Scene Graph Generation Using Depth, Spatial, and Visual Cues in 2D Images

AISWARYA S. KUMAR AND JYOTHISHA J. NAIR
Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India

Corresponding author: Jyothisha J. Nair (jyothishaj@am.amrita.edu)

This work was supported in part by the Council of Scientific and Industrial Research under Grant 09/942 (0014) 2K17.

ABSTRACT To understand an image or a scene properly, it is necessary to identify objects participating in the scene, their relationships, and various attributes that describe their properties. A scene graph is a high-level representation that confines all these features in a structured manner. Scene graph generation includes multiple challenges like the semantics of relationships considered and the availability of a well-balanced dataset with sufficient training examples. We tried to mitigate these problems by extracting two subsets, VG-R10 and VG-A16, from the popular Visual Genome dataset. Also, a framework (S2G) is proposed for generating scene graphs directly from images using depth and spatial information of object pairs. Evaluations on the scene graph generation model reveal that the proposed framework achieves better results on our data than the state-of-the-art.

INDEX TERMS Attribute prediction, relation prediction, scene graph generation, Visual Genome.

I. INTRODUCTION

Understanding an image is not an easy task, as a single image can be depicted in many ways. Similar is the case with videos. Though researchers are on the way to understand images with the help of phrases, scene graphs, and question-answers, they are only halfway to this objective. There had been lots of efforts to do relation extraction from unstructured text [25]–[27] and the extracted relations were further used for building knowledge bases as well. Similar to relation extraction, visual relationship detection (VRD) is considered as an intermediate task that extracts a triplet, <subject, predicate, object> from an image where the predicate is a relationship between subject and object. Some systems make use of these relations to generate scene graphs that can efficiently work on image retrieval [1] and image caption evaluation [2]. Fig. 1 shows an image and its scene graph where nodes represent the objects and attributes, and edges represent relations between the objects.

Visual relationships are of various types like action (‘eats’), comparative (‘taller_than’), spatial (‘in_front_of’), verb (‘playing’) and proposition (‘with’). Hence, relationship recognition is considered to be a broad and complex task in computer vision. Suppose an image contains n objects, then the maximum possible relationships in its corresponding scene graph are $n^2 - n$, for bidirectional relationships and $(1/2) \times (n^2 - n)$, for unidirectional relationships. Hence for m images, we have $m \times n$ objects and $m \times (n^2 - n)$ and $(m/2) \times (n^2 - n)$ relations in total for bidirectional and unidirectional relationships respectively. Labeling each of these predicates is hardly possible and a time-consuming one. Another challenge is that the space in which these relationships are defined is too huge to explore. It is difficult to consider all types of relationships at a single stretch. One possible way to reduce the complexity is to choose an arbitrary head to all similar relations. For example, ‘near’ can be considered as the head for relations like ‘next to’ and...
TABLE 1. Examples of polysemous predicates in Visual Genome.

| Predicates | Semantically similar predicates |
|------------|--------------------------------|
| on         | above, part of, growing on, on top of, attached to |
| has        | wearing, part of, with, carrying |
| near       | with, next to, at |

‘adjacent’ since all of these have a similar meaning. The next challenge is data imbalance that can occur while labeling the data, where the classes have an unequal number of instances. If all these can be taken care of, then the problems in visual relation recognition are almost solved.

It is a usual method to predict the relation by training a model with visual features of objects. These are not effective methods as the relation between two objects does not merely depend alone on their visual appearance. Another way to predict the relation is to incorporate other information like descriptions of images, knowledge bases, and the semantic embedding of object categories with visual characteristics. However, the systems become more complex in such situations and try to concentrate on objects irrelevant to the scene. Two objects do not need to be related if their bounding boxes intersect and may relate even if their bounding boxes do not overlap. Considering all these, we propose a system that can recognize relationships from images by utilizing depth and spatial information of object pairs and the visual cues and represent it as scene graphs. Most of the works concentrate either on object-object interaction, human-human interaction, or human-object interaction. In the real world, we cannot put such constraints on interactions, and hence, this work considers all three types of relationships with equal importance.

