Toward General Scene Graph: Integration of Visual Semantic Knowledge with Entity Synset Alignment

Woo Suk Choi
Seoul National University
wschoi@bi.snu.ac.kr

Kyoung-Woon On
Seoul National University
kwon@bi.snu.ac.kr

Yu-Jung Heo
Seoul National University
yjheo@bi.snu.ac.kr

Byoung-Tak Zhang
Seoul National University
AI Institute (AIIS)
btzhang@bi.snu.ac.kr

Abstract

Scene graph is a graph representation that explicitly represents high-level semantic knowledge of an image such as objects, attributes of objects and relationships between objects. Various tasks have been proposed for the scene graph, but the problem is that they have a limited vocabulary and biased information due to their own hypothesis. Therefore, results of each task are not generalizable and difficult to be applied to other down-stream tasks. In this paper, we propose Entity Synset Alignment(ESA), which is a method to create a general scene graph by aligning various semantic knowledge efficiently to solve this bias problem. The ESA uses a large-scale lexical database, WordNet and Intersection of Union (IoU) to align the object labels in multiple scene graphs/semantic knowledge. In experiment, the integrated scene graph is applied to the image-caption retrieval task as a down-stream task. We confirm that integrating multiple scene graphs helps to get better representations of images.

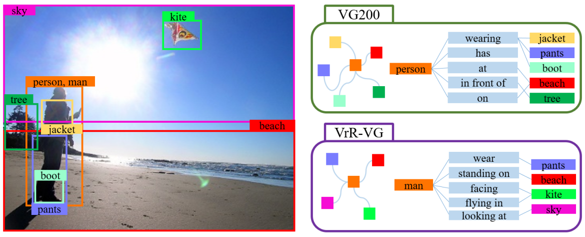

Figure 1: An example of scene graph for a common image from Visual Genome 200 (VG200) and Visually-Relevant Relationship (VrR-VG) dataset.

1 Introduction

Beyond detecting and recognizing individual objects, research for understanding visual scenes is moving toward extracting semantic knowledge to create scene graph from natural images. Starting with (Krishna et al., 2017), various studies have been proposed to generate this semantic knowledge from images (Zellers et al., 2018; Xu et al., 2017; Liang et al., 2019; Anderson et al., 2018). However, each study extracts only highly biased information from an image due to the limited vocabulary depending on their own hypothesis and the statistical bias of the dataset. For example, in (Anderson et al., 2018), the author conducted a study on extracting information of both object and attribute for each entity using 1,600 object and 400 attribute class labels. In addition, (Zellers et al., 2018; Xu et al., 2017) generate a relationship between objects in a form of triplet *(head entity - predicate - tail entity)* in an image by using 150 object and 50 predicate class labels. In (Liang et al., 2019), the author constructed a Visually-Relevant Relationships(VrR-VG) based on (Krishna et al., 2017) to mine more valuable relationships with 1600 objects and 117 predicate class labels. As such, each task defines and uses its own vocabulary, but the problem is that the vocabulary is limited. As shown in Figure 1,If some of objects in an image do not belong to the dataset-specific vocabulary, objects as well as relations are omitted frequently even though they are in an image. In addition, there are cases where the same object is defined with different vocabulary in a common image (e.g. man, person).

In this paper, we propose Entity Synset Alignment (ESA) to perform scene graph integration. With a large-scale lexical database WordNet and IoU, the ESA aligns the entity labels in scene graphs generated from each dataset. The contributions of the method proposed in this paper are as follows: 1) Scene graphs can be generated from raw image inputs, 2) integrating multiple scene graphs inferred from each dataset into one via ESA, 3) the qualitative results show that an integrated scene graph can extract richer semantic information in an image, 4) quantitative results show the significance of integrated scene graph by applying integrated scene graph to image-caption retrieval task.

2 Related Work

BottomUp-VG. Bottom-Up VG is a bottom-up attention model that extracts information of both object and attribute for each entity with 1,600 object and 400 attribute class labels from Visual Genome(VG).

VG200. VG200 introduced by (Xu et al., 2017) is a filtered version of the original VG scene graph dataset. It contains 150 object and 50 predicate class labels in 108,077 images, and consists of an average of 11.5 distinct objects and 6.2 predicates per image.

VrR-VG. Visually-Relevant Relationships (VrR-VG) introduced by (Liang et al., 2019) is constructed to highlight visually-relevant relationships using visual discriminator to learn the notion of visually-relevant.

WordNet. WordNet, a large lexical database of English, is an ontology that summarizes a relationship between words and has been integrated into the Natural Language ToolKit. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each representing intrinsic concept.

