Towards Captioning an Image Collection from a Combined Scene Graph Representation Approach

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

[1] Nagoya University, Nagoya, Aichi, Japan
[2] Kyoto University, Kyoto, Kyoto, Japan
[3] RIKEN, Seika, Kyoto, Japan
phueaksrii@cs.i.nagoya-u.ac.jp

Abstract. Most content summarization models from the field of natural language processing summarize the textual contents of a collection of documents or paragraphs. In contrast, summarizing the visual contents of a collection of images has not been researched to this extent. In this paper, we present a framework for summarizing the visual contents of an image collection. The key idea is to collect the scene graphs for all images in the image collection, create a combined representation, and then generate a visually summarizing caption using a scene-graph captioning model. Note that this aims to summarize common contents across all images in a single caption rather than describing each image individually. After aggregating all the scene graphs of an image collection into a single scene graph, we normalize it by using an additional concept generalization component. This component selects the common concept in each sub-graph with ConceptNet based on word embedding techniques. Lastly, we refine the captioning results by replacing a specific noun phrase with a common concept from the concept generalization component to improve the captioning results. We construct a dataset for this task based on the MS-COCO dataset using techniques from image classification and image-caption retrieval. An evaluation of the proposed method on this dataset shows promising performance.

Keywords: Multiple-image summarization · Image captioning · Scene graph captioning

1 Introduction

With an increasing number of images on the Web and on Social Media, it has become a challenge to describe and understand these images. Describing a collection of images with a short description is often easier to grasp overall contexts than describing them individually. For a single image, image captioning is a popular task that generates an image description in the form of a sentence. However, it cannot be easily adjusted to describe multiple images simultaneously. Image collection summarization [22,24,32] is a challenging new task which aims to generate a shared caption for all images in an image collection. However, existing

Fig. 1. Example of image summarization compared with image collection captioning: The left side shows the existing approach, in which the description is limited to words or noun phrases. The right side shows the proposed approach, in which an image collection is described in a single and refined sentence.

approaches are limited to summarizing an image collection only in the form of concept words or tag words. A recent work [24] presents a method for summarizing the texture, style, and material of similar objects in an image collection. Other works [22,32] summarize an image collection as a set of keywords. To be more informative in describing an image collection in semantic contexts, we propose image collection captioning to describe an image collection in the form of a single sentence. Fig. 1 shows the proposed task compared to the existing image summarization task.

Our approach aims to understand an image collection based on scene graph representations generated from the images in the collection. A scene graph is a popular means of describing a region-based image context by detecting objects in the image and their relationships. It has also been leveraged as a bottom-up mechanism for image captioning tasks [12]. A scene graph is a structured list of triplets consisting of a subject, a predicate, and an object. In the proposed method, with multiple scene graphs from images in the collection, we combine all scene graphs into a single scene graph representation. We then estimate all nodes and relations to find the most prominent combined context and generate a summarized scene graph for the captioning model to generate a phrase.

The challenge of the image collection captioning task is to generate a caption that can simultaneously describe all images in the image collection. Inspired by abstractive text summarization [7], we propose two components to generalize specific words to more general word choices with ConceptNet [23]. The first component is *Sub-Graph Concept Generation* that processes all image scene graphs in response to expanding the word concept following the ConceptNet. Incorporating the idea of word communities [1], we find the representative words in each community to be the word choices for the captioning. For example, when constructing a word community for "bird" and "bear," we find a representative word "animal" after expanding the concept. The second component is *Sentence Refinement*, which integrates the result of sub-graph concept generation and the captioning result to rebuild the caption, focusing on the noun phrase. When the captioning result is, e.g., "a polar bear standing in the snow near the water," the refined result becomes "animal standing in the snow near the water." Here, it replaces the phrase "a polar bear" with a more general word "animal," which is the knowledge gathered from the *Sub-Graph Concept Generation* component.

