a representative image from an image collection has been introduced in the album summarization task. To find the common occurring visual features as a representation, estimating visual features using clustering methods is a popular technique. In addition, by describing an image collection with keywords or tags, the summarization focuses on finding shared concepts of an image collection. However, this approach is still restricted to specific domains. Recently, an image collection captioning task [16] has been introduced to summarize the commonly occurring visual features and relations of an image collection into a caption.

The main discussion of an image collection summarization task is to find common visual features and relationships, such as objects, events, or concepts shared in an image collection. Finding shared concepts by integrating external knowledge into the summarization model has become popular in summarizing an image collection [17], [22]. In contrast, integrating external knowledge into a model can

[1] Nagoya University
[2] Kyoto University
[3] RIKEN
[a)] phueaksrii@cs.is.i.nagoya-u.ac.jp

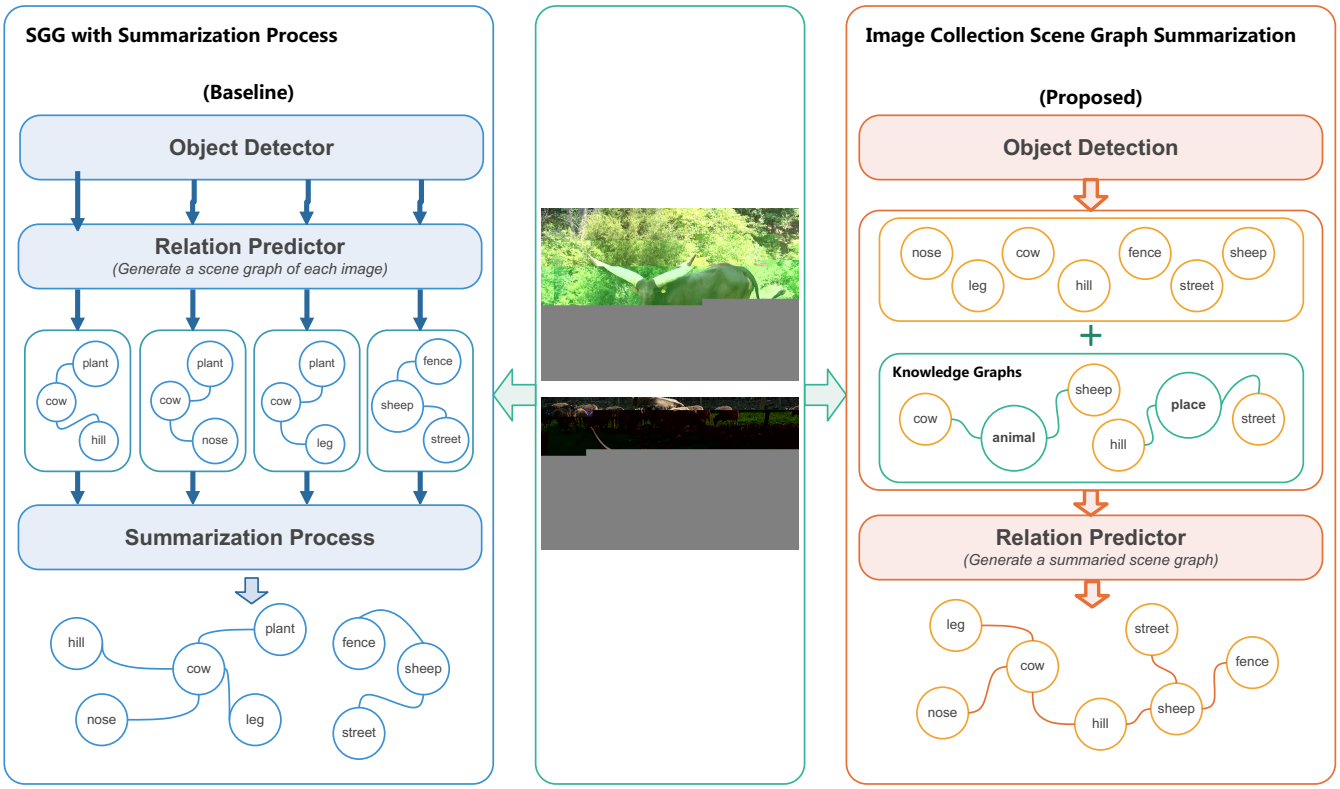

Fig. 1: Overview of the image collection scene graph summarization model, the proposed method (right) compared with the baseline method (left) with scene graph generation and summarization process. The proposed method takes advantage of knowledge graphs to relax the relation predictor of scene graph generation in generating a summarized scene graph.

by merging all scene graphs into a combined scene graph and then selecting a representative scene graph using graph theory in the process. To summarize an image collection, our approach aims to describe image collection in a scene graph form, focusing on integrating external knowledge into the learning process.