3 Method

As shown in Figure 2, we employ bottom-up attention (Anderson et al., 2018) model to generate only nodes containing information of both object and attribute, and CompTransR model to generate scene graphs from raw images. Entity Synset Alignment(ESA) integrates scene graphs generated from each dataset. We introduce a simple model, CompTransR, for scene graph generation in Section 3.1 and a scene graph integration technique, Entity Synset Alignment(ESA) in Section 3.2.

3.1 Compositional Translational Embedding

Compositional Translation Embedding combines the well-known Knowledge Graph embedding algorithms (i.e., TransR (Lin et al., 2015)) to learn the semantic relationships between two entities in a scene graph. Here, we apply transitive constraints to predict the semantic predicate labels in multiple symbolic subspaces by learning compositional representations of the relationships. As an entity feature, we extract visual, positional, and categorical features from a detected bounding box in a given image, and concatenate them into one. Then, entity features are transformed to head($\boldsymbol{h}$) and tail($\boldsymbol{t}$) features through single feed-forward neural network. The feature vectors of head and tail are projected into multiple latent relational subspaces. We aim to disentangle the semantic space of the sub-relation labels. The predicate representation $\boldsymbol{r}^s \approx \boldsymbol{t}^s - \boldsymbol{h}^s$ is defined on each latent relational space s. All $\boldsymbol{r}^s$ on the subspaces are summed out to predict predicate labels between two entities.

3.2 Entity Synset Alignment (ESA)

Algorithm 1: Entity Synset Alignment

```
Function ESA (A_obj_list, B_obj_list)
  obj_list=A_obj_list
for A_obj in A_obj_list do
    A_obj_synset = get_synset(A_obj);
    for B_obj in B_obj_list do
        B_obj_synset = get_synset(B_obj);
        if A_obj in B_obj_synset OR B_obj in
         A_obj_synset then
            iou = get_IoU(B_obj, A_obj);
            if iou is larger than 0.3 then
                pass_Flag=True;
                Break;
            end
        end
    end
    if pass_Flag is True then
        Continue;
    end
    obj_list.append(B_obj);
end
```

Entity Synset Alignment is an algorithm that integrates scene graphs generated from each dataset by using label alignment and Intersection of Union (IoU). In label alignment process, we use a synset, a set of synonym(lemma, hypernym, and hyponym) that shares a common meaning in WordNet, to align two entity labels. The method using synset compares whether an entity label in a scene graph is the same entity label in other scene graph, and aligns. If the entity label is same vocabulary or in the synset of entity label for other scene graph, then IoU calculation is implemented to check whether it indicates same entity. The detailed procedure is shown in Algorithm 1.

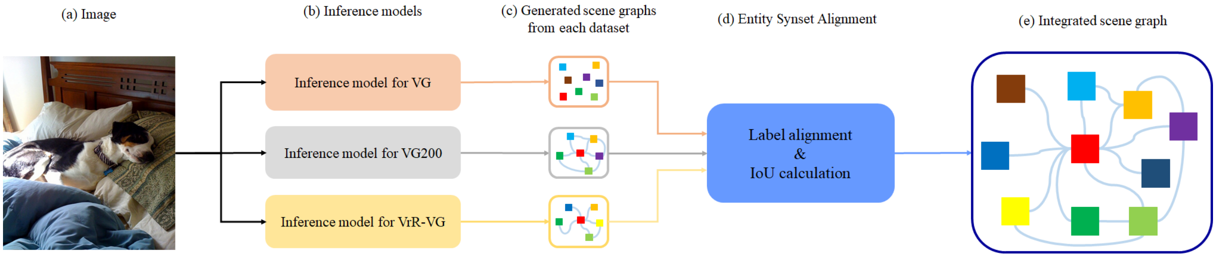
(a) Image
(b) Inference models
(c) Generated scene graphs from each dataset
(d) Entity Synset Alignment
(e) Integrated scene graph
Inference model for VG
Inference model for VG200
Inference model for VrR-VG
Label alignment & IoU calculation

Figure 2: An overview of framework which integrates visual semantic knowledge with Entity Synset Alignment(ESA). (a) A raw image goes into inference models as an input. (b) Inference models(Bottom-up attention and CompTransR) generate (c) scene graphs from each dataset(VG, VG200, VrR-VG). (e) Integrated scene graph is built as an output via (d) Entity Synset Alignment method.

4 Experiments

4.1 Scene Graph Statistics

In Table 1, we measure the average and max number of object, relation, and attribute with various combinations of scene graph datasets. Default VG200 has 12.53 average number of object and 62 max number of object, default BottomUp-VG has 26.35 average number of object and 55 for max, and default VrR-VG has 36.77 average number of object and 167 max number of object. The most key section of Table 1 is the average number of object and relation in integrating three datasets increased. This result implies that integrating three scene graphs into one scene graph can get more richer scene graph.

