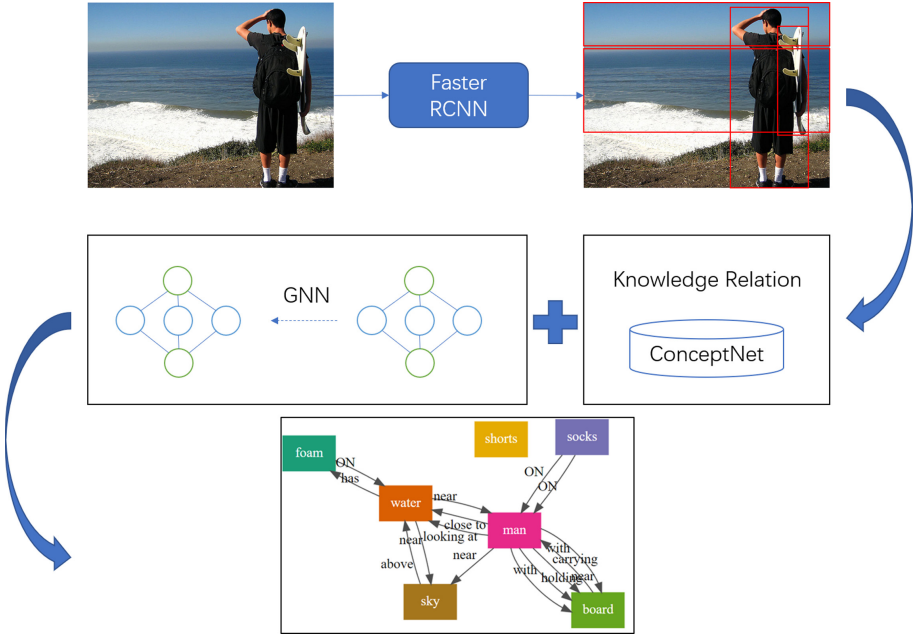


Fig. 3. Pipeline of our method. a) Faster RCNN is used for object recognition to obtain bounding-box and labels of objects in the scene. b) Train graph neural network model based on Visual Genome and knowledge base ConceptNet. c) The trained model predicts the relationship between objects and finally obtains the scene diagram corresponding to the image.

knowledge base, the representation of E_{head} and E_{tail} are equal to L_m and L_n in vision graph. $R_{head \rightarrow tail}$ and $R_{m,n}$ can also be assumed to be equivalent. After possessing the prior knowledge of the scene graph data set, the final prediction R^* is obtained by calculating the weighted sum of the initial relation of the data set and the relation retrieved from the knowledge base. So the task could be written as:

$$R^*_{m,n} = f_{visualgenome}((B_m, L_m), (B_n, L_n)) + f_{conceptnet}(E_m, E_n) \quad (5)$$

3.2 Data Source

ConceptNet is combined with word embedding (for example, word2vec) to facilitate word relevance evaluation (making the embedding of related words closer).

ConceptNet 5.5 was built from the following sources:

- Facts from Open Mind Common Sense (OMCS) and sister projects in other languages
- Use a custom parser (“Wikiparsec”) to extract information from parsing Wiktionary in multiple languages

- Purposeful games designed to gather common sense
- Open multilingual WordNet [15], Linked data representation for WordNet and its parallel multilingual project
- JMDict, A Multilingual Japanese dictionary
- OpenCyc, the upper word hierarchy provided by Cyc, represents common sense knowledge of the logic of systematic predicates
- A subset of DBpedia [3], is the fact web extracted from the Wikipedia infobox.

Combining these resources, ConceptNet contains 21 million edges and more than 8 million nodes. Its English vocabulary contains approximately 1,500,000 nodes, and it contains 83 languages, each of which contains at least 10,000 nodes.

Wiktionary is the largest input source for ConceptNet. It provides 18.1 million edges and is responsible for its vast multilingual vocabulary. However, most of the features of ConceptNet come from OMCS and a variety of purposeful games that express many different relationships between terms, Examples include PartOf (“A wheel is part of a car”) and UsedFor (“A car is used for driving”).

3.3 Data Fusion

Given an image I , use Faster-RCNN for target recognition, abstract N objects in current scene with bounding-box and label. Then pair them to $N * (n - 1)$ pairs. (O_m, O_n) and (O_n, O_m) have different meanings because of the difference between subject and object. These combinations are then entered into the extracted ConceptNet for knowledge search one by one. The retrieved relations are used to enhance or correct the original relations. Combined with the existing Visual Genome scene graph, a simple prediction model was used to output a new scene graph relationship based on knowledge base.

