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

Phueaksri Itthisak¹, Marc A. Kastner², Yasutomo Kawanishi³, ¹, Takahiro Komamizu¹, Ichiro Ide¹.

(1) Nagoya University, Japan, (2) Kyoto University, Japan, (3) RIKEN, Japan
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

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 \cite{1} dataset

Framework

Input

Scene Graph Generation

Multiple-Scene Graph Processing

Sub-Graph Concept Generalization

Captioning Model

Sentence Refinement

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 \[3\]

---


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

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

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

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

Results

animal standing in the snow near the water

group 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

## Evaluation Results

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

*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