Another important fact is that majority of the relations in the existing relationship dataset are polysemous, where a particular relation has different semantics in different contexts. The polysemous property of relations often leads to difficulties in identifying the exact meaning of these relations [28]. For example, the relation triplet <man, carry, bag> has the same meaning as <man, has, bag> and <man, with, bag>. To the best of our knowledge, there are two ways to overcome this issue. The first is to avoid mutually exclusive relations, and the second is to choose mostly used relations. We opt for the latter since the former requires high expertise and time. Table 1 shows some example predicates in Visual Genome (VG) dataset [11] that are semantically similar in different contexts.

Along with identifying relationships among objects, it is necessary to extract different attributes of these objects for proper understanding of an image [23], [24]. Attributes are visual properties of objects present in the image. It can be color, visual patterns like stripes or spots, material type such as wooden, or refer to the shape of the object. Besides, an object can have one or more attributes, and as the number of attributes increases, the image becomes more descriptive. Due to the substantial properties of these domains, there is no predefined number of relations or attributes that are required to represent an image meaningfully.

The main challenge in relation recognition is the non-availability of sufficient labeled data and long-tail distribution of classes. The problem with the long-tail distribution of class labels is that they do not provide an optimized solution using conventional supervised learning models. Considering these facts, we extracted two small subsets from Visual Genome. We also introduced the S2G model for generating scene graphs, without using any external knowledge. This consists of two modules - relation extraction and attribute prediction. The modules are analyzed individually, and results show that our model outperforms baseline models on the extracted dataset. In a nutshell, this paper focuses mainly on i) grounding an image into a scene graph with fewer types of relations and attributes and ii) dealing with the long-tail distribution of class labels in the existing dataset. The remaining part of the paper is organized as follows: Section 2 discusses the related works in image understanding, scene graph generation, and image retrieval. Section 3 deliberates the procedure for data extraction from the existing Visual Genome dataset. Section 4 details the proposed approach. Section 5 discusses the experimental results, and finally, section 6 concludes the paper.

II. RELATED WORKS

This section briefs some major milestones in relation prediction and scene graph generation. Muraoka et al. [3] trained a classifier to obtain positional relation between a pair of objects from images with the help of descriptions. Instead of handcrafting the labels, they automatically extract the relations using IBM model [4]. Further, a neural network recognizes the relationships with features like relative coordinates and the intersection of two objects. Muraoka et al. [3] used both images and sentences from the MSCOCO dataset for exploring the relationships and brilliantly solved the challenge of aligning the text and image objects. The work claims to be the first to explore open vocabulary relation between objects and demonstrate that spatial features contribute much towards relation recognition.

Due to insufficient relation labels, Ruichi et al. [5] make use of linguistic knowledge to predict the relationship jointly with the name of the subject and object and their relative spatial arrangement using deep neural networks. The model combines internal and external knowledge using a student-teacher framework which ensures better generalization. Liang et al. [6] propose a deep structural ranking that integrates multiple cues for identifying relationships in images. Further, they designed a ranking function by leveraging the labeled relationships to have higher class priority.

Zhang et al. [7] introduce a visual model that has object detection as well as recognition modules. Instead of training with different predicate images, the model learns a consistent translation vector in the relation space regardless of the subject-object combination. Che et al. [8] perform paragraph generation along with relation detection, which identifies significant regions of interest in an image and performs...
the detection with the help of a discriminant network. Yao et al. [9] explore the visual relationship between objects for caption generation in images. They have used Graph Convolutional Networks (GCN) and Long Short-Term Memory (LSTM) architectures to produce sentences from images by integrating semantic and spatial relationships among objects. The model achieves state-of-the-art performance on the MSCOCO dataset.

Zhang et al. [21] input an image to the visual module and extracts three visual embeddings corresponding to subject, predicate, and object. Similarly, word vectors of subject, predicate, and object are given as input to the semantic module, and three semantic embeddings are also collected. Then while training, the final loss is minimized by comparing the embeddings with previously designed losses. They claim to be the first to have evaluated on original visual genome dataset. Recently, Chen et al. [30] proposed a modified and improved model of Zellers et al. [19], utilizing statistical correlations between objects and their relations to build a structured knowledge graph. The model passes messages through the graph to capture the interplay between object pairs and their relations. In addition to that, the contextual cues are also explored by another graph with similar message propagation. Finally, the scene graph is generated based on these two graph neural networks. Moreover, Zellers et al. [19], Woo et al. [31], Wang et al. [33] and Ren et al. [34] utilize global context to explore relations between objects for building scene graphs whereas, Yang et al. [18] and [36] capture sub-graph structures in images.

