Attention-Based Scene Graph Generation: A Review

Afsana Airin, Rezab Ud Dawla, Ahmed Shabab Noor, Muhib Al Hasan, Ahmed Rafi Hasan,
Akib Zaman* and Dewan Md. Farid*
Department of Computer Science & Engineering, United International University
United City, Madani Avenue, Badda, Dhaka 1212, Bangladesh
Email: {aairin191172, rdawla191187, anoor193024, mhasan191083, ahasan191131}@bscse.uiu.ac.bd,
{akib, dewanfarid}@cse.uiu.ac.bd

***Abstract*—The automated creation of a semantic structural scene graph from an image or video is known as scene graph generation (SGG), which includes accurate labeling of all objects that are identified and the interconnections between them. Several SGG methods have been proposed employing deep learning techniques nowadays to achieve good results but most of the approaches failed to integrate the contextual information of pair of objects. Apart from the existing state of the arts of SGG, the attention mechanism is creating a new horizon in this field. This paper offers a thorough analysis of the most recent Attention-Based Scene Graph Generation techniques. In this paper, we have compared and tested five existing Attention-Based Scene Graph Generation methods. We have summarised the results of existing methods to understand progress in this field of Scene Graph Generation. Moreover, we have discussed the strengths of existing techniques and future directions of attention-based models in Scene Graph Generation.**

***Index Terms*—Attention Mechanism; Image Captioning; Scene Understanding; Scene Graph; Visual Relationship;**

I. Introduction

Digital images, movies, and other visual inputs may all be used as sources of information for embedded systems thanks to the area of artificial intelligence known as computer vision (CV). It analysis the extracted information and implements changes or makes recommendations in response to the findings. Scene understanding has been playing a significant role in the research growth of Computer Vision throughout the last several years through the inclusion of several cutting-edge technologies. Scene understanding is the process of perceiving a 3D dynamic scene through a network of sensors by detecting, analysing, and often in real-time [1]. Deep learning has considerably advanced this discipline during the last several decades. To capture the relationship between objects scene graph has emerged. The scene graph was first represented as a data structure that can depict the relationship between objects and objects instances [2]. To enhance the performance of scene understanding, not only object detection but also numerous neural networks have evolved to recognize visual correlations and perform picture captioning. [3] and their performance is quite nearby to human performance. The usage of scene understanding in real life is very impactful such as scene understanding for the eye-disabled human, vehicle detection, and tracking traffic. It is also contributing to the robotics field specifically in autonomous navigation [4], [5], reconnaissance [6], pose estimation, etc. Fig. 1 represents the SGG process.

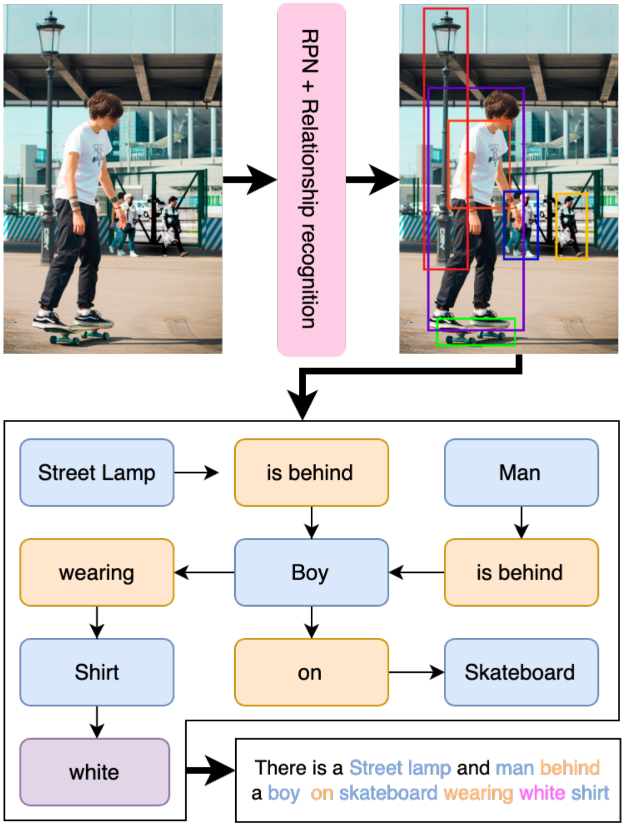
RPN + Relationship recognition
Street Lamp
is behind
Man
wearing
Boy
is behind
Shirt
on
Skateboard
white
There is a Street lamp and man behind a boy on skateboard wearing white shirt

Fig. 1: Simplified representation of Scene Graph Generation.