Due to this task being novel, there is no existing dataset available. Thus, we build an image collection dataset from the popular image caption dataset, MS-COCO [17]. However, the number of captions in the MS-COCO dataset is restricted to five captions per image. Correspondingly with the dataset limitation and the idea of summarizing scene graphs, we build a scene graph captioning model trained by a single image. Then, we transfer the model to our framework. To evaluate the proposed method, we compare it with text summarization methods, which are most similar to our work. These methods are evaluated on several automatic evaluation metrics designed to evaluate text generation and text summarization tasks. This work consists of the following:

- First, we propose a new challenging task, image collection captioning, which aims to describe an image collection in the form of a single sentence.
- Second, we propose a baseline for the novel task based on a combined scene graph captioning approach. We build a combined scene graph representing all images in the image collection and then generate a caption based on it.
- Last, we construct a dataset for this task based on the MS-COCO dataset by incorporating image classification and image-caption retrieval tasks.

2 Related Work

Image Captioning. Image captioning [12] is an image-to-text translation task that aims to describe the scene, location, objects, and interactions in an image in the form of a sentence. BEiT-3 [27] is the current state-of-the-art in vision-language task with multi-way transformers. Other methods [19,33] introduce a scene graph into their captioning models to improve the performance. In our work, we build upon this approach by extending this idea to scene graphs generated from multiple images and using their combined representation.

Multiple Image Summarization. Multiple image summarization was introduced recently, which aims to generate common keywords for an image collection. Samani et al. [22] propose a method to find the semantic concept of an image collection. They generalize word concepts in a specific domain and aim to find semantic similarities between them. Zhang et al. [32] present a method to generate a visual summary and a textual topic of an image collection by mapping and discovering the textual topics and corresponding images. Trieu et al. [24] propose extending the transformer-based architecture of a single image captioning model to generate a caption for an image collection, which aims to describe the texture, style, and material of the image collection in the form of a noun phrase. Due to the dataset limitation of this task, they also introduce a dataset construction method. By gathering images from Web pages, they collect 2.1 million image collections containing at least five images each. In our work, we propose a scene graph captioning model with a combined scene graph that can describe the interaction between the objects in the form of a sentence, which is more informative than implementing transformer-based methods.

Scene Graph Generation. Scene graph generation [9] is a method of describing object contexts in an image. Many scene graph generation methods start by finding object regions using Fast R-CNN [6] as an object detector. They then find the relationship between objects in both local and global contexts. Neural Motifs [29] is a model constructed from a stacked Motif Network (MotifNet), which is strongly predictive of relation labels on the Visual Genome Dataset and further evaluated on the MS-COCO Dataset. Their method includes three main predicting stages bounding regions, labels for regions, and relationships. RelDN [30] is a recent novel scene graph generation method focusing on improving the accuracy of relationship classification, raising entity confusion and loss over predicate classes. In our work, we make use of the Neural Motif network to detect the relationship between objects in an image using ResNet [11] as an object detector.

Text Summarization. Text summarization is a text-to-text generation task that aims to generate a short description from multiple documents. Text summarization tasks can be divided into two main paradigms: extractive summarization and abstractive summarization. Extractive summarization aims to identify the salient information that appears in documents. T5 [21] is a strong baseline of the supervised summarization model, which is pre-trained by the Wikipedia dataset[1]. Moreover, it is also introduced using unsupervised learning techniques. SUPERT [5] is an unsupervised multiple-document summarization model that evaluates the sentences in documents and selects one of them to be a topic. However, it is restricted to the document content. Meanwhile, abstractive summarization aims to rewrite the summarized sentence from sentences in the document by finding the semantic representation or generating a common word from a word corpus. XL-Sum [10] introduces the BBC news dataset and a multi-lingual abstractive summarization method that fine-tunes with the T5 model with their dataset. Our work is inspired by abstractive text summarization, as we aim to summarize an image collection into a generalized caption.

3 Proposed Method: Image Collection Captioning

Our proposed method starts with a collection of images. First, we generate a scene graph for each image. Next, all of the sub-graphs in the collection are combined. Word communities are also constructed from the sub-graphs to find common words. Lastly, we generate a caption from the summarized graph and refine the caption with the common words.