Recently, Xu et al. [17], Liao et al. [20], and Chen et al. [30] proposed message-passing systems that can pass messages between the nodes to update the scene graphs. Lu et al. [22], Sadeghi et al. [32], Liao et al. [35] and Plummer et al. [37] further utilize image captions or language cues of the object pairs to strengthen visual relationship detection. Similar to our model, Ding et al. [38] proposed a depth-guided relation prediction system to recognize spatial relations in images. The model uses linguistic information to frame out common sense knowledge of objects. In contrast to the existing systems that leverage language or context for relation prediction, we introduce a novel model to incorporate spatial and depth cues into visual features for relation recognition.

III. DATASET

Visual genome is a well-organized dataset developed specifically for exploring relations between objects in images. It contains 108,077 images, 75,729 unique object categories, 40,480 distinct relationships, 40,513 unique attributes, and other features like region descriptors and scene graphs. On average, one image has 22 objects, 18 relationships, and 16 attributes. Since the dataset is too broad for experiments, we collected a sample of images and further down-scaled the number of relations and attributes. The following sections detail on creation and analysis of the dataset.

| TABLE 2. Aggregate statistics for Visual Genome, extracted subsets (VG-R and VG-A), and cleansed subsets (VG-R10 and VG-A16). |
|---------------------------------------------------------------|
| Visual Genome | VG-R | VG-A | VG-R10 | VG-A16 |
| Unique objects | 75,729 | 9,420 | 200 |
| Object instances | 1,366,673 | 703,382 | 62,081 |
| Unique relationships | 40,480 | 4095 | 10 |
| Relationship instances | 1,534,448 | 112,227 | 25,221 |
| Unique attributes | 40,313 | 7,588 | 16 |
| Attribute instances | 1,670,182 | 112,060 | 63,070 |
| Objects per image | 21.24 | 33.87 | 14.1 |
| Relationships per image | 17.68 | 18.69 | 19.7 |
| Attributes per image | 16.08 | 18.67 | 18.5 |

A. SUBSET EXTRACTION FROM VISUAL GENOME

To extract subsets from Visual Genome, we pulled out sample images, relations, and attributes associated with those images (named as VG-R and VG-A) and cleansed them to form VG-R10 and VG-A16, respectively.

1) VG-R10: RELATIONSHIP DATASET

A sample of 6,000 images and 4,096 unique relationships associated with these images are collected. In total, 112,227 relationship instances are extracted from 6,000 images, and on further analysis, the following discrepancies are found in predicate labels.

- Unimportant labels (“have been”, “says”)
- Confusing labels (“I the”, “in at”, “is of a”)
- Similar labels (“next to” and “next to a”, “wears” and “wearing”)

Even after removing all such labels, it was observed that the label frequencies were scattered over a wide range (Fig. 2a). Also, in the existing 4,096 relations or predicates, hardly 1.7% have more than 100 instances. Hence, to avoid overfitting the learning model, we considered only those relations that have more than 1,250 instances. We got instances of 10 frequent relations and labeled them as VG-R10. Thus the new sample has a distribution as shown in Fig. 2b.

2) VG-A16: ATTRIBUTE DATASET

From the collected sample images, 112,060 attribute instances were extracted with distribution as in Fig. 3a. Most of the labels were confusing, say “very large” and “big”, whereas some were object names (“computer”) or relations (“wearing”). All such labels were removed to get 63,070 instances, and only those with more than 1,125 instances were chosen. Hence in total, we got instances of 16 frequent attributes. We labeled it as VG-A16 and had distribution as shown in Fig. 3b.