2.2 Scene Graph Generation

Scene graph generation [5] is a popular technique in describing relationships between objects in an image. The relationships of objects are generally represented in triplets which consist of subject, predicate, and object. In recent years, scene graph generation has been widely introduced and implemented on the Visual Genome dataset [10], a popular scene graph generation dataset. In addition, scene graph generation is also adapted to various applications, such as image captioning [12] and image retrieval [18], and has shown to advance their results. Various techniques are introduced in scene graph generation; Neural Motif [26] is built with Faster-RCNN with ResNext-101 [24], then computes and propagates through Bidirectional LSTM for predicting relations. From the long-tail problem of the scene graph dataset, the most recent work [21] aims to introduce a technique to solve the bias of the dataset.

2.3 Knowledge Base

Knowledge base is widely used to enrich models, especially text-generation models [25]. ConceptNet [19] and Wikipedia dataset[*1] are popular knowledge bases that are used in the generation process. ConceptNet is a knowledge graph that represents general knowledge and commonsense information, while Wikipedia is structured knowledge data with detailed information on each topic. In recent years, the knowledge graph has become a popular knowledge base on various generation processes, mainly focusing on capturing commonsense reasoning during the generation. To tackle the long-tail issues of scene graph generation mentioned above, integrating knowledge graphs to improve generation is widely introduced, and the result shows the advantage of using it. Moreover, a knowledge graph is additionally implemented in an image retrieval task which aims to reason on the semantic context and generalize the concepts inside an image [29].

3. Proposed Method

Scene graph generation has become a popular task to describe an image in a scene graph form. Using scene graph generation incorporating external knowledge to fnd shared concepts of an image collection has become a trend in recent years. However, identifying the exact shared concepts still relies on the summarization process and external knowledge.

Following the idea of building a scene graph summarization model by integrating commonsense knowledge into the relation predictor instead of summarizing all scene graphs

[*1] https://www.tensorfow.org/datasets/catalog/wikipedia/ (Accessed May 26, 2023)

from scene graph generation with the summarization process as shown in Fig. 1, we first introduce a *Scene Graph Summarization Model*, which focuses on relaxing the relation predictor. Next, we introduce *Knowledge Integration*, used in the scene graph summarization model. Lastly, we introduce the *Multiple Image Inference* for an image summarization.

3.1 Scene Graph Summarization Model

Following the idea of relaxing the relation predictor for an image summarization task, we adopt Neural motifs [26], consisting of an object detector, an object context, and a relation predictor incorporating external knowledge into the generation process.

Object Detector The first part is an object detector to detect a set of region proposals, Faster R-CNN [3] with RestNet-101 is used as a detector backbone. Following the scene graph generation, from each image; a set of region proposals $B = \{b_1, ..., b_n\}$ is predicted. Each region proposal consists of feature vectors $\mathbf{f}$ and object label probabilities for training. To generate a summarized scene graph, we modify an object detector for multiple images which will be explained in Section 3.3.

Object Context The second part is the object context predictor to construct a contextualized representation of a set of region proposals into a linear sequence, $[(b_1, \mathbf{f}_1, \mathbf{l}_1), ..., (b_n, \mathbf{f}_n, \mathbf{l}_n)]$. Then, the computation uses a bidirectional LSTM [4] as:

$$\mathbf{C} = \text{biLSTM}([\mathbf{f}_i; W_1 \mathbf{l}_i]_{i=1,...,n}); \quad (1)$$

where $\mathbf{C}$ is a set of object contexts, in which each object context contains the hidden state of each element in the linearization of B, W_1 is a parameter that maps to the distribution prediction represented in the matrix, and $\mathbf{l}_i$ is a probability vector of object labels. Each object context is used to decode a category label with an LSTM as:

$$\mathbf{h}_i = \text{LSTM}_i([\mathbf{c}_i; \mathbf{o}_{i-1}]); \quad (2)$$

$$\mathbf{o}_i = \text{argmax}(W_o \mathbf{h}_i) \in \mathbb{R}^{|C|} \text{ (one-hot)}; \quad (3)$$

where $\mathbf{c}_i$ is an object context vector, $\mathbf{h}_i$ is a hidden state that is used in the relation predictor. W_o is a parameter that maps to the hidden state.

