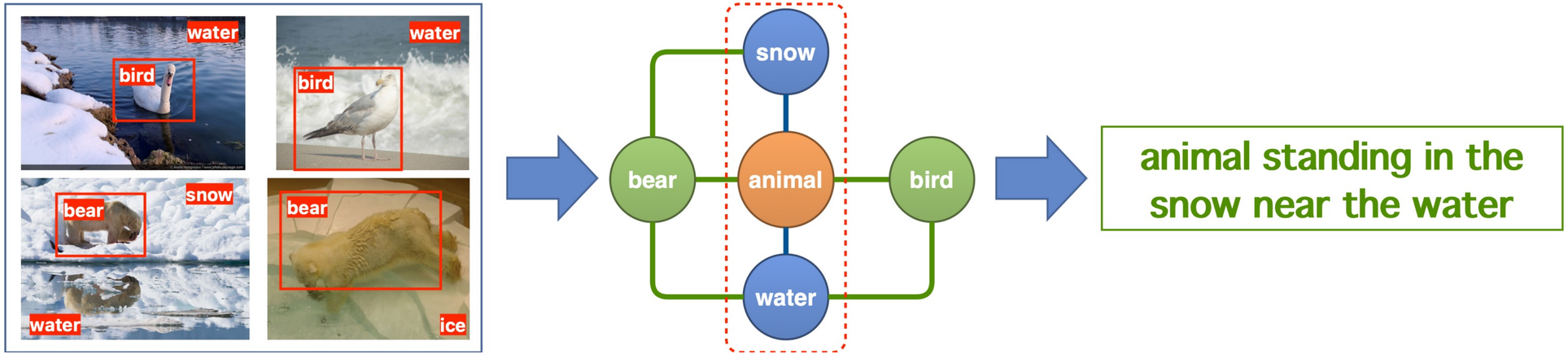
Toward Captioning an Image Collection from a Combined Scene Graph Representation Approach

Phueaksri Itthisak[1], Marc A. Kastner[2], Yasutomo Kawanishi[3, 1],
Takahiro Komamizu[1], Ichiro Ide[1].
(1) Nagoya University, Japan, (2) Kyoto University, Japan, (3) RIKEN, Japan

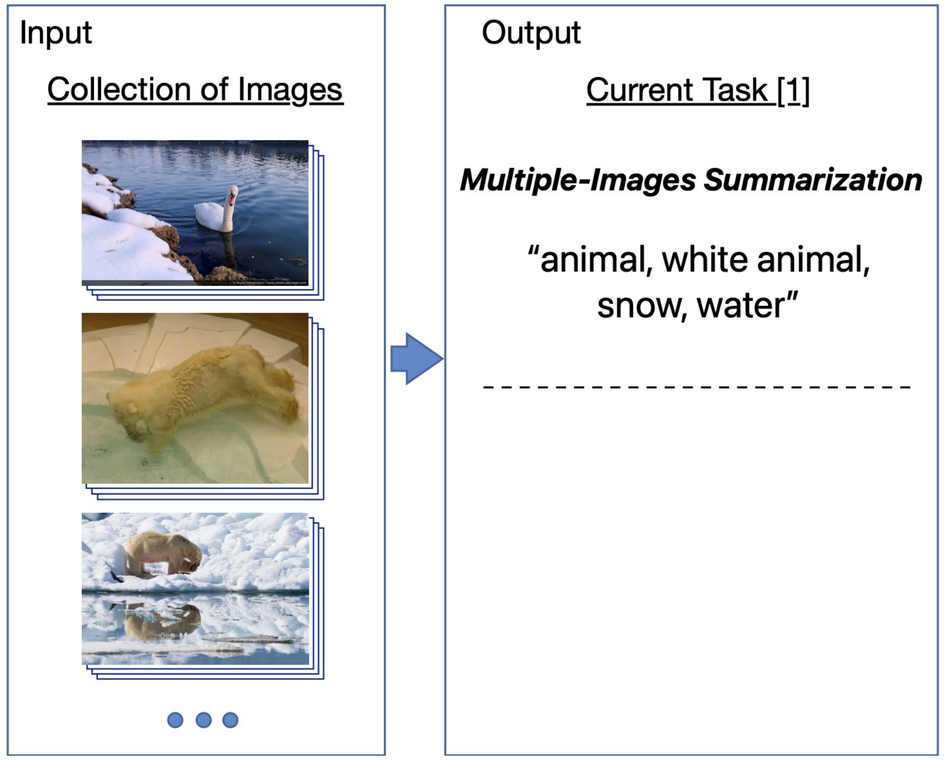
29th International Conference on Multimedia Modeling January 9, 2023