Both datasets were balanced before using in experiments to avoid overfitting of the prediction models. Even when we considered only the most common categories of predicates and attributes, the extracted dataset was still very rich with a mean of 14.1 objects, 18.5 attributes, and 19.7 relationships per image as shown in Table 2. Images were
A. S. Kumar, J. J. Nair: Scene Graph Generation Using Depth, Spatial, and Visual Cues in 2D Images

IV. OUR APPROACH: SCENE GRAPH GENERATOR (S2G)

We represent a scene graph as $S = (V, E)$ where $V = \{O, A\}$ is a set of nodes, $O = \{o_1, \ldots, o_n\}$ is a set of objects and $A = \{a_1, \ldots, a_m\}$ is a set of attributes. There may be a situation where multiple instances of a single object $o_i$ can appear in the scene. Hence, all object instances are assigned a unique identifier along with the object name. This is to ensure proper association between the bounding boxes and objects which may be represented as $o_i \rightarrow b_i$ where $B = \{b_1, \ldots, b_n\}$ is the set of bounding boxes associated with the objects.

S2G consists of three main modules: object detection, relation extraction, and attribute prediction.

A. OBJECT DETECTION

For a given image, objects were identified using Faster RCNN [13], a layered architecture that uses Region Proposal Network (RPN) by generating anchor boxes of different scales and aspect ratios. We used the same scales and aspect ratios mentioned in the original work.

Fig. 4 shows examples of object detection on test images. Faster RCNN had two outputs, object classifications, and bounding boxes, used as the input for relation extraction.
B. RELATION EXTRACTION USING VG-R10 DATASET

This section will discuss the proposed relation extraction module in detail. Fig. 5 depicts an overview of the relation extraction module. This module considers visual, spatial, and depth information of object pairs to predict their relation. Once the objects are detected, a depth map for the input image is generated (Fig. 6) using a multi-scale deep network [29].

During depth estimation, pixel intensity values in each of the bounding boxes are averaged to obtain depth scores for object regions in the depth map.

\[ ds_i = \frac{1}{p_i} \sum_{x,y} I(x, y) \]  

(1)

Then, the relative depth score of objects are calculated by,

\[ di = \frac{ds_i}{ds} \]

\[ D = [d_1 \ d_2 \ d_3 \ldots \ dn] \]

(2)

where \( ds \) is the depth score of the deepest object in the corresponding image. \( D \) forms the relative depth score vector for \( n \) objects.

We finetune the VGG16 base model with the training set in our VG-R10 dataset and the visual features of objects are obtained.

\[ F_{map}(O_i) = VGG16(O_i) \]

(3)

\[ V_{map}(O_i) = GAP(F_{map}(O_i)) \]

(4)

VGG16 outputs a feature map of dimension \( 7 \times 7 \times 512 \), which undergoes global average pooling operation. Thus, \( V_{map} \) has a dimension of \( 1 \times 1 \times 512 \). A spatial location encoder (SLE) module is added to obtain spatial relations between the objects. The module receives bounding box coordinates of objects and computes a spatial score vector for the pair, which is defined as

\[ SS(O_1, O_2) = (\Delta l, \Delta r, \Delta b, \Delta t) \]

\[ = (x_{1lb} - x_{2lb}, x_{1br} - x_{2br}, y_{1bl} - y_{2bl}, y_{1tl} - y_{2tl}) \]

(5)

\( x_{bl}, x_{br}, \) and \( x_d \) are the x-coordinates and \( y_{bl}, y_{br} \) and \( y_d \) are the y-coordinates of bottom-left, bottom-right and top-left
corner of bounding box respectively. A two-layer perceptron is used for relation prediction and it expects four inputs, a pair of object regions along with their relative depth scores, relative depth scores of all objects in the scene, and spatial score vector. The feature vector formation for the perceptron is as follows.

\[ f(O_i, O_j) = (V_{map}(O_i) \oplus d_i) \oplus (V_{map}(O_j) \oplus d_j) \oplus D \oplus SS(O_i, O_j) \]  

(6)

This is repeated for all the object pairs in an image. For an image \( I \), we consider a pair of objects \( <o_i, o_j> \), \( \forall i < j \) and the relation model predicts the predicate, \( p_{ij} \) that connects the corresponding object pair. Since S2G assumes unidirectional relation between objects, a relation in the form \( <o_i, p_{ij}, o_j> \) means that \( o_i \) is related to \( o_j \) with predicate, \( p_{ij} \) and not vice versa.