Scene graphs are structural representations that accurately describe objects, characteristics of objects, and interactions between paired items in order to capture exact semantics [7]. Objects, attributes, and relations are the main elements of a scene graph [8]. A topological representation of a scene that encodes the objects and their relationships is called a full scene graph [2]. An image has two fundamental building components: subject and object which can be located with bounding boxes. Briefly, a group of visual connection triplets may serve as a representation of a scene graph <*subject, predicate, object*>. A scene graph's major objective is to build a structured representation that will help those who struggle with visual and semantic perception grasp visual situations completely by parsing one picture or a series of images. The three stages of creating a scene graph are feature extraction, contextualization, and graph building and reasoning as shown in Fig. 1. Over the years, several models have been introduced to construct a scene graph such as CNN-based SGG [7], RNN/LSTM-based SGG [9], and GNN-based SGG [10]. CNN-based SGG LinkNet [7] developed an auto-encoder architecture where there are direct connections from encoders to decoders to pass information. Similarly, Another prominent area of study is RNN-LSTM techniques, which naturally excel at collecting context in the scene graph and inferring from structured data [9]. An intuitive approach named GNN-based SGG [11] involved a network structure where neighboring nodes might successfully learn local information. Most of the methods follow the same process.

When understanding a scene, we not only need to consider the object pair, rather we need also to consider the feature of the predicate and the influence that the rest of the scene has on that object pair. However, only object features are taken into account by traditional SGG methods, even though it is well-researched that predicate features and contextual information have a great impact on overall scene understanding [12]. That being the case, attention mechanism-based SGG methods solve this problem by integrating contextual information among the subjects and objects. Since the introduction of the Recurrent Attention Model (RAM) [13] for categorization of images, attention mechanisms grew rapidly. The generation of scene graphs has two objectives. One involves refining local gestures, while the other combines global contextual data. In this application, the attention mechanism effectively represents feature information and allows feature refining. Attention is applied at the feature representation stage to provide more insightful object regions appearance and unions of object pairs. Attention have been applied in the spatial domain [14], channel domain [15], or more generally in their combined domain [16]. In this study, we have reviewed the most recent research where the attention mechanism is being utilized to achieve scene understanding through the generation of scene graphs. We have tested the performance of the five existing Attention-Based Scene Graph Generation methods.

The remainder of the article is organized as follows: The overview of the most current attention-based state-of-the-art SGG models is highlighted in Section II. Section III highlights the summary of the performance metrics. Section IV discusses about the strengths and future directives of the reviewed models. Finally, Section V concludes the article summarising our work along with discussing the future scopes.

II. Exploration of State-of-the-Art

The process of creating a scene graph often starts from the bottom up with the grouping of things into triplets, which are then connected to create the whole scene graph. Evidently, the main purpose of the task is to identify the triplets of the subject, predicate, and object known as <*S, P, O*>. In this work, We focus on SGG based on the context embedding using a different variant of attention mechanism. The attention mechanism allow models to concentrate on the most important elements of the input [17]. A scene graph with an iterative message-passing architecture has two pathways: combining global contextual data with local feature refinement. Both the feature representation stage and the feature refining stage may use attention techniques. There are two approaches to creating a scene graph [18]. It uses a two-step sequential strategy to identify items in the first approach. The associations between each pair of items are then established via the use of a categorization problem. The other approach uses joint reasoning of the objects and their relationships depend on proposed object regions. For both approaches, first, the objects are detected in the image and grouped into pairs and used in a union area. Fig. 2 illustrates the generalized architecture of the SGG using the attention-based models. Given an image as input, in the first step, the objects of the image are detected and the feature is extracted with Faster R-CNN. Then the models ATR-Net, RelMN, Relational transformer, MSTG, and RSAN use different types of attention mechanisms such as multi-head attention, self-attention, and fuse attention. Different models generate different outputs: A sparse graph with a corresponding relationship, text, and image captions.

Object detection is the very first step of scene graph generation to detect and locate objects in an image. In order to create subject/object and union suggestions, a scene graph generation approach initially uses a Region Proposal Network (RPN), it was created using the image's original annotations. Here, for object detection, all the papers used Faster R-CNN [19]. The following subsection describes the attention-based methods applied in the SGG after the object detection part.