Motivation: Image Collection Captioning

- With the increasing number of images and image collections
  - Describing a collection of images with a short description
  - Grasping the common context from an image collection



Describing an Image and Multiple-Images



- **Current tasks**
  - ***Image captioning***

    Describe an image with a single sentence
  - ***Multiple-images summarization***

    Describe multiple images with concept words or noun phrases in specific domains
- **Proposed task**
  - ***Image collection captioning***

    Describe the commonly occurring contexts of an image collection

[1] Samani, Z.R., et al.: A knowledge-based semantic approach for image collection summarization. Multimed. Tools Appl. 76(9), 11917–11939 (2017)

Difficulties and Solutions

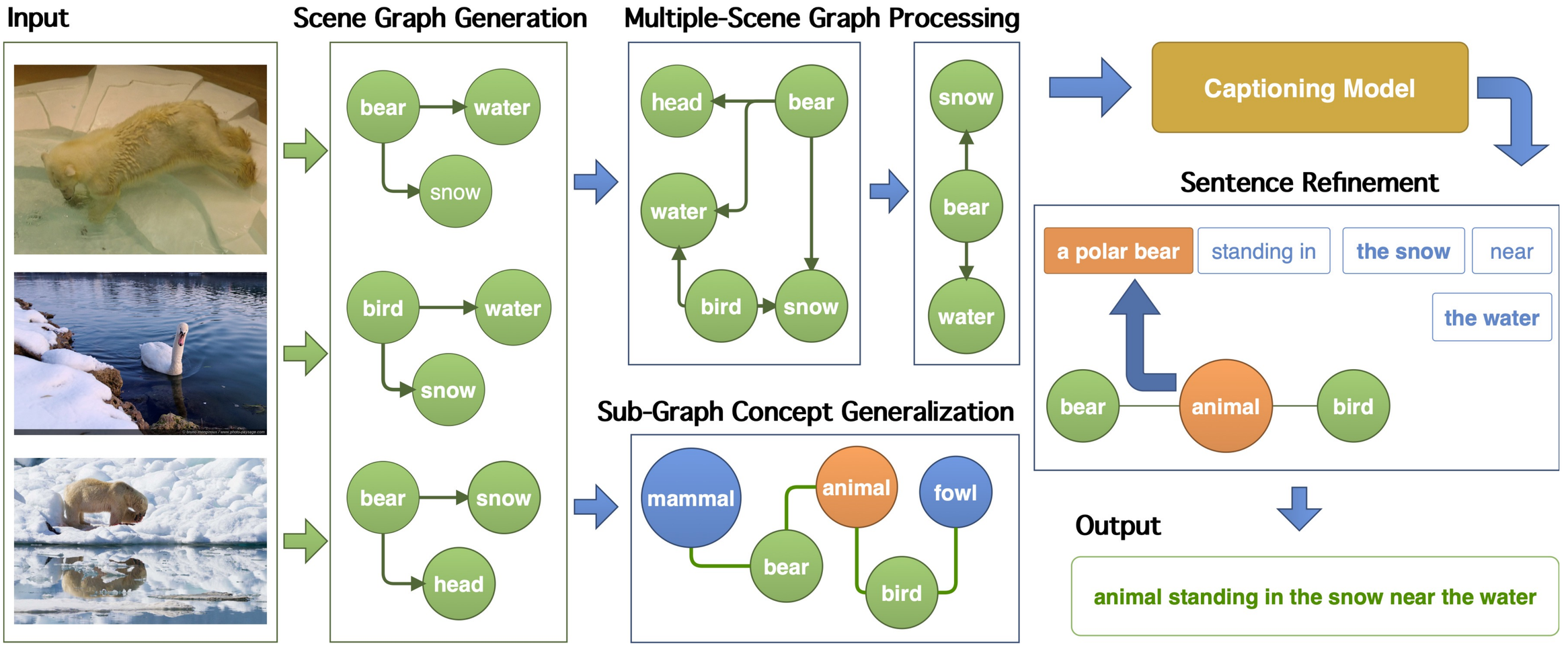
- **Difficulties**
  - How to estimate the most prominent context of an image collection
  - How to generalize specific concepts in each image of an image collection
- **Solutions**
  - ***Multiple-Scene Graph Processing*** that merges image scene graphs to generate a representative scene graph
  - ***Sub-Graph Concept Generalization*** that finds common concept words by refining the final caption incorporating external knowledge