Following this idea, the proposed method consists of five components which are shown in Fig. 2 and discussed in detail in the following sections. The first one is *Scene Graph Generation* which extracts image features and generates a scene graph for each image (sub-graph). Next, all the sub-graphs are parallelly passed into two components: *Multiple Scene Graph Processing* to merge and

[1] https://www.tensorflow.org/datasets/catalog/wikipedia/ (accessed Sept. 9, 2022)

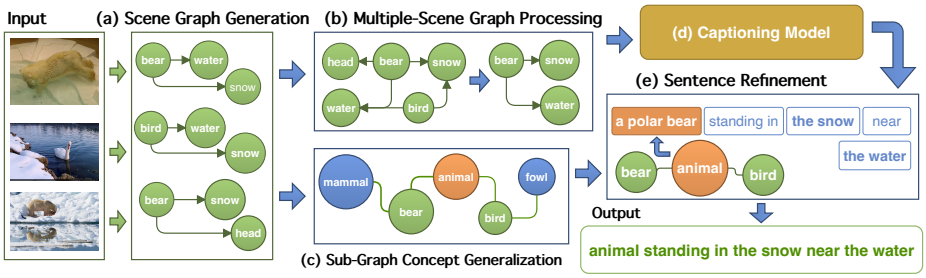

Fig. 2. Overview of the proposed method, consisting of five components: (a) *Scene Graph Generation* extracts a scene graph for each image. (b) *Multiple-Scene Graph Processing* combines all scene graphs and finds a representative graph. (c) *Sub-Graph Concept Generalization* finds word communities from all scene graphs and generates a common word. (d) *Captioning Model* generates the initial caption for the representative graph. (e) *Sentence Refinement* rephrases the caption with the common words.

select part of the combined graph, and *Sub-Graph Concept Generalization* to find general concept words through the word communities. Then, the *Captioning Model* generates a sentence based on the representative scene 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

Following the existing work on image captioning model leveraging scene graphs, we use the current state-of-the-art scene graph generation method with ResNet101 [11] + Neural Motif [29] as a scene graph parser in the proposed method. This model is retrained on the Visual Genome dataset [15], a popular practice for scene graph captioning. A recent image captioning work [2] shows that manually cleaning up duplicate labels of the Visual Genome dataset from 2,500/1,000/500 to 1,600/400/20 of objects/attributes/relations can improve image captioning performance. We also follow this idea. The result of scene graph generation is represented in a directed graph, which includes subjects, predicates, and objects.

3.2 Multiple-Scene Graph Processing

We build the multiple-scene graph processing module as a feature selection of all sub-graphs from the *Scene Graph Generation Component*. All the sub-graphs are merged into a single directed graph, as shown in Eq. 1, in which G is the merged graph and g_i is a directed graph represented as a set of triplets. In the merging process, we count the occurrence of each feature and the number of edges to be the weight in the selection step.

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

Next, we select the top-n nodes and the top-m relations from the sub-graphs (with n=36 and m=100 used in the following experiments, which are feature numbers of our captioning model). In preliminary experiments, we found that implementing betweenness centrality, which refers to the summarization of the fraction of the shortest path in finding the center nodes, is the most efficient method compared with other aspects, which is represented as:

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

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

3.3 Sub-Graph Concept Generalization

Next, we discuss how to generalize specific contexts within the sub-graphs. For example, given an image collection of animals in which each image contains "bear" or "bird," our idea is to generalize these words as a common word: "animal." We build the *Sub-Graph Concept Generalization Component* to find a common concept. A popular text-based semantic network named ConceptNet [23] is employed to extend the word relations of specific words and then select a general word to represent them.

Inspired by text analysis based on word synonym relationships, we build a community of words and find the representative word in each word community. The process is shown in Fig. 3. First, all the object words that appear in the sub-graph are lemmatized. We then incorporate ConceptNet to expand the word relationships of each node based on *synonym* and *isA* relations. In a preliminary experiment, we found that finding a concept word from more different numbers of expandable relations results in generating a more general word. We hence limit the maximum number of each relation to ten.

