Towards Captioning an Image Collection using Scene Graph

PHUEAKSRI Itthisak[1,a)] Marc A. KASTNER[2] Yasutomo KAWANISHI[3]
Takahiro KOMAMIZU[1] Ichiro IDE[1]

Abstract

Most content summarization models are targeted to summarize the text content of a set of texts, and it is still challenging to summarize the visual content of a collection of images. In this presentation, we propose a method for summarization of the visual content of an image collection by combining scene graphs of multiple images and generating a single caption that describes the image collection. We also present a method to find a common context word to improve the description of the image collection by incorporating ConceptNet. In this method, we build word relations of different words, such as synonym words and category words, to find the representative word in each word relation. The proposed method is evaluated on the MS COCO dataset compared with other text generation methods, showing a promising direction for this research.

1. Introduction

The recent increase of images on the Web and on Social Media became a challenge to describe their visual contents in text. For this, image captioning is a popular task that generates an image description as a sentence. However, current image captioning methods are limited to a single image. Even for methods that aim to summarize the visual content of a collection of images, their output is restricted to tags that have limited description ability. In this presentation, in order to better describe the visual content of a collection of images, we propose a method to summarize it in one sentence.

Scene graph generation is a popular method to describe the relationships between objects and actions in images [1], [4], [5]. A scene graph is a set of edges consisting of subject, predicate, and object, generated from an image. We combine multiple scene graphs of images in an image collection into a combined scene graph and then generate a caption from it as a summary of the visual content in the image collection.

With the idea of captioning the summary of an image collection, we propose the image summarization framework as illustrated in Fig. 1. Given a collection of images, we first extract features and scene graphs of each image. We then design two processes to merge the scene graphs of all images and generate a summarized scene graph for the whole image collection. Next, we build word communities and find the most representative word in each word community. Then we generate a sentence from the summarized scene graph. To improve the generalization of the sentence, we finally refine the generated sentence by implementing noun phrase replacement with the representative word of each community.

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

2. Related Work

2.1 Scene Graph Generation

Scene graph generation [3] is a method used to describe image contexts. Many scene graph generation methods mainly start by finding the object regions by using Fast R-CNN [14] as an object detector. They then find the relationship between objects in both local context and global context, for example, Neural Motif [2], and RelDN [6]. Their model backbone is implemented with many states-of-the-art object detectors such as VGG [16], and ResNet [15].

2.2 Text Generation

For text generation methods, we consider two categories; text-to-text generation [22] and image-to-text generation [1], [4], [5]. Many methods are proposed for the former, but the main idea summarizes a text article into sentences in two main aspects; extractive and abstractive. The extractive summarization is proposed in many methods such as T5 [21] which uses supervised text generation based on Wikipedia knowledge, and SUPERT [20] which is unsupervised text summarization method based on evaluating similarity scores between sentences in an article. Meanwhile, image-to-text generation methods describe images by sentences, mostly known as image captioning models. However, image captioning is typically one-one (image to sentence) or one-many (image to paragraph), but few researchers have studied at many-one (multiple images to sentence). Therefore our work mainly focusses on many-one type generation. We follow the idea of scene graph captioning, that many methods are proposed.