Contributions

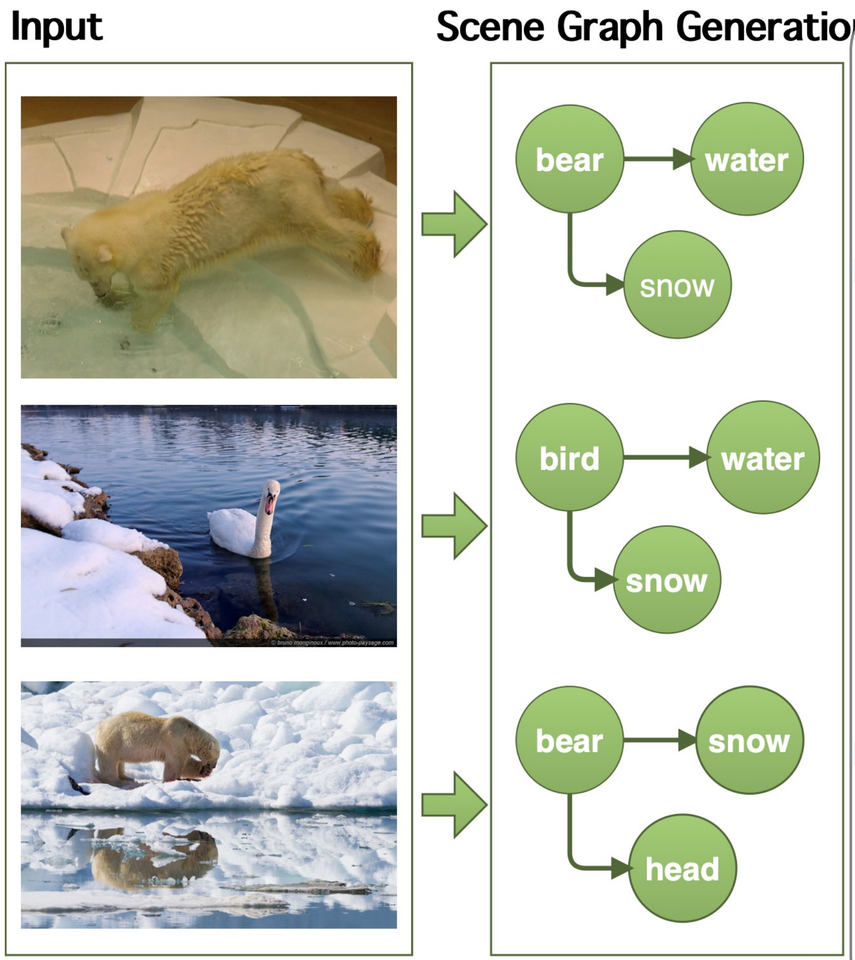
- Build a framework to generate a single caption for an image collection
- Propose a scene graph processing method and a concept generalization method to build a combined scene graph representation and then generate a caption based on it
- Construct a dataset by augmenting the MS-COCO [1] dataset

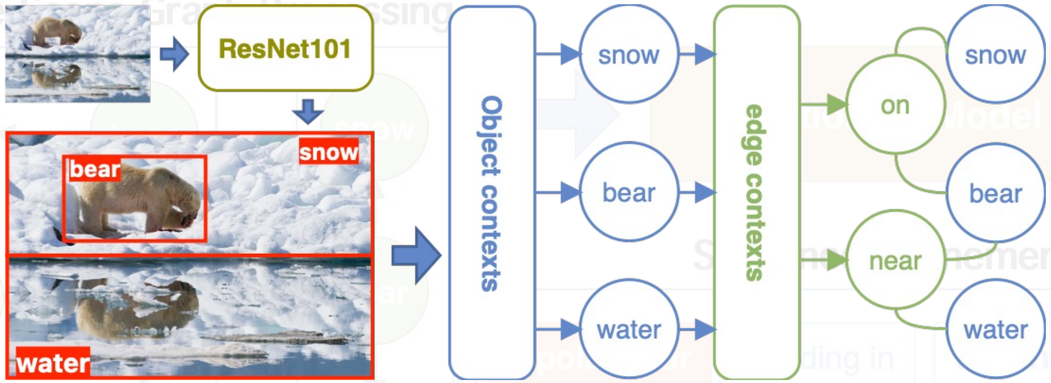
[1] Lin, T.Y., et al.: Microsoft COCO: Common objects in context. In: 13th Euro. Conf. Comput. Vis. vol. 5, pp. 740–755 (2014)