Following the success of the Recurrent Attention Model (RAM) [13] for image classification, attention mechanisms quickly took off. The attention mechanism is used in the SGG [20] dealing with overlapping and noisy object areas. The attention mechanism mainly focuses on the small but important part of the graph to extract more relevant information eliminating less relevant information. The self-Attention mechanism combines an object's multi-modal features to provide a more thorough representation. On the other hand, Context-Aware Graph parsing is used by attention to understand contextual characteristics. Other types of attention mechanisms are single-head attention and multi-head attention. Gkanatsios et al. [21] proposed ATR-Net where they used multi-head attention. ATR-Net focuses on every edge to solve the following two issues: assuming the two items are connected, setting a probability to the predicate P(P|related), and predicting a relevance score that determines the likelihood of interaction. In this method, the ultimate predicate score stands as $P(Pa) = P(Pa \mid related)P(related)$ and the edge's score stands as $P(Sa, Pa, Oa) = P(Sa)P(Pa)P(Oa)$. The model ranks the score of edges after merging the Faster-RCNN's scores and predicate scores. For every predicate class Pa, a low-dimensional attention vector (AP) is computed by fusing language and spatial features. The attention matrix of all predicates, which has the same number of rows as classes, is denoted as A. The authors calculate weights by conducting attention pooling. The weights are $W_S(A)$, $W_P(A)$,

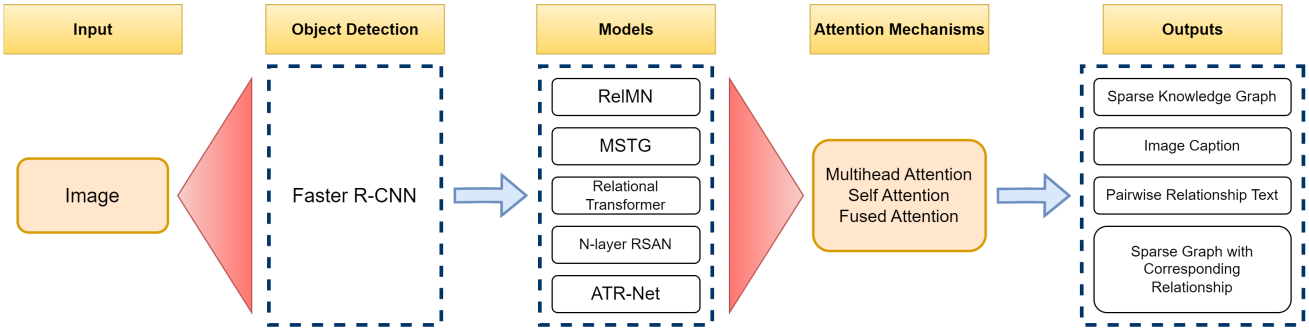

Fig. 2: A Generalised architecture of Scene Graph Generation Based on Attention Mechanism

and $W_O(A)$ for the subject, predicate, and object features, respectively on this attention. After computing the attention weights $W_S(A)x_S$, $W_P(A)x_P$, and $W_O(A)x_O$ are determined as the projections of their visual feature Sa, Pa, and Oa. The VTransE equation is reformulated into:

$$W_P(A)x_P \approx W_O(A)x_O - W_S(A)x_S \tag{1}$$

For each object pair, both relevance and predicate are needed to be decided. The procedure is viewed as two independent tasks and learn per-task attention weights.

Similar to this, Tian et al. introduced the Multi-level Semantic Tasks Generation (MSTG) [15] architecture to carry out visual tasks including visual connection recognition and picture captioning. In the image captioning part the model used a high-level semantic attention mechanism and fused spatial to improve the accuracy. To produce the caption, the attention mechanism dynamically picks image characteristics from the caption regions' image feature vector denoted as relevant depending on the LSTM network's hidden state. The model chooses the previous hidden state of the LSTM network as well as the feature vector and name attribute vector of the target objects in the picture in the decoder. Then establishes the feature vector's and name attribute vector's current weights. In order to focus on each item in the picture, the model employs Softmax normalization to determine the weight distribution of characteristics relating to the target object.