In Fig. 4, The input of the predictive model mainly consists of two parts. The first part comes from Visual Genome, including visual features of objects from bounding-box boxes I_m, I_n and text features of object labels V_m, V_n . The second part comes from the knowledge base, which features the knowledge K provided by ConceptNet. These features are transformed into high-dimensional vectors as the input of a multi-layer perceptron neural network, which outputs predictive relational label R^* . The network is trained on new data sets annotated according to the rules described above.

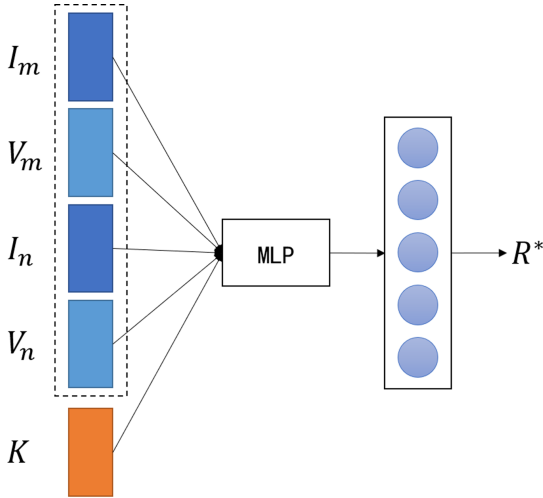


Fig. 4. Structure of visual and knowledge prediction model. I_m and I_n represent the image characteristics of subject and object. V_m and V_n represent the word vector of subject and object. K represents the relation searched from ConceptNet.

4 Experiments

4.1 Data Preparation

Visual Genome data system consists of the following types of files:

1. 108k pictures. (5.4 Million Region Descriptions, 1.7 Million Visual Question Answers, 3.8 Million Object Instances, 2.8 Million Attributes, 2.3 Million Relationships)
2. image_data.json - hyperlink and size of every image.
3. scene_graphs.json - All descriptions of scene graph in every image.
4. region_descriptions.json - Semantic relationship description in every images.

Then we batch process the data with pictures as the unit, and mark scene graph annotations in scene_graphs.json into .json files. Convert the pictures one by one according to the number. We also implemented two meaningful visualization components: graphviz-based scene diagrams and matplotlib-based natural language annotations.

4.2 Scene Graph Generation

There are three main types of scene map prediction tasks at present. As shown in Table 1, Predicate Classification (PredCls)'s input are real labels and bounding-box. Scene Graph Classification (SGCls) using unlabeled ground reality bounding-box while Scene Graph Detection(SGDet) only take the image itself

Table 1. Scene graph tasks.

Task	Recall	Input	Output
Predicate classification	PredCls	Image, Box, Object	Relation
Scene graph classification	SGCls	Image, Box	Object, Relation
Scene graph prediction/Generation	SGGen	Image	Box, Object, Relation

as input. We have verified Visual Genome and our new version under the three types of recall: R@20, R@50, and R@100.

The results of Table 2 show that the effect of the model based on the new data set is improved to some extent, because the annotation combined with knowledge is more reasonable and more consistent with human cognition. However, because of change of both train and test set, the improvement effect is not obvious, and the advantages of the new dataset are not fully reflected in the scene graph prediction index. However, this should not be taken to mean that the real quality of scenario diagram predictions is not refined, and we will verify downstream tasks in future work.

Table 2. Result of scene graph prediction.

Visual genome								
SGGen			SGCls			PredCls		
R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
25.38	32.56	37.25	35.36	38.98	39.89	59.16	65.59	67.37
Ours								
SGGen			SGCls			PredCls		
R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
25.49	32.78	37.66	35.63	39.12	40.27	59.86	66.39	68.02

4.3 Environment

In scene graph prediction, our hardware environment is consist of NVIDIA GeForce GTX 2080. There are 8 steps in training, each step takes about 8 h.

5 Conclusion

In this paper, we propose a method to combine knowledge base information with scene graph data. The original data and the knowledge base data are sent in and out of an MLP network at the same time to generate new scene graph data. The prediction accuracy of scene graphs combined with knowledge has been improved to a certain extent. However, because this improvement in conformity with the

cognitive significance is not directly presented in the data set fitting, in the future we plan to verify the effect of our method on more downstream tasks such as image captioning and visual question answer. On the other hand, we will also try to integrate more different types and domains of knowledge bases into the scene graph data, looking forward to further performance improvements.

Acknowledgement. This work was supported by Major Project of the New Generation of Artificial Intelligence (No. 2018AAA0102900).