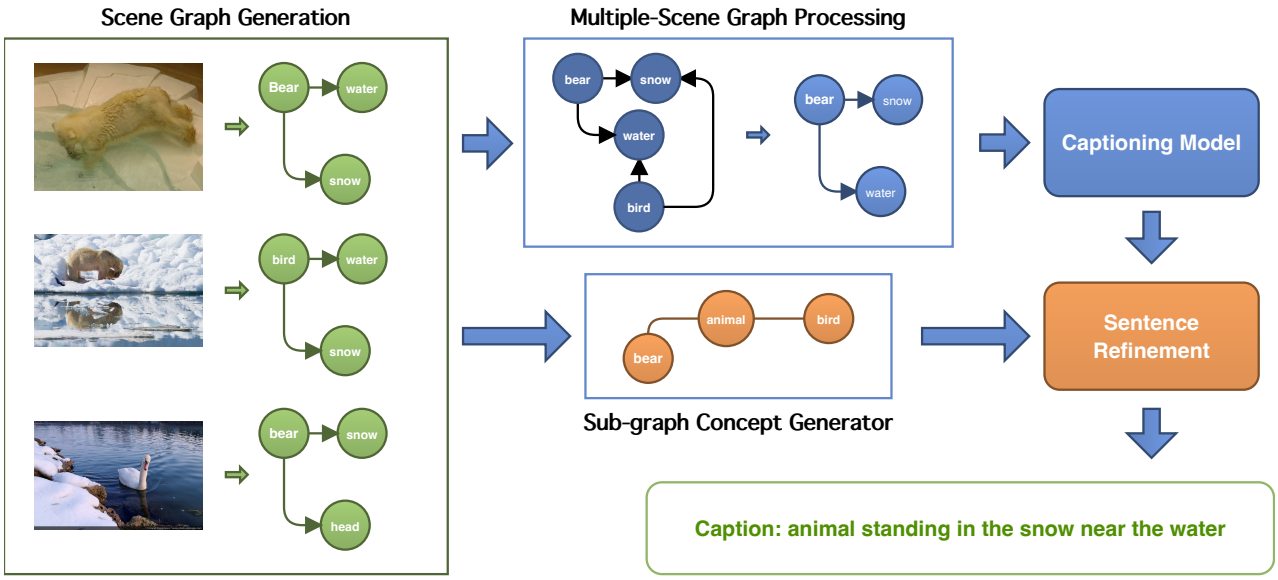

Fig. 1: Overview of the proposed method consists of five components: The first one is the Scene Graph Generator, which extracts a scene graph for each image. All scene graphs are passed into Multiple-Scene Graph Processing and Sub-graph Concept Generator. The former combines the scene graphs to find the representative graph, while the latter finds word communities from the scene graphs and representative in each community. Captioning Model generates the initial caption for the representative graph. In the final step, the initial caption and sub graph concept is passed into the Sentence Refinement module to output the final caption.

3. Proposed Method

The proposed method consist of five components. The first one is the Scene Graph Generation which extracts image features and scene graphs. All scene graphs are next passed into two modules: Multi-Scene Graph Processing to merge and select the representative graphs, and Sub-Graph Concept Generation to find general concepts of words in scene graphs through detecting word communities. Then a captioning model generates a sentence based on the representative graph. The generated sentence from the captioning model and the community word graphs are finally passed to the Sentence Refinement to output the final caption.

3.1 Scene Graph Generation

We use one of the current state-of-the-art ResNet101 [15]+Neural Motif [2] as a scene graph parser that is trained by the Visual Genome dataset [8] which is a popular practice for scene graph captioning. Recent works in image captioning show that the caption can be improved by manually cleaning up some duplicate labels [5]; we hence follow by reducing the label of the Visual Genome dataset from 2,500/1,000/500 to 1,600/400/20 of objects/attributes/relations.

3.2 Multiple-Scene Graph Processing

All scene graphs are merged into a single directed multigraph as shown in Eq.1. We also count the occasion of each node and the number of edges in the merging process to use in the selection step.

$$G = \bigcup_{i=1}^{n} g_i \tag{1}$$

From general graph theory, we consider graph characteristics in three aspects; degree centrality, closeness centrality, and betweenness centrality. In preliminary experiments, we found that implementing betweenness centrality to find the center node is the most efficient method compared with other centrality methods as shown in Eq. 2 We next selected 36 nodes and 100 relations from all the scene graphs by considering the ranking and relation between nodes.

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2}$$

where $\sigma_{st}(v)$ is the number of paths passing through v, and σ_{st} is a total number of the shortest paths from node s to t.