C. ATTRIBUTE PREDICTION USING VG-A16 DATASET

Similar to the relation module, attribute edges of objects are extracted using the pre-trained VGG16 network. The model is initially trained in an unsupervised manner using a pair-wise approach [12] and then fine-tuned with the VG-A16 dataset. Feature generation is same as in (3) and (4) where as the final feature vector is represented as

\[ f(O_i) = V_{map}(O_i) \]  

(7)

The extracted feature vector, \( f(O_i) \) is given as input to a two-layer perceptron for attribute prediction. 16 frequently occurring attributes include labels like white, black and green. Each object \( o_i \) can have multiple attributes. For example, ‘a cow with black and white spots’. Here a single object cow can have three different attributes; spotted, black, and white. Therefore, \( o_i \rightarrow A' \) where \( A' \subseteq A \). When the object has zero attributes, \( A' = \emptyset \).

D. SCENE GRAPH FORMULATION

The scene graph, \( S \) is generated in three steps.

1) For each object \( o_i \), a node is created.

2) As the relation module predicts the predicate \( p_{ij} \), an edge is added in between the corresponding pair of objects \( <o_i, o_j> \).

3) As the attribute module predicts the attribute, a node and an edge is added from the corresponding object node, \( o_i \).

Fig. 7 shows the output of scene graph generation on sample images. As discussed earlier, the generated scene graph is represented as \( S = (\{O, A\}, E) \). We used a notation, \( \gamma = (\gamma_{rel}, \gamma_{attr}) \) for representing significance value of prediction. This parameter restricts unimportant edges and nodes from getting added to the graph. The threshold values for these parameters are set through continuous experiments. We randomly chose 50 images from the test set and gave them as input to S2G with varying significance values. The generated scene graphs were given to different users, and for each image, it was asked to mark the reasonable one

V. EXPERIMENTS AND RESULTS

From VG-R10, 10,000 and 2,500 instances were split as training and testing data. Furthermore, from training images, 2,000 were taken for validation. The relation model was trained over 70 epochs with stochastic gradient descent as the optimizer. The initial learning rate was chosen to be 0.0001 with a decay factor of \( 1e^{-6} \). Similarly, from VG-A16, 11,520 samples were taken training, 2,880 for validation, and 3,600 for testing. Initially, the attribute model gave only 28.43 and 30.12 for recall@50 and recall@100, respectively. Hence, the model was trained in a pair-wise manner [12] initially and then fine-tuned using our dataset. This helped in improving the recall values by about 25%.

A. QUANTITATIVE ANALYSIS

Table 3 shows the performances of different approaches on the VG-R10 dataset. We evaluated the models on 2,500 test images in VG-R10 and used recalls under the top 50 and 100 predictions as metrics to measure their performances. Recall@k computes the fraction of times the correct rela-
A. S. Kumar, J. J. Nair: Scene Graph Generation Using Depth, Spatial, and Visual Cues in 2D Images

**FIGURE 8.** Scenario of overall system. Phase 1 creates nodes for corresponding objects, phase 2 adds detected relations as edges and finally phase 3 appends the attributes for each object. The scene graph is shared between all the phases.

**TABLE 3.** Comparison of relation prediction models on the VG-R10 dataset.

| Method          | Recall@50 | Recall@100 |
|-----------------|-----------|------------|
| MSDN [16]       | 42.3      | 48.2       |
| Message passing [17] | 44.8      | 53.0       |
| CSC [33]        | 53.2      | 57.9       |
| Graph-RCNN [18] | 54.2      | 59.1       |
| Neural-Motifs [19] | 65.2      | 67.1       |
| KERN [30]       | 65.8      | 67.8       |
| HCNet [34]      | 66.4      | 68.8       |
| SGRN [20]       | 66.5      | 68.7       |
| L-SVRU [21]     | 68.4      | 70.4       |
| Ours(S2G-V)     | 71.0      | 73.9       |
| Ours(S2G-V+D)   | 72.3      | 74.2       |
| Ours(S2G-V+S)   | 71.7      | 73.4       |
| Ours(S2G-V+D+S) | 73.2      | 76.5       |