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: a nucleus for a web of open data. In: Aberer, K., et al. (eds.) ASWC/ISWC -2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76298-0_52
2. Berant, J., Chou, A., Frostig, R., Liang, P.: Semantic parsing on freebase from question-answer pairs. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1533–1544 (2013)
3. Bizer, C., et al.: DBpedia—a crystallization point for the web of data. *J. Web Semant.* **7**(3), 154–165 (2009)
4. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, pp. 1247–1250 (2008)
5. Cohen, W.W., Sun, H., Hofer, R.A., Siegler, M.: Scalable neural methods for reasoning with a symbolic knowledge base. arXiv preprint [arXiv:2002.06115](https://arxiv.org/abs/2002.06115) (2020)
6. Dhingra, B., Zaheer, M., Balachandran, V., Neubig, G., Salakhutdinov, R., Cohen, W.W.: Differentiable reasoning over a virtual knowledge base. arXiv preprint [arXiv:2002.10640](https://arxiv.org/abs/2002.10640) (2020)
7. Gao, L., Wang, B., Wang, W.: Image captioning with scene-graph based semantic concepts. In: Proceedings of the 2018 10th ICML, pp. 225–229 (2018)
8. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6904–6913 (2017)
9. Ji, S., Pan, S., Cambria, E., Marttinen, P., Philip, S.Y.: A survey on knowledge graphs: representation, acquisition, and applications. *IEEE Trans. Neural Netw. Learn. Syst.* (2021)
10. Johnson, J., et al.: Image retrieval using scene graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3668–3678 (2015)
11. Krishna, R., et al.: Visual genome: connecting language and vision using crowd-sourced dense image annotations. arXiv preprint [arXiv:1602.07332](https://arxiv.org/abs/1602.07332) (2016)
12. Li, Z., Ding, X., Liu, T.: Constructing narrative event evolutionary graph for script event prediction. arXiv preprint [arXiv:1805.05081](https://arxiv.org/abs/1805.05081) (2018)
13. Liang, X., Hu, Z., Zhang, H., Lin, L., Xing, E.P.: Symbolic graph reasoning meets convolutions. *Adv. Neural. Inf. Process. Syst.* **31**, 1853–1863 (2018)
14. Liang, Y., Bai, Y., Zhang, W., Qian, X., Zhu, L., Mei, T.: VRR-VG: refocusing visually-relevant relationships. In: Proceedings of the IEEE/CVF ICCV, pp. 10403–10412 (2019)

15. Miller, G.A.: Wordnet: a lexical database for English. *Commun. ACM* **38**(11), 39–41 (1995)
16. Narasimhan, M., Lazebnik, S., Schwing, A.G.: Out of the box: reasoning with graph convolution nets for factual visual question answering. *arXiv preprint [arXiv:1811.00538](https://arxiv.org/abs/1811.00538)* (2018)
17. Pan, B., et al.: Spatio-temporal graph for video captioning with knowledge distillation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10870–10879 (2020)
18. Qi, M., Wang, Y., Li, A.: Online cross-modal scene retrieval by binary representation and semantic graph. In: *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 744–752 (2017)
19. Ren, H., Hu, W., Leskovec, J.: Query2box: reasoning over knowledge graphs in vector space using box embeddings. *arXiv preprint [arXiv:2002.05969](https://arxiv.org/abs/2002.05969)* (2020)
20. Shih, K.J., Singh, S., Hoiem, D.: Where to look: focus regions for visual question answering. In: *Proceedings of the 2019 CVPR*, pp. 4613–4621 (2016)
21. Speer, R., Chin, J., Havasi, C.: Conceptnet 5.5: an open multilingual graph of general knowledge. In: *Thirty-First AAAI Conference on Artificial Intelligence* (2017)
22. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: *Proceedings of the IEEE/CVF CVPR*, pp. 3716–3725 (2020)
23. Wan, H., Luo, Y., Peng, B., Zheng, W.-S.: Representation learning for scene graph completion via jointly structural and visual embedding. In: *IJCAI, Stockholm, Sweden*, pp. 949–956 (2018)
24. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Philip, S.Y.: A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **32**(1), 4–24 (2020)
25. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5410–5419 (2017)
26. Xu, K., Li, J., Zhang, M., Du, S.S., Kawarabayashi, K.I., Jegelka, S.: What can neural networks reason about? *arXiv preprint [arXiv:1905.13211](https://arxiv.org/abs/1905.13211)* (2019)
27. Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D.: Graph R-CNN for scene graph generation. In: *Proceedings of the ECCV*, pp. 670–685 (2018)
28. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4651–4659 (2016)
29. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: scene graph parsing with global context. In: *Proceedings of the CVPR*, pp. 5831–5840 (2018)
30. Zhang, M., Liu, X., Liu, W., Zhou, A., Ma, H., Mei, T.: Multi-granularity reasoning for social relation recognition from images. In: *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1618–1623. *IEEE* (2019)
31. Zhao, B., Meng, L., Yin, W., Sigal, L.: Image generation from layout. In: *Proceedings of the 2019 CVPR*, pp. 8584–8593 (2019)