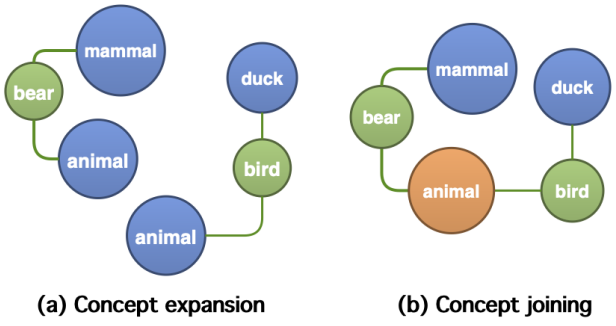

Fig. 2: Example of building a word community (a) Concept expansion by implementing ConceptNet, and (b) Concept joining with the same word.

3.3 Sub-graph Concept Generation

From various words extracted from the image collection, we try to generalize the visual content. In order to summarize the common content in an image collection, we make use of a popular text-based semantic network named ConceptNet [7] to generalize specific words into a general word. From the idea of text analysis based on word synonym relationship to build a community of words by a graph-based method [24], we propose a method focusing on nouns. First, the object words of the scene graph are lemmatized. Then ConceptNet is used to find the relationship between each node. From the experiments of recent text analysis that focus on synonyms [24], we also expand the concept over *isA* relation of ConceptNet to achieve a more accurate matching of common concepts as shown in Fig. 2. After mapping, non-degree nodes are dropped, and then sub-graphs are extracted from the whole graph. To estimate the representativeness of the common concept of each sub-graph, we encode all nodes by GloVe word embedding [9] then calculate the Euclidian distance and cosine similarity between each node. In the experiments, we select a word by calculating the highest node degree using cosine similarity as a weight to find each sub-graph word concept.

3.4 Captioning Model

The captioning model consists of Graph Convolutional Network (GCN) and the Attention-based LSTM model. We build a GCN to process the triplet of subject, predicate, and object features. Each feature is extracted from the Scene Graph Generation process, whose dimension is 1,024. The GCN maps the relationships between subject and predicate and between object and predicate. In the experiment, we test the number of graph layers around two and four layers to update mapping node and relation features in the graph. Next, we build the attention-based LSTM model [5] following the top-down LSTM captioning with two layers of attention-based LSTM in which both layer's sizes are set as 512. In training, we implement a learning rate decay of 0.8 for every eight epochs, initial learning rate of 0.0008, dropout of 0.5, and implement Adam optimization [25].

3.5 Sentence Refinement

To improve the caption, we modify the beam search of sentence generation to generalize the caption by mainly focussing on processing nouns in the result. We hence implemented noun phrase mapping with the sub-graph of concept community from the sub-graph concept generation by implementing POS tagging. In a preliminary experiment, we tested the beam size between 3 to 10, and found 5 as the best for generating and estimating the final caption.

4. Evaluation

4.1 Dataset

We use the MS COCO dataset [17] for the experiment. We further split the dataset following the Karpathy split [19], including 118K images for training, 5K images for assessment, and 5K images for testing. From the testing set, we considered testing image collection generation in three dimensions: image, caption, and both. Here, we implement VSE++ [18] to the testing set that considers both image and caption.

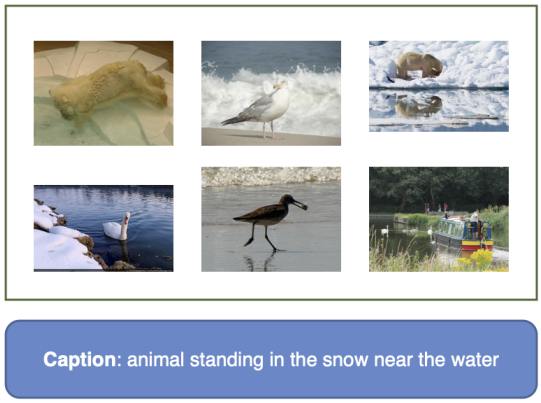

Fig. 3: Example of a caption generated by the proposed method with different main objects.

Table 1: Evaluation metrics for the result in Fig. 3 of the proposed model based on without concept generation (No CG) and with concept generation (CG) compared with SUPERT [20] and T5 [21].