On the contrary, Koner et al. [14] proposed a Relation transformer model with an introduction of effective propagation of context utilizing a transformer encoder-decoder architecture across all nodes and edges. After detecting the object using Faster-RCNN, the initial feature of each node is obtained by applying a linear projection layer on concatenated features. Here, the attention mechanism is used to extract the global and local context of edges. For each edge, they first extracted the initial edge features by taking its visual features and spatial features and applying a linear projection layer onto them. Then, they applied self-attention in two ways: Edge-to-Nodes (E2N) and edge-to-edge (E2E). E2N generates the necessary global context for an edge and E2E generates the local context as defined in Equation 2.

$$Attention(X, O, P) = softmax(\frac{XO^T}{\sqrt{d_k}})P \tag{2}$$

In Equation 2, d_k represents the scaling factor and query X, keys O, and values P are a set of learnable matrices. Each value's weight is established by dividing a query matrix by the key that it belongs to, and the result is derived as a weighted sum of the values. Similarly, A relational self-attention unit is presented by Li et al [22] that simultaneously models relation contexts and objects. The vanilla self-attention model serves as the foundation of the relational self-attention paradigm. A collection of m input features X is transformed into m output features Z by the vanilla model. First, three parallel linear layers are supplied with the input features X to produce the query X, keys O, and values P. The output features shown in 2 are calculated by an attention function where the multiplication of X and O are divided by the scaling factor$\sqrt{d_k}$ and the overall value of P is summed up with it. For the self-attention module, a generalized version of the self-attention mechanism with a relationship prior is used. The fundamental associations between x are encoded by the three-order tensor R.

$$Z = A(X, O, P, R) = softmax(\frac{XO^T}{\sqrt{d_k}} + \phi R)P \tag{3}$$

The relational self-attention (RSA) block was produced by combining the feed-forward network (FFN) with the relational MHA (RMHA) module. By deeply cascading n RSA blocks, the deep relational self-attention network (RSAN) architecture is created. The modified characteristics were recovered by feeding the input features into a deep RSAN model that took into consideration both node-level and edge-level interactions. The image is represented by four distinct kinds of heterogeneous features: visual, spatial, linguistic, and relational. The bounding box data from the identified items is used to get the spatial characteristics. Finally, the spatial features of two objects are used to create the relation features. All four feature representations are fused to obtain fused features Z. Both the fused features and relation features R are fed into all the RSA blocks in the RSA network. The object and relation contexts eventually interact when the RSA blocks are stacked, characterizing more precise object and relation semantics. On a similar note, a novel integrated sparse deep sparse graph attention network [16] for SGG is proposed to achieve effective message passing and get detailed contextual information. The whole process consists of three distinct parts:

Year	Model	Input	Model Description	Type of attention	Output	VG			
						SGGen/SGCls		PredCls	
						R@50	R@100	R@50	R@100
2019	ATR-Net	Fully connected graph	Faster R-CNN + Word embedding	Multi-Head	A sparse graph with corresponding relationship	22.7	28.14	**81.63**	**89.03**
2021	MSGT	Set of object pair and relationship region	Faster R-CNN + ROI Pooling + GGNN+ Attention mechanism	Fused Attention mechanism	Image Caption	40.63	41.36	52.84	58.72
2021	RSAN	An arbitary image	Faster R-CNN + N layer RSAN + multitask classifier	Multi-head	Caption for pairwise relationship	37	9.8	66.1	67.9
2021	Relation Transformer	A single image	Faster R-CNN + (N2N, E2N, E2E) attention + Predicate Classification	Self Attention Mechanism	Image Caption	**43.6**	**43.7**	68.5	68.5
2022	Deep Scene Graph Attention network	Set of subject and object pairs	Faster R-CNN + RelMN + GAT	Multihead	Sparse knowledge graph	40.2	41.1	67.7	69.3

TABLE I: Recent work on Scene Graph Generation (SGG) based on attention mechanism along with model definition and performance on VG datasets (All values in percentage).

the Bounding Box Module, the Sparse Graph Module, and the Message Passing Graph Module. In order to classify edges into foreground and background categories, the sparse graph module makes use of a Relationship Management Network (RelMN). To create a sparse graph, partial background edges and foreground edges are automatically chosen. Then, it chooses all object pairings whose foreground edges are expected. The attention mechanism is mainly applied in the message-passing graph module. When this occurs, the sparse network's nodes and edges concurrently learn both object and relationship characteristics. The node features and the edge features learn their weights based on the inherent weight and attention. Encoding the inherent relationships between classes is aided by the attention weights of nodes and edges, as well as the intrinsic weight from prior statistical probability. With the assistance of the multi-head attention mechanism and the prior probability matrix of category correlations, the model concentrated on retrieving statistical co-occurrence information and contextual signals from the massive dataset. By using the output characteristics, the model categorized the objects and connections. Particularly, the output node features were used to categorize the object categories, while the output edge features are used to categorize the relationships.