Framework


Input
Scene Graph Generation
Multiple-Scene Graph Processing
bear
water
snow
bird
water
snow
bear
snow
head
Sub-Graph Concept Generalization
mammal
animal
fowl
bear
bird
head
bear
water
bird
snow
snow
bear
water
Captioning Model
Sentence Refinement
a polar bear
standing in
the snow
near
the water
bear
animal
bird
Output
animal standing in the snow near the water

Framework: Scene Graph Generation



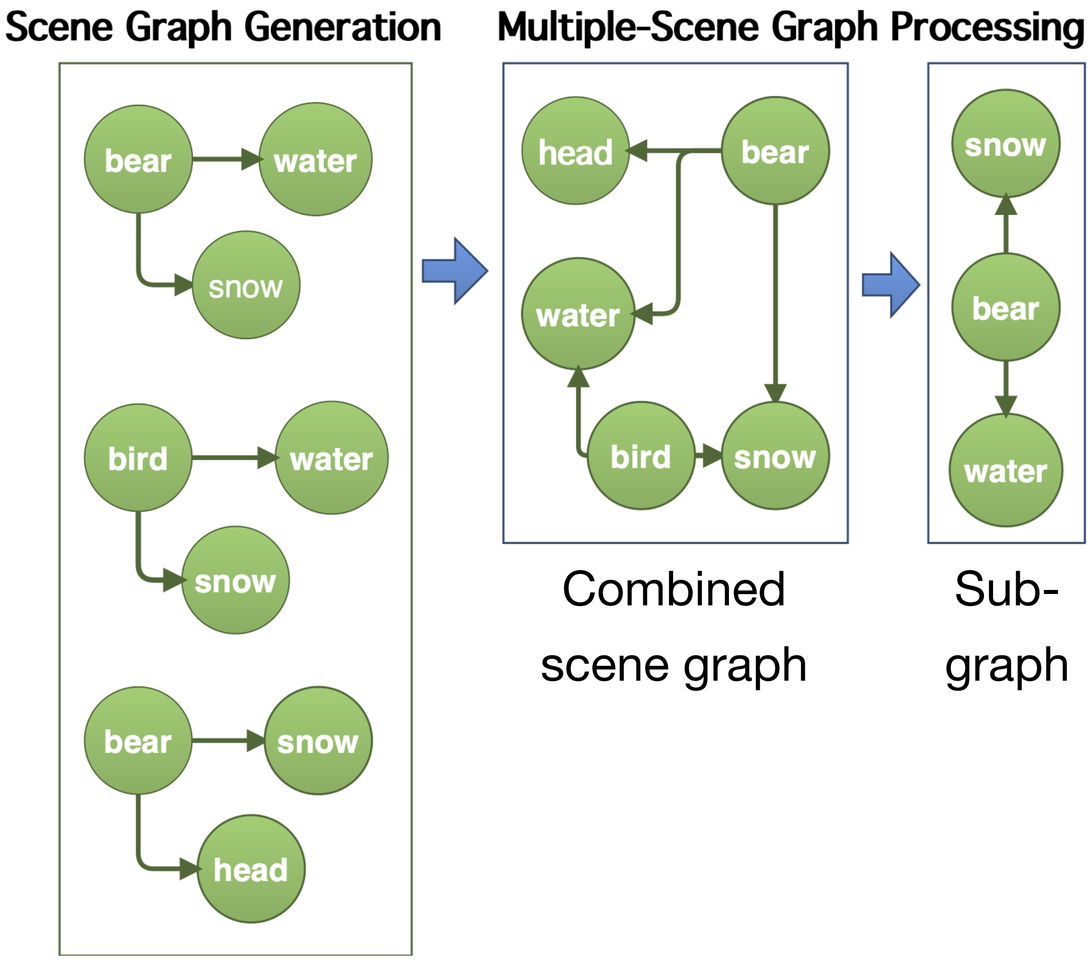


- Model Architecture
  - Scene graph parser: Neural Motif [1]
  - Backbone: ResNet101 [2]
- Pre-trained on Visual Genome dataset [3]