Method	SUPERT [20]	T5 [21]	Proposed (No CG)	Proposed (CG)
ROUGE-1	**0.3328**	0.2999	0.3273	0.2365
ROGUE-2	0.0901	0.0728	**0.1279**	0.0748
ROUGE-L	0.2892	0.2711	**0.3557**	0.2117
CIDERBtw	0.4418	0.2523	**0.6732**	0.3936
WEEM4TS	0.1075	0.0874	0.1088	**0.1132**

4.2 Evaluation Metric

In the image captioning and text summarization field, various evaluation metrics are proposed. Since the proposed method mainly focusses on summarization, we use ROUGE-1, ROUGE-2, and ROUGE-L [13] as the evaluation metric. However, it is limited to the evaluation of abstractive summarization, we consider other evaluation metrics including CIDErBtw [10], which is the similarity evaluation between sentences, and WEEM4THS [11] which is a metric for evaluating abstractive summarization.

4.3 Result

5,000 images in the testing set are analyzed as result; we first generate the image collection similarity by using VSE++ with $K = 5$ to find five similar images. We then generate the caption for each image collection. Examples of results are shown in Fig. 3. To evaluate the proposed method, we compare it with supervised and unsupervised summarization models and show the improvement in summarization. The results are shown in the Table 1. We can see that the proposed method perform better than ex-

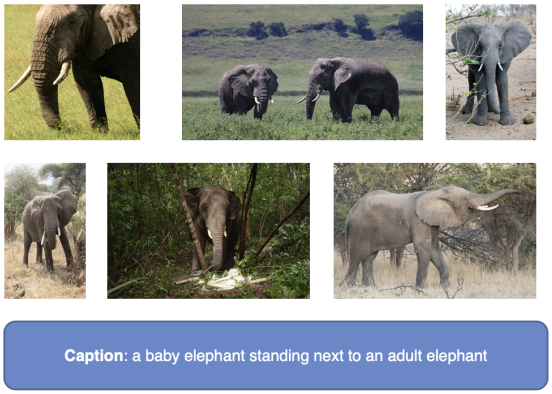

Fig. 4: Example of a caption generated by the proposed method with the same main objects.

isting methods except for ROGUE-1. For the proposed method, concept generation did not contribute except for WEEM4TS, the abstractive summarization evaluation that applies calculating word similarity.

We found from the experiment that the concept generalized in Sub-graph concept generation component can keep the main specific context if it can represent the whole image collection context shown in Fig. 4.

5. Conclusion and Discussion

We proposed a captioning method to describe the content of an image collection as a text caption. Inspired by text summarization in generating the summary, the proposed method shows improvement in the abstractive of the summarized image collection caption by finding the related words using graph theory and word communities. We further showed improvement in the caption generation based on summarization metric evaluation by CIDErBtw focusing on the similarity between sentences and WEEM4TS focusing on evaluating abstractive summary.

In the future, we will compare the proposed method with other summarization models, both the supervised model such as T5 [21] and the unsupervised model, for example, SUPERT [20]. We also notice that the proposed method is limited in fair comparison with other state-of-the-art text summarizations. We hence consider evaluating our methods by humans in the future. Moreover, establishing an evaluation method for the task is also needed. By reviewing abstractive summarization and evaluation methods, extending evaluation vocabulary in the evaluation process may be a suitable evaluation process such as BERTScore [23] and ROUGE-G [26].

Acknowledgement

Parts of this work were supported by JSPS KAKENHI (21H03519).

References

[1] Zhong, Yiwu, et al. “Comprehensive image captioning via scene graph decomposition.” In Proc. 16th European Conf. on Computer Vision. Vol. 14, pp. 211–229, 2020.

[2] Sankar, Aravind, Xinyang Zhang, and Kevin Chen-Chuan Chang. “Motif-based convolutional neural network on graphs.” arXiv preprint, arXiv:1711.05697, 2017.

[3] Han, Xiaotian, et al. “Image scene graph generation (SGG) benchmark.” arXiv preprint, arXiv:2107.12604, 2021.

[4] Milewski, Victor, Marie-Francine Moens, and Iacer Calixto. “Are scene graphs good enough to improve image captioning?.” arXiv preprint, arXiv:2009.12313, 2020.