Relation Predictor The obtained object contexts by the previous process are used in the relation predictor, which builds for generating a representation, a set of regions as B and objects using a bidirectional LSTM as:

$$\mathbf{E} = \text{biLSTM}([\mathbf{c}_i; W_2 \mathbf{o}_i]_{i=1,...,n}); \quad (4)$$

where $\mathbf{E}$ is a set of edge contexts from the relation predictor, in which each edge context contains the states of the bounding box region and W_2 is a mapping parameter of $\mathbf{o}_i$. Each edge context is combined with the knowledge embedding and predicts the relation of each pair as:

$$\mathbf{g}_{i,j} = (W_h \mathbf{e}_i) \circ (W_t \mathbf{e}_j) \circ \mathbf{f}_{i,j}; \quad (5)$$

$$\mathbf{r}_{i,j} = \text{argmax}([\mathbf{g}_{i,j}; \mathbf{e}_{\text{kb}}^{(i,j)}] W_r); \quad (6)$$

where $\mathbf{e}_i$ and $\mathbf{e}_j$ are edge context vectors of head and tail, W_h and W_t are parameters of heads and tails, $\mathbf{f}_{i,j}$ is a feature vector for the union of two bounding boxes, W_r is a parameter that maps to the relation predictor, and $\mathbf{e}_{\text{kb}}$ is a knowledge embedding vector which will be explained in Section 3.2.

3.2 Knowledge Integration

From the idea of integrating knowledge into the relation prediction of the scene graph generation, we build a 1-hop word embedding knowledge graph from ConceptNet. From the various aspect of the relation representation in ConceptNet, we build a knowledge graph focusing on semantic relations, such as "$relatedTo$", "$similarTo$", and "$synonym$."

To build a 1-hop knowledge graph, we first give a class pair from an object detector as (x, y) and then employ the connection between (x, y) as (V_x, V_y). Lastly, we retrieve a set of all possible paths from V_x to V_y as $P_{(x,y)}$

Next, we build a graph convolutional network using the GlobalSortPool operator [28] to encode a knowledge graph into a knowledge feature vector as:

$$\mathbf{e}_{\text{kb}}^{(x,y)} = \text{GlobalSortPool}(\mathbf{N}^{(x,y)}); \quad (7)$$

where $\mathbf{N}^{(x,y)}$ is all embedding nodes from $P_{(x,y)}$, encoded by GloVe word embedding [15].

3.3 Multiple Image Inference

To generate a summarized scene graph of an image collection, we modify an object detector to parse multiple images into a relation predictor. Based on a single image scene graph generation, we build an object detector backbone to detect image features, $\mathbf{f}_i$, and proposals b_i, Then, we combine all image features and all proposals as:

$$F = \{[\mathbf{f}_{i,1}, ..., \mathbf{f}_{n,m}]_{i=1,..,n}\}; \quad (8)$$

$$P = \{[p_1, ..., p_m]_{i=1,..,m}\}; \quad (9)$$

where F is a set of feature vectors of all images, m is the number of region proposals, and P is a set of proposals of all images. All predicted sets of region proposals $P = \{p_1, ..., p_m\}$ and each region proposal consists of feature vector $\mathbf{f}$ and an object label probability are used in an object context for the summarization process.

4. Experiment

4.1 Dataset

Due to lacking annotation in image summarization datasets, we used two datasets in the experiment consisting of the MS-COCO dataset [11] and the Visual Genome dataset [10]. In the training process and the preliminary evaluation, we used VG200 [27], which is based on the Visual Genome dataset consisting of 50 relationships and is balanced in category frequency. It contains 101,174 images

Table 1: Evaluation on an image of the VG200 compared to other baselines

MODEL	SGDets			SGCls			PredCls		
	R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
VCTree [20]	0.2453	0.3193	0.3621	0.3401	0.3748	0.385	0.5902	0.6542	0.6718
Neural Motif [26]	0.2548	0.3278	0.3716	0.3687	0.4018	0.4102	0.5846	0.6518	0.6701
Proposed	0.2277	0.2864	0.3262	0.3594	0.3863	0.3943	0.4335	0.4974	0.5167

from the MS-COCO dataset. To experiment on a scene graph image collection summarization task, we built a testing set of an image collection by extending from the MS-COCO testing dataset using VSE++ [2], which is an image retrieval task by estimating the similarity of image contexts and image captions. Following the Karpathy split [7] on the MS-COCO dataset, our initial testing set was selected from 5,000 images of the MS-COCO testing set. Then, we retrieved 5 images to build each collection, making our testing set contain 5,000 collections with 6 images each.