After expanding the concept, a word graph community is generated by joining all the expanded sub-graphs, and non-degree nodes are dropped. To estimate the representativeness of the node to be the common concept of each sub-graph, we encode all nodes using GloVe word embedding [20] and then calculate the similarity between each node as the distance. When calculating the weight, we select a word by calculating the highest node degree using cosine similarity as a weight to find each sub-graph word concept. We determine the representative in each word community by considering the average shortest distance node over all nodes. We implement the improved closeness centrality, which can estimate the graph with many connections [28] as:

$$C(u) = \frac{n-1}{N-1} \frac{n-1}{\sum_{v=1}^{n-1} d(u,v)}, \tag{3}$$

where $C(u)$ is the closeness centrality of node u, n is the number of all reachable nodes, N is the number of nodes in the graph, $d(u,v)$ is the distance between nodes u and v.

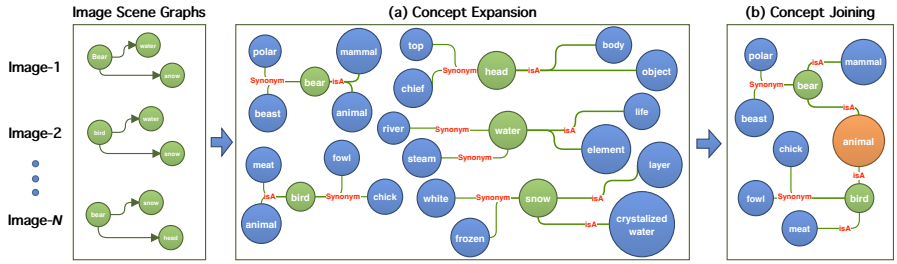

Fig. 3. Example of constructing word communities. (a) *Concept Expansion* expands word concepts of each word in sub-graphs by incorporating with ConceptNet [23]. (b) *Concept Joining* joins all the same words together and selects the central word node as the representative.

3.4 Captioning Model

The captioning model consists of a Graph Convolutional Network (GCN) [19] and the Attention-based LSTM model [2]. We build the GCN to process the triplet of *subject*, *predicate*, and *object* features. Each feature is extracted from the scene graph generation process, which consists of 36 subject features, 36 object features, and 100 predicate features. Each feature dimension is 1,024, which is the feature size of the *Scene Graph Generation* output. Next, we build the attention-based LSTM model following the top-down LSTM captioning with two layers of attention-based LSTM, both layers with a size of 512.

3.5 Sentence Refinement

To improve the caption, we modify the beam search of sentence generation to generalize the caption, mainly focusing on general noun words of the caption result. First, word tokens are extracted from the caption and labeled with NLTK POS tagging [18]. Next, a noun phrase is found and labeled with its object component. Finally, the object component of each noun phrase is mapped with the *Sub-Graph Concept Generation Component* to replace the word with the representative of the word community in the sub-graph concept. In the following experiment, we select a beam size of five for generating the final caption.

4 Dataset Construction

The proposed method aims to generate a caption for an image collection, i.e., a sentence describing multiple images. Due to this task being novel, there is no existing dataset for this. The MS-COCO dataset [17] is a popular image captioning dataset which is closest to this task. However, it is typically used only for the single-image captioning task in which each image is captioned with one or more sentences.

In our work, we build upon the MS-COCO dataset by estimating the semantic contents of images and captions and use this to augment the dataset towards

image collection captioning. The dataset contains numerous similar images, and the annotated labels are not distinct [26]. It is thus straightforward to use it to generate image collections. To construct a dataset for the proposed task, we implement and compare two approaches based on image classification and image-caption retrieval to estimate the semantic contents of the 5K testing.

4.1 Image Classification Approach

This approach refers to the common concepts gathered from scene graphs. We classify image classes using an image classification model and use that knowledge to collect similar images to construct image collections. In our experiment, we use ResNet101 [11] pre-trained with the ImageNet dataset [3]. First, the top-five classes of each image are predicted. Then, the intersection of classes between the images is found. The prediction scores of each class in each image collection are ranked, and the top-five prediction scores are selected, thus limiting the number of images for each collection. Each of the 880 classes forms the ImageNet concept classes, in which each class contains at least 5 images. The ground-truth captions for the image collection are the concatenated set of all captions of its composite images. Thus, we end up with 15 and 25 sentences from the original description, which are used for evaluation.