[5] Anderson, Peter, et al. “Bottom-up and top-down attention for image captioning and visual question answering.” In Proc. 2018 IEEE Conf. on Computer Vision and Pattern Recognition. pp. 6077–6086, 2018.

[6] Zhang, Ji, et al. “Graphical contrastive losses for scene graph parsing.” In Proc. 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. pp. 11535–11543, 2019.

[7] Speer, Robyn, Joshua Chin, and Catherine Havasi. “Conceptnet 5.5: An open multilingual graph of general knowledge.” In Proc. 31st AAAI Conf. on Artificial Intelligence. pp. 4444–4451, 2017.

[8] Krishna, Ranjay, et al. “Visual Genome: Connecting language and vision using crowdsourced dense image annotations.” Int. J. of Computer Vision, 123.1. pp. 32–73, 2017.

[9] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. “GloVe: Global vectors for word representation.” In Proc. 2014 Conf. on Empirical Methods in Natural Language Processing. pp. 1532–1543, 2014.

[10] Wang, Jiuniu, et al. “Compare and reweight: Distinctive image captioning using similar images sets.” In Proc. 16th European Conf. on Computer Vision. Vol.1, pp. 370–386, 2020.

[11] Hailu, Tulu Tilahun, Junqing Yu, and Tessfu Geteye Fantaye. “A framework for word embedding based automatic text summarization and evaluation.” Information 11.2, pp. 78–100, 2020.

[12] Papineni, Kishore, et al. “BLEU: A method for automatic evaluation of machine translation.” In Proc. 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318, 2002.

[13] Lin, Chin-Yew. “ROUGE: A package for automatic evaluation of summaries.” In Proc. ACL2004 Workshop on Text Summarization Branches Out. pp. 74–81, 2004.

[14] Girshick, Ross. “Fast R-CNN.” In Proc. 16th IEEE Int. Conf. on Computer Vision. pp. 1440–1448, 2015.

[15] He, Kaiming, et al. “Deep residual learning for image recognition.” In Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition. pp. 770–778, 2016.

[16] Simonyan, Karen, and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition.” arXiv preprint, arXiv:1409.1556 2014.

[17] Lin, Tsung-Yi, et al. “Microsoft COCO: Common Objects in Context.” In Proc. 13th European Conf. on computer vision, Vol.5 pp. 740–755, 2014.

[18] Faghri, Fartash, et al. “VSE++: Improving visual-semantic embeddings with hard negatives.” arXiv preprint, arXiv:1707.05612, 2017.

[19] Karpathy, Andrej, and Li Fei-Fei. “Deep visual-semantic alignments for generating image descriptions.” In Proc. 2015 IEEE Conf. on Computer Vision and Pattern Recognition. pp. 3128–3137, 2015.

[20] Gao, Yang, Wei Zhao, and Steffen Eger. “SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization.” arXiv preprint, arXiv:2005.03724, 2020.

[21] Raffel, Colin, et al. “Exploring the limits of transfer learning with a unified text-to-text transformer.” arXiv preprint, arXiv:1910.10683, 2019.

[22] Gupta, Som, and Sanjai Kumar Gupta. “Abstractive summarization: An overview of the state of the art.” Expert Systems with Applications 121. pp.49–65, 2019.

[23] Zhang, Tianyi, et al. “BERTScore: Evaluating text generation with BERT.” arXiv preprint, arXiv:1904.09675, 2019.

[24] Alrasheed, Hend. “Word synonym relationships for text analysis: A graph-based approach.” PLOS ONE, 16.7, e0255127, 2021.

[25] Kingma, Diederik P., and Jimmy Ba. “Adam: A method for stochastic optimization.” arXiv preprint, arXiv:1412.6980, 2014.

[26] ShafieiBavani, Elaheh, et al. “A graph-theoretic summary evaluation for ROUGE.” In Proc. 2018 Conf. on Empirical Methods in Natural Language Processing. pp. 762–767, 2018.