4.2 Training Strategy

With the limitation in scene graph summarization datasets, we first trained the scene graph generation on a single image of the Visual Genome dataset. After training, we extended the model to a multiple-image inference process. In the training phase, we trained the model with the training split of VG200 with 0.12 of the initial learning rate. We used Adam [8] for optimization and cross-entropy loss as the loss function. We observed $SGDet$ recall to select the best checkpoint for an image collection scene graph summarization task.

4.3 Evaluation

The evaluation of the proposed method consists of two phases. As we modify the relation prediction of the scene graph generation model, we first evaluated the proposed method on VG200 compared with the baseline to ensure that the proposed method still sustains good results on a single image scene graph generation. We observed *Scene Graph Classification Recall (SGCls Recall)*, *Predicate Classification Recall (PredCls Recall)*, and *Scene Graph Detection Recall (SGDet Recall)* on a single image. We lastly evaluated the proposed method with an image collection.

4.4 Results

We will first see the result of the proposed method on image scene graph generation on VG200, which shows the proposed method with no bias. Based on the evaluation in scene graph generation with $SGCls$, $PredCls$, and $SGDet$ as shown in Table 1, the proposed method still sustains an image scene graph generation task. The results show that enhancing the relation predictor still keeps good results on a single image scene graph generation. Due to focusing on SGDet Recall to select the best checkpoint, this experiment shows that the proposed method achieves results not much lower than the baseline.

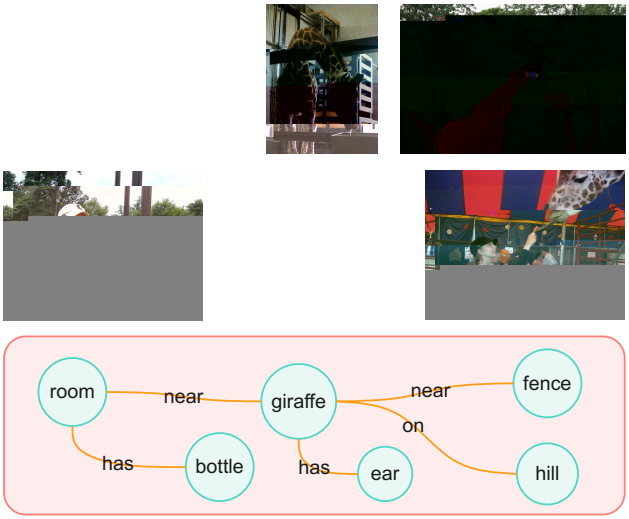

Fig. 2: Example of the proposed method showing the summarization of an image collection into a scene graph.

Next, we will see the output of an image collection scene graph summarization. An example is shown in Fig. 2. It demonstrates that the proposed method can generate a summarized scene graph containing the main object \girafe" and image locations from most images.

5. Conclusion

We introduced a novel scene graph summarization model incorporating external knowledge following the idea that aims to relax the relation predictor in the training process for an image collection. A preliminary experiment demonstrated promising results of the proposed method. In addition, the result further showed a good result on an image collection of the MS-COCO dataset. In the future, we plan to introduce a new evaluation method on scene graph summarization that focuses on the context coverage of a summarized scene graph on a single scene graph and implement the proposed method on an annotated MS-COCO panoptic scene graph summarization dataset [14].

Acknowledgments

This work was supported in part by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, through Grant-in-Aid for Scientifc Research under Grant JP21H03519. The computation was carried out using the General Projects on the supercomputer \Flow" at Information Technology Center, Nagoya University.

References

[1] Celikkale, B. et al.: Generating Visual Story Graphs with Application to Photo Album Summarization, *Signal Proc.: Image Commun.*, Vol. 90, p. 116033 (2021).

[2] Faghri, F. et al.: VSE++: Improving Visual-Semantic Embeddings with Hard Negatives, *29th Brit. Mach. Vis. Conf.* (2018).

[3] Girshick, R.: Fast R-CNN, *16th IEEE Int. Conf. Comput. Vis.*, pp. 1440{1448 (2015).

[4] Graves, A. et al.: Long Short-Term Memory, *Supervised Sequence Labelling with Recurrent Neural Netw.*, pp. 37{45 (2012).