**TABLE 4.** Test accuracy of S2G model variants on VG-R10 dataset.

| Predicate      | S2G-V | S2G-V+D | S2G-V+S | S2G-V+D+S |
|----------------|-------|---------|---------|-----------|
| in             | 76.2  | 76.2    | 76.4    | 76.4      |
| has            | 72.7  | 65.1    | 62.6    | 62.6      |
| of             | 63.4  | 63.4    | 63.4    | 63.5      |
| wearing        | 70.0  | 70.8    | 71.2    | 74.9      |
| with           | 66.8  | 67.6    | 66.9    | 69.0      |
| on a           | 70.9  | 73.6    | 70.5    | 76.2      |
| near           | 74.3  | 75.8    | 72.3    | 77.4      |
| on top of      | 67.1  | 67.4    | 74.3    | 74.6      |
| in front of    | 69.0  | 75.8    | 66.9    | 78.8      |
| behind         | 61.3  | 64.7    | 62.1    | 67.3      |

**Table 3** includes different variants of our model named S2G-V, S2G-V+D, and S2G-V+S to study the role of visual, depth, and spatial information in images. S2G-V+D+S is the proposed model. Table 4 reveals the performance of the relation model variants. Note that in Table 4, predicates like ‘on a’, ‘near’, ‘on top of’, ‘in front of’, and ‘behind’ receive more attention in the proposed model than in the variants. That is, certain predicates depend more on spatial and depth information than others.

The main highlight of this work is that only a small subset of the original Visual Genome data was used for the scene graph generation. Even though VG is an excellent dataset for image understanding, it is highly noisy and refining this huge data manually requires much effort. Both datasets that were used for experiments are fully refined and balanced. This indeed boosts the performance of models. Another highlight is that the proposed model uses no additional knowledge like text phrases or scene graphs for learning. A detailed performance analysis was done on the variants of the relation...
model. Each of 10 predicates is evaluated separately, and test accuracies are plotted in Table 4. From the table, it can be noticed that S2G-V+D+S performs better on predicates like ‘on top of’ and ‘behind’ which requires both spatial and depth knowledge for understanding a scene. Another important fact is the presence of polysemous relations. For example, ‘has’ can be interpreted as ‘wearing’, ‘with’, or ‘carrying’. As discussed earlier, most of the relations in the real world are polysemous and thus can be used interchangeably. Due to this property, even though we consider most frequent relations, those relations represent a wide variety of predicates in Visual Genome.

B. ANALYSIS ON SCENE GRAPH GENERATOR

SGGen [17] and SGGen+ [18] are two widely used metrics evaluating scene graphs. The former is a triplet-based metric that considers \(<subject, predicate, object>\) triplets, while the latter is an augmented version of the former with the addition of singletons; objects, and predicates. SGGen returns 1 if all elements in the triplet have been correctly identified, whereas SGGen+ is more lenient. The computation of SGGen+ is formulated as:

$$\text{Recall} = \frac{C(O) + C(P) + C(T)}{N} \quad (8)$$

where $$C()$$ is a counting operation, and hence $$C(O)$$ is for object nodes recognized correctly; $$C(P)$$ is for predicate; $$C(T)$$ is for triplet and finally, $$N$$ is the number of entries (sum of the number of objects, predicates, and relationships) in the ground truth graph.

The problem with this evaluation method is that it does not consider the attributes of objects. Hence we modified SGGen+ to include object attributes and the singletons and triplet. The modified metric is named ASG and computed as:

$$\text{Recall} = \frac{C(O) + C(P) + C(T) + C(A) + C(D)}{N_{RA}} \quad (9)$$

where $$C(A)$$ is the number of correctly recognized attributes and $$C(D)$$ is for duplet \(<object, attribute>\) and hence $$N_{RA}$$ modifies to be the sum of the number of objects, predicates, relationships and attributes along with duplets in the ground truth. Ground truth scene graphs were manually created for proper evaluation, as the ground truths from Visual Genome were not up to the mark. Table 5 details the performance of state-of-the-art on scene graph generation task.