[1] Zellers, R., et al.: Neural motifs: Scene graph parsing with global context. In: 2018 IEEE Conf. Comput. Vis. Pattern Recognit. pp. 5831–5840 (2018)
[2] He, K., et al.: Deep residual learning for image recognition. In: 2016 IEEE Conf. Comput. Vis. Pattern Recognit. pp. 770–778 (2016)
[3] Krishna, R., et al.: Visual Genome: Connecting language and vision using crowd- sourced dense image annotations. Int. J. Comput. Vis. 123(1), 32–73 (2017)

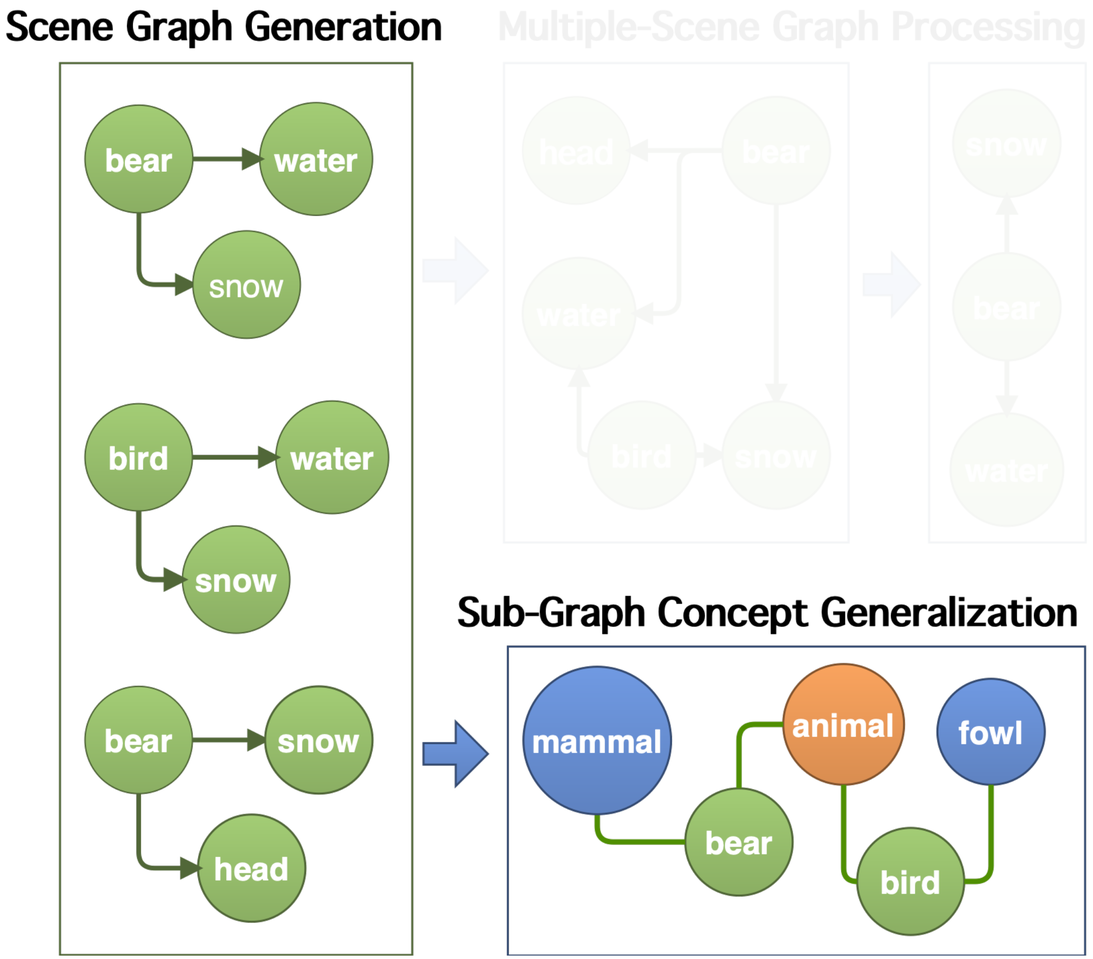
Framework: Multiple-Scene Graph Processing