4.2 Image-Caption Retrieval Task

To verify the usefulness of our algorithm, we suggest an image-caption retrieval task (Kiros et al., 2014) as an application of scene graphs. The image-caption retrieval task needs visual-semantic embeddings, which is obtained by mapping the image features and caption features into joint embedding space. A general approach for this task is to obtain image features and caption features with pretrained model (such as VGGNet (Simonyan and Zisserman, 2014) for images and S-BERT (Reimers and Gurevych, 2019) for captions), then to learn mapping both to joint embedding space for maximizing similarities. In our case, we substitute image features from the pre-trained CNN model to scene-graphs and learn the representations of scene-graphs with simple 2-layer Graph Convolution Networks (Kipf and Welling, 2016). Following (Faghri et al., 2017), we use the *Max of Hinge* loss for training:

$$l_{MH}(i,c) = \underset{c'}{max}[\alpha + s(i,c') - (i,c)]_+ \\ + \underset{i'}{max}[\alpha + s(i',c) - (i,c)]_+ \tag{1}$$

where i and c are image features and caption features in joint embedding space, $s(x,y)$ is inner-product similarity function for x and y, $[x] \equiv max(x,0)$ and α serves as a margin parameter.

4.3 Results

4.3.1 Qualitative Results

Figure 3 shows each generated scene graph for an image and an integrated scene graph generated. In each scene graph, person is presented as *person* in BottomUp-VG, but *woman* in VG200 and VrR-VG. Furthermore, *phone* and *tree(s)* nodes are in BottomUp-VG and VrR-VG, but not in VG200. On the other hand, BottomUp-VG and VrR-VG have *grass* node but not in VG200. In integrated scene graph, each node has an attribute of each object such as color and some entities such as person or tree are aligned via ESA. For the setting of qualitative results, we limit the number of relation(predicate) between objects to top 20 in generated each scene graph.

4.3.2 Quantitative Results

To obtain both captions and scene-graphs for images, we select subset of images, called VG-COCO, belongs to both MS COCO dataset (Lin et al., 2014) (for captions) and Visual Genome (VG) dataset (Krishna et al., 2017) (for scene graphs). We manually split the VG-COCO dataset with 24,763 train, 1,000 validation and 1,470 test images. To evaluate the performance of image-caption retrieval task, we introduce $Recall@K(R@K)$, i.e., the fraction of

Table 1: The average and max number of object, relation and attribute with various combinations of scene graph datasets.

Method	Number of object		Number of relation		Number of attributes	
	Avg.	Max	Avg.	Max	Avg.	Max
VG200	12.53	62	50.0	50	0.0	0
VrR-VG	36.77	167	50.0	50	0.0	0
BU-VG	26.35	55	0.0	0	26.35	55
VG200 ∧ VrR-VG	37.00	167	100	100.0	0.0	0
VG200 ∧ BU-VG	27.21	66	44.39	50	26.35	55
VrR-VG ∧ BU-VG	42.04	141	29.57	50	26.35	55
VG200 ∧ VrR-VG ∧ BU-VG	41.95	127	79.67	100	26.35	55

Table 2: Quantitative results for our method on image-to-caption retrieval(caption retrieval) and caption-to-image retrieval(image retrieval) task. BU-VG is an abbreviation of BottomUp-VG.

	Method	Caption Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
CNN based	ResNet-152	26.9	65.1	79.4	24.2	36.4	39.9
GCN based	VG200	22.2	57.6	73.2	19.7	34.6	39.5
	VrR-VG	28.1	66.2	80.4	23.2	37.2	40.9
	BU-VG	27.0	65.4	80.6	23.1	37.0	40.7
	VG200 ∧ VrR-VG	29.3	67.6	81.9	23.4	37.4	41.0
	VG200 ∧ BU-VG	29.4	68.7	82.8	24.1	37.5	41.1
	VrR-VG ∧ BU-VG	27.9	70.5	83.2	23.7	37.7	41.4
	VG200 ∧ VrR-VG ∧ BU-VG	27.2	70.0	82.4	24.7	37.7	41.0

queries for which the correct item is retrieved in the closest K points to the query in the embedding space. We adopt R@1, R@5, R@10 metrics, as used in (Faghri et al., 2017).

First, to understand the effectiveness of scene graph based approach, we compare graph based method (GCN based) to CNN based model (Resnet-152). ResNet-152 trains the whole CNN networks, starting from pretrained model parameters. Here, we note that graph based method shows superior performance than the CNN based model, even though the graph based model exploits the simple two-layer graph convolution operations.

Second, we evaluate our proposed method with various combinations of VG200, VrR-VG and BottomUp-VG. The results show that integrated scene graph generally works better than default scene graph. The overall quantitative results for image-caption retrieval are presented in Table 2.