We analyzed the performance of scene graph generation based on Recall@50 and Recall@100. MSDN [16], Message passing [17], Graph-RCNN [18] and Neural-motifs [19] were evaluated using SGGen+ only, since they didn’t value attributes of objects in images. [15] reports a lower recall when compared to other methods because it was unable to perform well on predicate detection tasks which got reflected in both ASGGen+ and SGGen+ scores. The proposed model outperforms the frequency baseline indicating that this
FIGURE 10. Sample images from VRD (a)–(c) and their corresponding scene graphs (d)–(f).

FIGURE 11. Sample images from VrR_vg dataset (a)–(c) and their corresponding scene graphs (d)–(f).
approach can efficiently generate semantic scene graphs from images.

C. QUALITATIVE ANALYSIS

The proposed system is tested on three popular datasets: Visual Genome (VG), Visual Relationship Dataset (VRD), and Visually-Relevant Relationships (VrR-VG), Fig. 9, 10 and 11 show example images and the generated graphs using proposed framework. In Fig. 9a, our model captures relations between ‘5_dining table’ and the remaining objects (except ‘1_cake’). It recognizes a ‘red cup’ but is confused with overlapping objects. For example, the ‘5_dining table’ gets the color of ‘3_cup’. In Fig. 9c, ‘3_chair’ and ‘1_cat’ are predicted as ‘red’ and ‘brown’ respectively but the relation prediction performs poorly. Most of the predicted attributes and the relation between ‘2_horse’ and ‘3_person’ in Fig. 10a are correct, but the system confuses with ‘1_horse’ and ‘3_person’. Further, in Fig. 11a, ‘3_elephant’ is identified as larger than the other two and in Fig. 11b, we find ‘1_cup, near 2_cup’ and ‘2_cup’ to be ‘blue’.

With the available 10 relations, more relations are found when compared to ground truths in Visual Genome. However, as expected, in some cases, the prediction was senseless. For example, if there exist relations like <man, on, chair> and <jacket, on, man> in an image, it is not necessary that the relation <jacket, on, chair> should also exist. Though ideally, it is correct, usually humans neglect such kinds of irrelevant relations. It is also noted that object detection is a critical phase in this framework. The result of relation and attribute prediction models depends on the output of the object detection phase.

VI. CONCLUSION

Despite advancements in computer vision techniques such as object recognition, computers are still not able to achieve acceptable performance on tasks such as image understanding, image captioning, and image retrieval. Various challenges involved in extracting visual knowledge from images were studied, and we came up with a straightforward approach to understand images using the underlying relationships and attributes of objects. Two datasets are extracted from Visual Genome, namely VG-R10 and VG-A16, and could achieve benchmark results on these datasets even without using any external knowledge. In the future, we intend to use these scene graphs to build an efficient image retrieval system by incorporating high-level semantics.