4.2 Image-Caption Retrieval Approach

This approach considers both semantic image contents and semantics of the captions annotated to each image. In the following experiment, VSE++ [4] can query the top K images in the embedding space by estimating their visual-semantic embedding. We generate 5K collections for our testing set and limit the query number to five, which results in each image collection in our testing set containing six images. The ground-truth captions for the image collection are the concatenated set of all captions of its composite images. Thus, we end up with 30 sentences from the original description, which are used for the evaluation.

5 Evaluation

5.1 Captioning Model Pre-Training Strategy

Due to the limits of the dataset, we first train and evaluate our captioning model for single images using the MS-COCO dataset [17]. Afterwards, the single image captioning model is integrated with the proposed image collection captioning framework, as discussed in Sect. 3. In our training phase, we follow the Karpathy split [13] in which the sizes of the training image set is 118K, the validation image set is 5K, and the testing image set is 5K images. We implement a learning rate decay of 0.8 for every eight epochs, the initial learning rate of 0.0008, dropout 0.5, employing Adam optimization [14], cross-entropy loss, and multi-label margin loss. The best checkpoint of a single caption with CIDEr [25] evaluation is used as the captioning model in the image collection captioning framework.

Fig. 4. Examples of the proposed method on image collections built with the image classification approach. The results show that the proposed method can extract important features from the image collections and generate a general caption describing the most occurred image contents in an image collection. Images with a border indicate that they fit the image collection caption, while those without indicate outliers.

5.2 Evaluation Metrics

Various evaluation metrics are introduced in the image captioning and text summarization field. Due to our method aiming to perform image captioning over summarization of multiple images, we use ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L) [16] as text summarization evaluation metrics. We further implement CIDErBtw [26], a similarity evaluation between sentences based on the CIDEr metric. However, these evaluation metrics are limited in evaluating abstract contexts. Thus, we use BERTScore [31], which is a text generation metric based on calculating word similarity score, and WEEM4TS [8], which is a metric introduced to evaluate abstractive summarization by evaluating the similarity between the summary and the ground-truth using word embedding.

5.3 Results

We evaluate the proposed method on 5K images from each testing dataset created by the two dataset construction approaches. A caption is generated for each image collection and is compared with the ground-truth caption set of each image collection using automatic evaluation metrics by averaging the scores. We evaluate the proposed method by comparing it to extractive and abstractive text summarization models. However, we realize that automatic evaluation is limited in abstractive summarization. To make the evaluation clear, we ablate two methods, *with* and *without* the *Sub-Graph Concept Generalization (CG)* component, and compare their results.

Image Classification Dataset We first show results on the dataset constructed with the image classification approach, which consists of 880 image

Table 1. Evaluation of the results of image collections built with the image classification approach compared to SUPERT [5], T5 [21], and XL-Sum [10].

Method		R-1 ↑	R-2 ↑	R-L ↑	BERTScore ↑	CIDErBtw ↑	WEEM4TS ↑
SUPERT [5]		0.3116	0.0823	0.2848	0.5889	0.4095	0.0746
T5 [21]		0.2938	0.0728	0.2601	0.5710	0.3002	0.0573
XL-Sum [10]		0.1873	0.0284	0.1632	0.4630	0.1004	0.0616
Proposed	w/o CG	**0.3254**	**0.0912**	**0.2958**	**0.5999**	**0.5308**	0.1088
	w/ CG	0.3077	0.0823	0.2777	0.5895	0.4755	**0.1132**

Fig. 5. Examples of the proposed method for image collections built with the image-caption retrieval approach. The results show that the proposed method can extract important features from the image collection and normalize the word concepts in captioning results. Images with a border indicate that they fit the image collection caption, while those without indicate outliers.

collections, in Fig. 4. It shows that the proposed method detects the most frequently occurring contents, ignoring less appearing contents. Then, it generates a general caption to describe most image contents in an image collection. The evaluation results are shown in Table 1, which shows that the proposed method achieved the best result on overall automatic evaluations.