[5] Han, X. et al.: Image Scene Graph Generation (SGG) Benchmark, *Comput. Res. Reposit. arXiv Preprint*, arXiv:2107.12604 (2021).

[6] Johnson, J. et al.: Image Retrieval Using Scene Graphs, *2015 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3668{3678 (2015).

[7] Karpathy, A. et al.: Deep Visual-Semantic Alignments for Generating Image Descriptions, *2015 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3128{3137 (2015).

[8] Kingma, D. P. et al.: Adam: A Method for Stochastic Optimization, *3rd Int. Conf. Learn. Representat.* (2014).

[9] Kohonen, T.: *Self-Organizing Maps*, Vol. 30, Springer Science & Business Media (2012).

[10] Krishna, R. et al.: Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations, *Int. J. Comput. Vis.*, Vol. 123, No. 1, pp. 32{73 (2017).

[11] Lin, T.-Y. et al.: Microsoft COCO: Common Objects in Context, *13th Euro. Conf. Comput. Vis.*, Vol. 5, pp. 740{755 (2014).

[12] Milewski, V. et al.: Are Scene Graphs Good Enough to Improve Image Captioning?, *Joint Conf. 59th Annu. Meeting Assoc. Comput. Linguist. and 11th Int. Joint Conf. Nat. Lang. Process.* (2020).

[13] Park, H.-S. et al.: A Simple and Fast Algorithm for K-Medoids Clustering, *Expert Syst. Appl.*, Vol. 36, No. 2, pp. 3336{3341 (2009).

[14] Pasini, A. et al.: Semantic Image Collection Summarization with Frequent Subgraph Mining, *IEEE Access*, Vol. 10, pp. 131747{131764 (2022).

[15] Pennington, J. et al.: GloVe: Global Vectors for Word Representation, *2014 Conf. Empir. Methods Nat. Lang. Process.*, pp. 1532{1543 (2014).

[16] Phueaksri, I. et al.: Towards Captioning an Image Collection from a Combined Scene Graph Representation Approach, *29th Int. Conf. Multimed. Model.*, Vol. 1, pp. 178{190 (2023).

[17] Samani, Z. R. et al.: A Knowledge-Based Semantic Approach for Image Collection Summarization, *Multimed. Tools Appl.*, Vol. 76, No. 9, pp. 11917{11939 (2017).

[18] Schroeder, B. and Tripathi, S.: Structured Query-Based Image Retrieval using Scene Graphs, *2020 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 178{179 (2020).

[19] Speer, R. et al.: ConceptNet 5.5: An Open Multilingual Graph of General Knowledge, *31st AAAI Conf. Artif. Intell.*, pp. 4444{4451 (2017).

[20] Tang, K. et al.: Learning to Compose Dynamic Tree Structures for Visual Contexts, *2019 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 6619{6628 (2019).

[21] Tang, K. et al.: Unbiased Scene Graph Generation from Biased Training, *2020 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3716{3725 (2020).

[22] Trieu, N. et al.: Multi-Image Summarization: Textual Summary from A Set of Cohesive Images, *Comput. Res. Reposit. arXiv Preprint*, arXiv:2006.08686 (2020).

[23] Wang, B. et al.: Hierarchical Photo-Scene Encoder for Album Storytelling, *33rd AAAI Conf. Artif. Intell.*, Vol. 33, pp. 8909{8916 (2019).

[24] Xie, S. et al.: Aggregated Residual Transformations for Deep Neural Networks, *2017 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1492{1500 (2017).

[25] Yu, W. et al.: Knowledge-Enriched Natural Language Generation, *2021 Conf. Empir. Methods Nat. Lang. Process.*, pp. 11{16 (2021).

[26] Zellers, R. et al.: Neural Motifs: Scene Graph Parsing with Global Context, *2018 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 5831{5840 (2018).

[27] Zhang, J. et al.: Graphical Contrastive Losses for Scene Graph Parsing, *2019 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 11527{11535 (2019).

[28] Zhang, M. et al.: An End-to-End Deep Learning Architecture for Graph Classifcation, *32nd AAAI Conf. Artif. Intell.*, Vol. 32, pp. 4438{4445 (2018).

[29] Zhang, W. et al.: Joint Optimisation Convex-Negative Matrix Factorisation for Multi-Modal Image Collection Summarisation based on Images and Tags, *IET Comput. Vis.*, Vol. 13, No. 2, pp. 125{130 (2019).