- Merge image scene graphs into a directed graph
- Estimate the centrality of a combined scene graph
- Select nodes and relationships and represent them as a sub-graph



Framework: Sub-Graph Concept Generalization

- Build word communities to find the representative of the community
- Employ ConceptNet [1] to extend synonyms and related words
- Find the representative of each word community by estimating the centrality of the community



[1] Speer, R., et al.: ConceptNet 5.5: An open multilingual graph of general knowledge. In: 31st AAAI Conf. Artif. Intell. pp. 4444–4451 (2017)

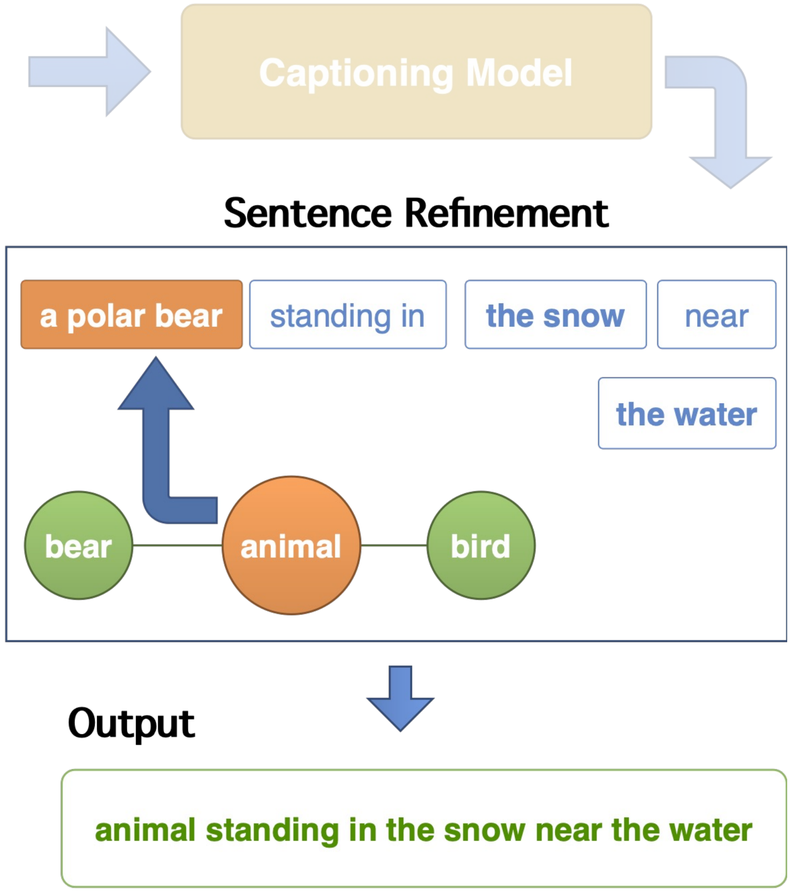
Framework: Captioning Model

- Graph Attention Network [1]
  - Graph Convolution Network
  - Attention-based LSTM
- Training
  - Train and validate with a single image on the MS-COCO dataset [2]
  - Transfer a single image captioning model to an image collection captioning framework

[1] 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)
[2] Lin, T.Y., et al.: Microsoft COCO: Common objects in context. In: 13th Euro. Conf. Comput. Vis. vol. 5, pp. 740–755 (2014)

Framework: Sentence Refinement

- Extract the caption into a set of words
- Label each word with NLTK POS tagging [1]
- Map a noun phrase with the representative word of word communities



[1] Loper, E., et al.: NLTK: The natural language toolkit. In: 42nd Annu. Meeting Assoc. for Comput. Linguist. vol. 1, pp. 63–70 (2002)