Image-Caption Retrieval Dataset We next show results on the dataset constructed with the image-caption retrieval approach in Fig. 5. It shows that the prediction relates to the most frequent content in the image collection, and unrelated contents are ignored. They also keep the main specific content if it can describe the image collection. Table 2 shows the evaluation results, which show that the proposed method beats text summarization methods in this novel task.

The experiments above show a novel image collection captioning baseline compared with text summarization methods. However, we also found a limitation of automatic evaluation metrics when the captioning results are refined. The proposed method without refining the captioning results, achieved better evaluation scores in ROUGE, BERTScore, and CIDerBtw, due to the nature of these metrics focusing on text similarity. However, WEEM4TS is a novel metric for abstractive summarization evaluation and shows a promising direction for the image collection captioning task being more suitable for our task of generalizing across images.

Table 2. Evaluation of the results of image collections built with the image-caption retrieval approach compared to SUPERT [5], T5 [21], and XL-Sum [10].

Method		R-1 ↑	R-2 ↑	R-L ↑	BERTScore ↑	CIDErBtw ↑	WEEM4TS ↑
SUPERT [5]		0.3756	0.1105	0.3231	0.6166	0.7016	0.1083
T5 [21]		0.3441	0.1037	0.3025	0.6057	0.5524	0.1031
XL-Sum [10]		0.2148	0.0367	0.1833	0.4678	0.1023	0.0860
Proposed	w/o CG	**0.3782**	**0.1265**	**0.3409**	**0.6270**	**0.7955**	0.1062
	w/ CG	0.3517	0.1105	0.3140	0.6093	0.7156	**0.1096**

6 Conclusion

We introduced a challenging new task which aims to produce a single fitting caption for a collection of images. The key idea was to generate a scene graph for each image in the collection, combine them to generate a generalized combined representation, and then generate a caption. The proposed method showed potential for transferring a single-image captioning model to image collection captioning. Inspired by text summarization methods generating a summary, the proposed method improves the abstractiveness of the summarized image collection caption by finding generalized words using graph theory and word communities. We additionally introduced the prospect of using an augmented version of the MS-COCO dataset, a popular image captioning dataset, in the image collection captioning task. The results are promising and pioneering steps toward captioning an image collection with a shared description. In the future, we plan to work also on a more challenging dataset and also improve the captioning model focusing on estimating the overall semantic context of an image collection incorporating external knowledge. Our project can be found at https://www.cs.is.i.nagoya-u.ac.jp/opensource/nu-icc/

Acknowledgements Parts of this work were supported by JSPS Grant-in-aid for Scientific Research (21H03519) and a joint research project with National Institute of Informatics.

References

1. Alrasheed, H.: Word synonym relationships for text analysis: A graph-based approach. PloS One **16**(7), e0255127 (2021)
2. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: 2018 IEEE Conf. Comput. Vis. Pattern Recognit. pp. 6077–6086 (2018)
3. Deng, J., et al.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Comput. Society Conf. Comput. Vis. Pattern Recognit. pp. 248–255 (2009)
4. Faghri, F., et al.: VSE++: Improving visual-semantic embeddings with hard negatives. In: 29th Brit. Mach. Vis. Conf. (2018)
5. Gao, Y., et al.: SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In: 58th Annu. Meeting of the Assoc. for Computat. Linguist. pp. 1347–1354 (2020)
6. Girshick, R.: Fast R-CNN. In: 16th IEEE Int. Conf. Comput. Vis. pp. 1440–1448 (2015)
7. Gupta, S., et al.: Abstractive summarization: An overview of the state of the art. Expert Syst. Appl. **121**, 49–65 (2019)
8. Hailu, T.T., et al.: A framework for word embedding based automatic text summarization and evaluation. Information **11**(2), 78–100 (2020)
9. Han, X., et al.: Image scene graph generation (SGG) benchmark. Comput. Res. Reposit. arXiv preprint arXiv:2107.12604 (2021)
10. Hasan, T., et al.: XL-Sum: Large-scale multilingual abstractive summarization for 44 languages. In: Findings Assoc. Comput. Linguist.: ACL-IJCNLP 2021. pp. 4693–4703 (2021)