III. Performance Evaluation

In this section, we have described the performance evaluation on different datasets of the models. We initially provided a few standard evaluation methods and standards that are widely used to evaluate SGG.

A. Tasks

The main task is to extract visual and contextual information from a given image through scene graphs. The generation of scene graphs is a cumulative workflow of some sub-tasks; object detection with object coordinates and feature values, identifying their category labels, anticipating relationships between each pair of these items, and classifying them. Several of the following common sub-tasks were studied by the bulk of prior research using their SGG models.

*1) **Phrase Detection (PhrDet)**:* : It generates an output as a text label in a *subject-predicate-object* manner. Side by side the relationships between the subject and object are localized by a bounding box.

*2) **Predicate Classification (PredCls)**:* : Region proposal network generates a set of object pairs. From those sets of objects, PredCls does is, it determines whether pairings interact, then categorize the predicate of each pair.

*3) **Scene Graph Classification (SGCls)**:* : It takes some localised objects as input. Then it predicts object categories and predicates in every pairwise relationship.

*4) **Scene Graph Generation (SGGen)**:* : From a given image it detects objects, extracts feature appearance from them, generates a relationship between them known as a predicate, and predicts the predicate between the object pairs with a constraint of 50 percent overlapping of the bounding boxes of the *subject* and *object* with their ground truth.

B. Metrics

The traditional criterion for measuring SGG performance is ***Recall@k***. *Recall@k* determines the percentage of top k confident relationship predictions when the right relationship is predicted. There are many categories of recall@k metrics among them *R@50* and *R@100* are used commonly. On the other hand, some works use *R@20*. It is a more challenging metric than others used for more comprehensive evaluation. In this metric, k is the super parameter that determines the number of predictions allowed per object pair. Some techniques only allow only one relationship to be derived for a given object pair when computing *R@K*. On the contrary, some methods do not put any constraint on *R@K*. Recalls can be of two types macro recall and micro recall. Another modified version of R@K is mean Recall@K (mR@K). It collects each predicate individually before averaging R@K across all predicates.

C. Result Comparison

A total of 14 datasets are used in the performance evaluation of the five papers. Among them, we utilized the Visual Genome (VG) dataset to compare the state-of-the-art models. The reason for choosing the VG dataset as it was used to

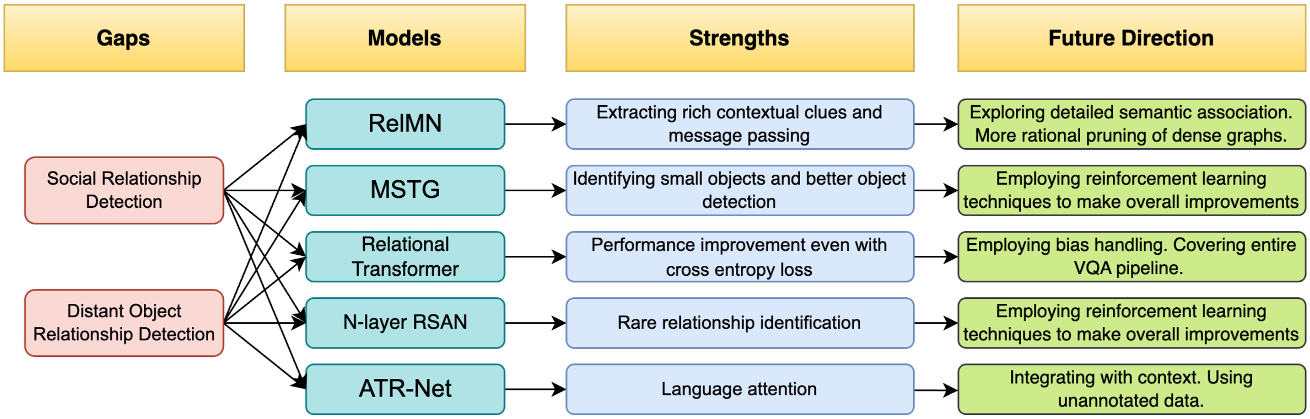
Gaps
Models
Strengths
Future Direction
Social Relationship Detection
Distant Object Relationship Detection
RelMN
MSTG
Relational Transformer
N-layer RSAN
ATR-Net
Extracting rich contextual clues and message passing
Identifying small objects and better object detection
Performance improvement even with cross entropy loss
Rare relationship identification
Language attention
Exploring detailed semantic association. More rational pruning of dense graphs.
Employing reinforcement learning techniques to make overall improvements
Employing bias handling. Covering entire VQA pipeline.
Employing reinforcement learning techniques to make overall improvements
Integrating with context. Using unannotated data.