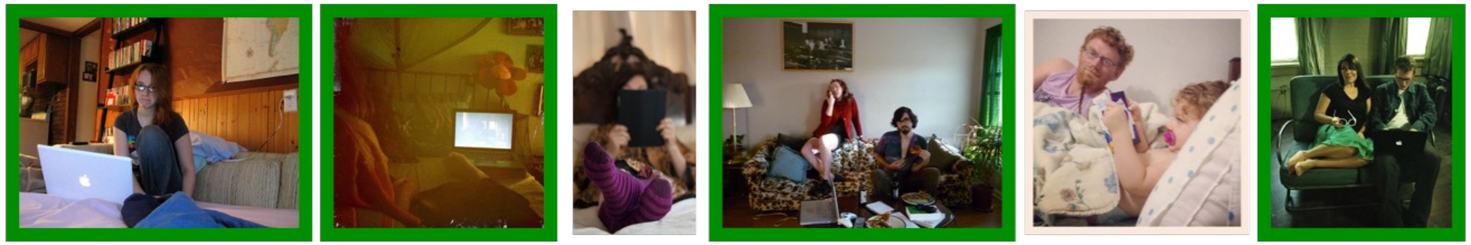
Experimental Dataset

- Build a dataset based on the MS-COCO [1] dataset
  - **Image-Text Retrieval Approach**
    - Considers the semantics of both image contents and captions by estimating visual semantics embedding
    - Implement VSE++ [2] to query the top-*k* images
    - 5,000 testing collections with 6 images/collection

[1] Lin, T.Y., et al.: Microsoft COCO: Common objects in context. In: 13th Euro. Conf. Comput. Vis. vol. 5, pp. 740–755 (2014)
[2] Faghri, F., et al.: VSE++: Improving visual-semantic embeddings with hard neg- atives. In: 29th Brit. Mach. Vis. Conf. (2018)

Results

animal standing in the snow near the water



person sitting on a couch with a laptop

Evaluation

- **Evaluation Metrics**
  - Summarization: ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L) [1], WEEM4TS [2], BERTScore [3]
  - Distinctiveness: CIDerBtw [4]
- **Comparison methods**
  - Text Summarization models: SUPERT [5], T5 [6], XL-Sum [7]
  - Summarize ground-truth captions of each collection into a single sentence

[1] Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: ACL-04 Workshop on Text Summarization Branches Out. pp. 74–81 (2004)
[2] Hailu, T.T., et al.: A framework for word embedding based automatic text sum- marization and evaluation. Information 11(2), 78–100 (2020)
[3] Zhang, T., et al.: BERTScore: Evaluating text generation with BERT. In: 9th Int. Conf. Learn. Representat. (2020)
[4] 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)
[5] Gao, Y., et al.: SUPERT: Towards new frontiers in unsupervised evaluation met- rics for multi-document summarization. In: 58th Annu. Meeting of the Assoc. for Computat. Linguist. pp. 1347–1354 (2020)
[6] 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)
[7] 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)

Evaluation Results

Metrics	SUPERT	T5	XL-Sum	Proposed (w/o CG)	Proposed (w/ CG)
ROUGE-1 (↑)	0.376	0.344	0.215	**0.378**	0.352
ROUGE-2 (↑)	0.111	0.104	0.037	**0.127**	0.111
ROUGE-L (↑)	0.323	0.303	0.183	**0.341**	0.314
WEEM4TS (↑)	0.108	0.103	0.086	0.106	**0.110**
BERTScore (↑)	0.617	0.606	0.468	**0.627**	0.609
CIDErBtw (↑)	0.702	0.552	0.102	**0.796**	0.716

*CG is Sub-Graph Concept Generalization

- **Summarization**
  - Proposed methods outperform text summarization methods
  - Proposed *CG* is not shown to be effective when evaluated by text-similarity-based metrics (*ROUGE-1/2/L and BERTScore*)
  - Proposed *CG* is shown to be effective when evaluated by word embedding-based metric (*WEEM4TS*)
- **Distinctiveness**
  - Proposed methods outperform text summarization methods

Conclusion

- **Summary**
  - Introduced a new challenging task of “**Image Collection Captioning**”
  - Introduced a framework to generate a shared caption for an image collection by scene graph and text summarization
  - Built an augmented version of the MS-COCO dataset for this task
- **Future work**
  - Improve the captioning model by estimating the overall semantic contexts of an image collection incorporating external knowledge
  - Work on a more challenging dataset by extending and augmenting from the existing dataset