11. He, K., et al.: Deep residual learning for image recognition. In: 2016 IEEE Conf. Comput. Vis. Pattern Recognit. pp. 770–778 (2016)
12. Hossain, M.Z., et al.: A comprehensive survey of deep learning for image captioning. ACM Comput. Survey **51**(6), 1–36 (2019)
13. Karpathy, A., et al.: Deep visual-semantic alignments for generating image descriptions. In: 2015 IEEE Conf. Comput. Vis. Pattern Recognit. pp. 3128–3137 (2015)
14. Kingma, D.P., et al.: Adam: A method for stochastic optimization. In: 3rd Int. Conf. Learn. Representat. (2014)
15. Krishna, R., et al.: Visual Genome: Connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. **123**(1), 32–73 (2017)
16. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: ACL-04 Workshop on Text Summarization Branches Out. pp. 74–81 (2004)
17. Lin, T.Y., et al.: Microsoft COCO: Common objects in context. In: 13th Euro. Conf. Comput. Vis. vol. 5, pp. 740–755 (2014)
18. Loper, E., et al.: NLTK: The natural language toolkit. In: 42nd Annu. Meeting Assoc. for Comput. Linguist. vol. 1, pp. 63–70 (2002)
19. Milewski, V., et al.: Are scene graphs good enough to improve image captioning? In: Joint Conf. 59th Annu. Meeting Assoc. Comput. Linguist. and 11th Int. Joint Conf. Nat. Lang. Process. (2020)
20. Pennington, J., et al.: GloVe: Global vectors for word representation. In: 2014 Conf. Empir. Methods Nat. Lang. Process. pp. 1532–1543 (2014)
21. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(140), 1–67 (2020)
22. Samani, Z.R., et al.: A knowledge-based semantic approach for image collection summarization. Multimed. Tools Appl. **76**(9), 11917–11939 (2017)
23. Speer, R., et al.: ConceptNet 5.5: An open multilingual graph of general knowledge. In: 31st AAAI Conf. Artif. Intell. pp. 4444–4451 (2017)
24. Trieu, N., et al.: Multi-image summarization: Textual summary from a set of cohesive images. Comput. Res. Reposit. arXiv preprint arXiv:2006.08686 (2020)
25. Vedantam, R., et al.: CIDEr: Consensus-based image description evaluation. In: 2015 IEEE Conf. Comput. Vis. Pattern Recognit. pp. 4566–4575 (2015)
26. Wang, J., et al.: Compare and reweight: Distinctive image captioning using similar images sets. In: 16th Euro. Conf. Comput. Vis. vol. 1, pp. 370–386 (2020)
27. Wang, W., et al.: Image as a foreign language: BEiT pretraining for all vision and vision-language tasks. Compt. Res. Reposit. arXiv preprint arXiv:2208.10442 (2022)
28. Wasserman, S., et al.: Social Network Analysis: Methods and Applications, vol. 8. Cambridge University Press, Cambridge, UK (1994)
29. Zellers, R., et al.: Neural motifs: Scene graph parsing with global context. In: 2018 IEEE Conf. Comput. Vis. Pattern Recognit. pp. 5831–5840 (2018)
30. Zhang, J., et al.: Graphical contrastive losses for scene graph parsing. In: 2019 IEEE Conf. Comput. Vis. Pattern Recognit. pp. 11535–11543 (2019)
31. Zhang, T., et al.: BERTScore: Evaluating text generation with BERT. In: 9th Int. Conf. Learn. Representat. (2020)
32. Zhang, W., et al.: Joint optimisation convex-negative matrix factorisation for multimodal image collection summarisation based on images and tags. IET Comput. Vis. **13**(2), 125–130 (2019)
33. Zhong, Y., et al.: Comprehensive image captioning via scene graph decomposition. In: 16th Euro. Conf. Comput. Vis. vol. 14, pp. 211–229 (2020)