REFERENCES

[1] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, “Image retrieval using scene graphs,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 3668–3678.
[2] P. Anderson, B. Fernando, M. Johnson, and S. Gould, “SPICE: Semantic propositional image caption evaluation,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 382–398.
[3] M. Muraoka, S. Maharanj, M. Saito, K. Yamaguchi, N. Okazaki, T. Okatani, and K. Inui, “Recognizing open-vocabulary relations between objects in images,” in Proc. 30th Pacific Asia Conf. Lang., Inf. Comput., 2016, pp. 239–248.
[4] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, “The mathematics of statistical machine translation: Parameter estimation,” Comput. Linguistics, vol. 19, no. 2, pp. 263–311, Jun. 1993.
[5] R. Yu, A. Li, V. I. Morariu, and L. S. Davis, “Visual relationship detection with internal and external linguistic knowledge distillation,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1974–1982.
[6] K. Liang, Y. Guo, H. Chang, and X. Chen, “Visual relationship detection with deep structural ranking,” in Proc. 32nd AAAI Conf. Artif. Intell., 2018, pp. 1–8.
[7] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, “Visual translation embedding network for visual relation detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 5532–5540.
[8] W. Che, X. Fan, R. Xiong, and D. Zhao, “Paragraph generation network with visual relationship detection,” in Proc. 26th ACM Int. Conf. Multimedia, Oct. 2018, pp. 1435–1443.
[9] T. Yao, Y. Pan, Y. Li, and T. Mei, “Exploring visual relationship for image captioning,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 684–699.
[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 580–587.
[11] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” Int. J. Comput. Vis., vol. 123, no. 1, pp. 32–73, 2017.
[12] A. S. Kumar and J. J. Nair, “Pair wise training for stacked convolutional autoencoders using small scale images,” J. Intell. Fuzzy Syst., vol. 36, no. 3, pp. 1987–1995, Mar. 2019.
[13] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 91–99.
[14] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556.
[15] X. Liang, L. Lee, and E. P. Xing, “Deep variation-structured reinforcement learning for visual relationship and attribute detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2014, pp. 580–587.
[16] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, “Scene graph generation from objects, phrases and region captions,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1261–1270.
[17] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, “Scene graph generation by iterative message passing,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 5410–5419.
[18] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, “Graph R-CNN for scene graph generation,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 670–685.
[19] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, “Neural motifs: Scene graph parsing with global context,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5831–5840.
[20] W. Liao, C. Lan, W. Zeng, M. Ying Yang, and B. Rosenhahn, “Target-tailored source-Transformation for scene graph generation,” 2019, arXiv:1904.02104.
[21] J. Zhang, Y. Kalantidis, M. Rohrbach, M. Paluri, A. Elgammal, and M. Elhoseiny, “Large-scale visual relationship understanding,” in Proc. AAAI Conf. Artif. Intell., vol. 33, 2019, pp. 9185–9194.
[22] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, “Visual relationship detection with language priors,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 852–869.
A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, “Describing objects by their attributes,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 1778–1785.

H. Chen, A. Gallagher, and B. Girod, “Describing clothing by semantic attributes,” in Proc. Eur. Conf. Comput. Vis. Berlin, Germany: Springer, 2012, pp. 609–623.

G. Veena, R. Hemanth, and J. Harreesh, “Relation extraction in clinical text using NLP based regular expressions,” in Proc. 2nd Int. Conf. Intell. Comput., Instrum. Control Technol. (ICICICT), vol. 1, Jul. 2019, pp. 1278–1282.

R. R. K. Menon, D. Joseph, and M. R. Kaimal, “Semantics-based topic inter-relationship extraction,” J. Intell. Fuzzy Syst., vol. 32, no. 4, pp. 2941–2951, Mar. 2017.

V. Hariharan, M. A. Kumar, and K. P. Soman, “Relation extraction using convolutional neural networks,” in Proc. Int. Conf. ISMAC Comput. Vis. Bio-Eng. Cham, Switzerland: Springer, 2018, pp. 937–944.

J. Luo, J. Zhao, B. Wen, and Y. Zhang, “Explaining the semantics capturing capability of scene graph generation models,” Pattern Recognit., vol. 110, Feb. 2021, Art. no. 107427.

D. Eigen and R. Fergus, “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 2650–2658.

T. Chen, W. Yu, R. Chen, and L. Lin, “Knowledge-embedded routing network for scene graph generation,” in Proc. IEEE/CVF Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 6163–6171.

S. Woo, D. Kim, D. Cho, and I. S. Kweon, “LinkNet: Relational embedding for scene graph,” in Proc. Adv. Neural Inf. Process. Syst., 2018, pp. 1–11.

M. A. Sadeghi and A. Farhadi, “Recognition using visual phrases,” in Proc. CVPR, Jun. 2011, pp. 1–57.

W. Wang, R. Wang, S. Shan, and X. Chen, “Exploring context and visual pattern of relationship for scene graph generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 8188–8197.

G. Ren, L. Ren, Y. Liao, S. Liu, B. Li, J. Han, and S. Yan, “Scene graph generation with hierarchical context,” IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 2, pp. 909–915, Feb. 2021.

W. Liao, B. Rosenhahn, L. Shuai, and M. Y. Yang, “Natural language guided visual relationship detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2019, pp. 1–10.

Y. Li, W. Ouyang, B. Zhou, J. Shi, C. Zhang, and X. Wang, “Factorizable Net: An efficient subgraph-based framework for scene graph generation,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 335–351.