Fig. 3: Gaps, strengths and future directions of SGG models.

test all the models by their respective authors which we are comparing in this study. Table I highlights the comparative results. We split the table into 7 parts. Where the last part shows the result on the VG dataset. Along with the Recall metric, Scene graph classification (SGGen/SGCls) and predicate classification are mentioned in the table. Here, ATR-Net used the Recall metrics and scored 89.03 (R@100) in predicate classification. It achieved high score than the other methods due to its attentional scheme that enables ATR-Net to fully use visual elements. The other methods also used similar metrics to evaluate their performance. MSTG scored 58.84 at R@50 in the predicate classification and 41.31 at R@50 in the SGCls. The recall rate of other attention-based methods also shows that their performance is quite close to the existing state of the arts. Overall the performance of ATR-Net is more impressive than others.

IV. Discussion

Through the analysis of state-of-the-art attention-based SGG models, we identify strengths and weaknesses of the works. We assume these works can be extended further in a multitude of ways and a summary of our findings is illustrated in Figure 3.

A. Explored Gaps

Upon going through several SGG literature we have found some relatively less explored areas in the recent works, which include *Distant Object Relationship Detection* and *Social Relationship Detection*.

1) Distant Object Relationship Detection: The spatial distance between items, which ensures that no relationship will exist between two distant objects, greatly influences the selection of possible effective relationships and the formation of the final relationships in the scene graph. These kind of relationships are more prevalent in larger images [23]. These relationships should be considered to increase the robustness of the scene graph.

2) Social Relationship Detection: Scene graphs are a useful tool for studying human-human interaction, and social ties may be inferred from these kinds of relationships. We think that social relationship identification is a highly essential area of research since it may be utilised to better comprehend the situations. A greater range of real-world practical application values are produced by SGG models built on large-scale datasets since they can even extract hidden social interactions from the visual data [24].

B. Strengths

As far as strengths are concerned, for the work seen in ATR-Net [21] it is evident that language attention is its strongest point. In the work of Relation Transformers [14] it is seen that even with simple cross-entropy loss, their novel context propagation for both objects and edges significantly improves performance. But one of its downfalls is the object relation semantics. It sometimes fails to predict the predicate in a subject-object pair accurately. They show that this is due to a bias in the dataset but improvement is necessary here. The RSAN architecture [22] which uses stacked self-attention blocks is seen to be the better model when it comes to finding rare relationships. RSAN also has a lower percentage of producing incorrect scene graphs. In the RelMN architecture, [16] it is seen that due to the multi-head attention mechanism relationship detection is significantly improved. The RSAN also employs the multi-head attention mechanism in its inner workings. The MSTG [15] architecture shows that the addition of fused attention increases the performance of model-generated description. Additionally, it improves the overall efficiency of object detection and is capable of accurately identifying tiny things. In order to simultaneously improve the accuracy of object recognition, scene graph generation, and photo captioning, this approach leverages mutual correlation and replenishment of numerous semantic properties. The three independent semantic layers' characteristics are updated regularly by the model.

C. Future Directions

Analysing the explored gaps and strengths of the developed models, we assume that a multitude of future directions can be explored in this research arena. Highlighting future directions include:

- ATR-Net concept could be integrated with context and investigating ways to use unannotated data to improve single-instance prediction.
- Data balancing techniques for handling the bias could be fruitful for Relation Transformers. They can also integrate cutting-edge SGG into their architecture to cover the entire visual question-answering (VQA) pipeline.
- Considering RelMN, exploration of detailed semantic associations between object pairs and more rational pruning of dense graphs for high-level vision tasks can be a good future work.