5 Conclusion

In this paper, we present a simple and efficient method to integrate multiple visual semantic knowledge into general scene graph. With a large-scale lexical database WordNet and IoU, the ESA aligns the entity labels in scene graphs generated from each dataset. The integrated scene graph has richer information and is less biased. To evaluate our proposal, we conduct the image-caption retrieval task as a down-stream task and show better performance than each scene graph. For future work, we plan to integrate more diverse visual semantic knowledge such as Human-object interaction (Gkioxari et al., 2018).

Acknowledgments

This work was partly supported by the Institute for Information and Communications Technology Promotion (2015-0-00310-SW.StarLab, 2017-0-01772-VTT, 2018-0-00622-RMI, 2019-0-01367-BabyMind) and Korea Institute for Advancement Technology (P0006720-GENKO) grant funded by the Korea government.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for

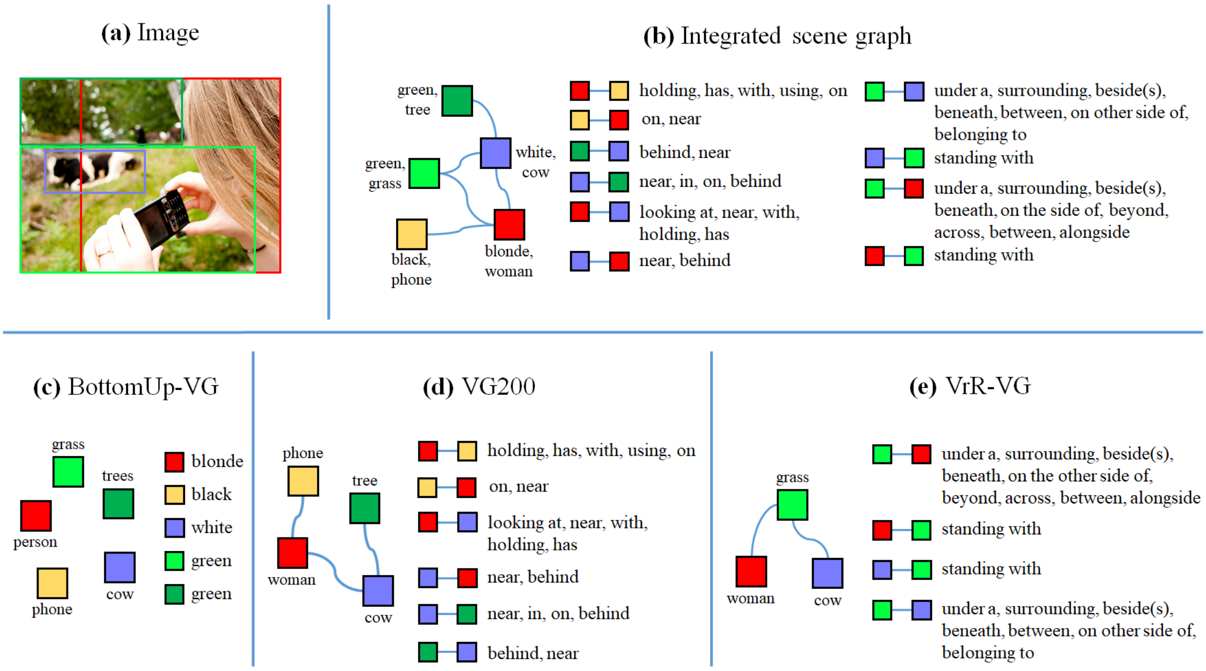
(a) Image
(b) Integrated scene graph
green, tree
green, grass
white, cow
black, phone
blonde, woman
holding, has, with, using, on
on, near
behind, near
near, in, on, behind
looking at, near, with, holding, has
near, behind
under a, surrounding, beside(s), beneath, between, on other side of, belonging to
standing with
under a, surrounding, beside(s), beneath, on the side of, beyond, across, between, alongside
standing with
(c) BottomUp-VG
grass
trees
person
phone
cow
blonde
black
white
green
green
(d) VG200
phone
tree
woman
cow
holding, has, with, using, on
on, near
looking at, near, with, holding, has
near, behind
near, in, on, behind
behind, near
(e) VrR-VG
grass
woman
cow
under a, surrounding, beside(s), beneath, on the other side of, beyond, across, between, alongside
standing with
standing with
under a, surrounding, beside(s), beneath, between, on other side of, belonging to

Figure 3: Qualitative results for our Entity Synset Alignment(ESA) method with Top 20 relations. Each scene graph (c),(d),(e) generated from inference models are combined into an integrated scene graph (b) for an image (a).

image captioning and visual question answering. In *CVPR*.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. 2018. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

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):32–73.

Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, and Tao Mei. 2019. Vrr-vg: Refocusing visually-relevant relationships. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10403–10412.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Twenty-ninth AAAI conference on artificial intelligence*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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

Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5419.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5831–5840.