During the SGG, the implementation of Reinforcement Learning (RL), which may be used to fine-tune attention scores in either the spatial or temporal domains, is another fascinating potential direction for the future. RL makes it possible to make iterative improvements through trial and error, which provides a significant boost in the performance of deep learning tasks [25]. We make the assumption that RL is capable of successfully filtering the attention score, extracting more coarse-grained information from the image, and developing more pertinent links among the subject, predicate, and object triplet.

V. Conclusion

Attention-based modules recently made a positive impression on solving problems that require neural networks with larger context propagation ability. Thus, the task of solving SGG has also been approached using it. In this paper, the five recent attention-based SGG are described in the Exploration of State Of The Art section. We summarized the methods in order to convey the scope of improvements in SGG to the concerned researchers in this field. Moreover, the discussed strengths and future directions are expected to be useful to extend the growth of attention-based models in SGG research. In the future, we like to extend our work to develop a comprehensive guideline for SGG using the fine-tuning of attention scores by Reinforcement Learning.

References

[1] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[2] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3668–3678.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[4] A. Zaman, M. S. Majib, S. A. Tanjim, S. M. A. Siddique, F. Ashraf, S. Islam, A. H. M. M. Morshed, S. T. Shahid, I. Hasan, O. Samir *et al.*, "Phoenix: Towards designing and developing a human assistant rover," *IEEE Access*, vol. 10, pp. 50 728–50 754, 2022.

[5] A. Zaman, M. S. Majib, S. A. Tanjim, S. M. A. Siddique, S. Islam, M. S. Aadeeb, N. I. Khan, R. Haque, M. R. U. Islam, M. R. F. Faisal *et al.*, "Uvc-purge: A novel cost-effective disinfection robot for combating covid-19 pandemic," *Ieee Access*, vol. 10, pp. 37 613–37 634, 2022.

[6] J. Dijk, A. W. van Eekeren, O. R. Rojas, G. J. Burghouts, and K. Schutte, "Image processing in aerial surveillance and reconnaissance: from pixels to understanding," in *Electro-Optical Remote Sensing, Photonic Technologies, and Applications VII; and Military Applications in Hyperspectral Imaging and High Spatial Resolution Sensing*, vol. 8897. SPIE, 2013, pp. 76–92.

[7] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2017, pp. 1–4.

[8] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5831–5840.

[9] Y. Chen, Y. Wang, Y. Zhang, and Y. Guo, "Panet: A context based predicate association network for scene graph generation," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 508–513.

[10] Y. Teng and L. Wang, "Structured sparse r-cnn for direct scene graph generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 437–19 446.

[11] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

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

[13] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," *Advances in neural information processing systems*, vol. 27, 2014.

[14] R. Koner, P. Sinhamahapatra, and V. Tresp, "Scenes and surroundings: Scene graph generation using relation transformer," *arXiv preprint arXiv:2107.05448*, 2021.

[15] P. Tian, H. Mo, and L. Jiang, "Scene graph generation by multi-level semantic tasks," *Applied Intelligence*, vol. 51, no. 11, pp. 7781–7793, 2021.

[16] H. Zhou, Y. Yang, T. Luo, J. Zhang, and S. Li, "A unified deep sparse graph attention network for scene graph generation," *Pattern Recognition*, vol. 123, p. 108367, 2022.

[17] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[18] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 335–351.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[20] T. Chen, W. Yu, R. Chen, and L. Lin, "Knowledge-embedded routing network for scene graph generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6163–6171.

[21] N. Gkanatsios, V. Pitsikalis, P. Koutras, and P. Maragos, "Attention-translation-relation network for scalable scene graph generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[22] P. Li, Z. Yu, and Y. Zhan, "Deep relational self-attention networks for scene graph generation," *Pattern Recognition Letters*, vol. 153, pp. 200–206, 2022.

[23] J. Zhang, Y. Kalantidis, M. Rohrbach, M. Paluri, A. Elgammal, and M. Elhoseiny, "Large-scale visual relationship understanding," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 9185–9194.

[24] H. Yoo, T. Eom, J. Seo, and S.-I. Choi, "Detection of interacting groups based on geometric and social relations between individuals in an image," *Pattern Recognition*, vol. 93, pp. 498–506, 2019.

[25] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.