Story Generation using Scene Graphs

Term Project *for*

AI60007: Graph Machine Learning Foundations and Applications

Group-7

Debjoy Saha (18EC35008)

Shubhesh Anand (18QE30002)

Divyanshu Sheth (18QE30008)

Objective

- To generate stories from sequence of images.

Story: I went to the party last week. the chef was preparing the food. He was very happy to see him. they had a great time. After the ceremony was over, everyone gathered together to talk about their plans .

Previous Work

- Diverse and Relevant Visual Storytelling with Scene Graph Embeddings

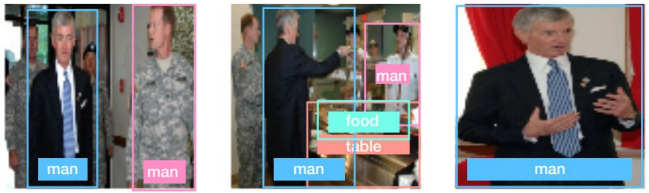
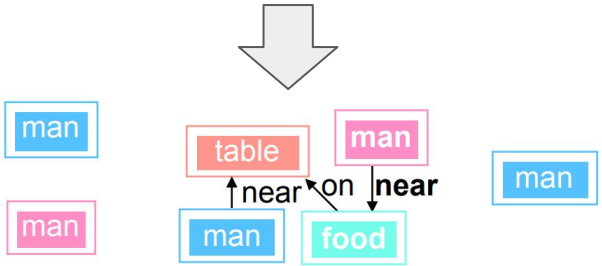
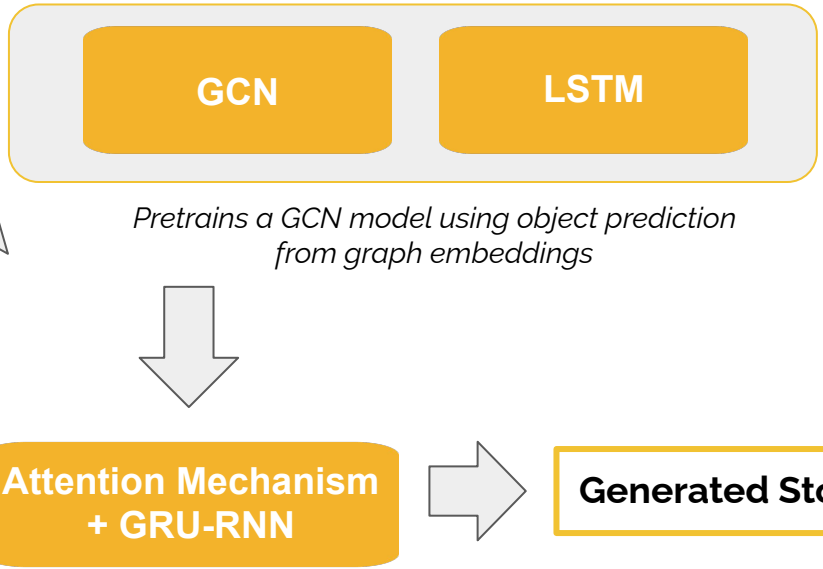


Image Sequence extracted from video



Scene Graphs generated and augmented



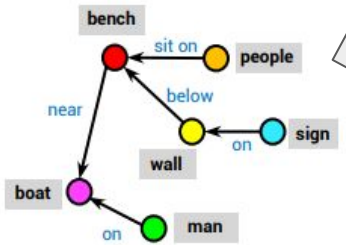
Pretrains a GCN model using object prediction from graph embeddings

Sequence Model for language generation

Related Work

- Sub-GC: Comprehensive Image Captioning via Scene Graph Decomposition

Image



Scene Graph generated

GCN

Graph Convolutional Network (trained)

Subgraph-1

Subgraph-2

Subgraph-3

Subgraphs extracted using Subgraph Proposal Network (sGPN)

For each subgraph

Attention Mechanism + LSTM

Sequence Model for language generation

Generated Caption

Sub-GC Details

Graph Convolution Network Training in Sub-GC

- Data Preparation
  - MotifNet is used to extract Scene-Graphs from the Object Detection outputs - Extracted scene-graphs contain nodes as objects and edges as relationship between object pairs.
  - Sub-graphs are extracted from the scene graph by using neighbor sampling.
  - Nodes and Edges are augmented with visual and text features.
- Sub-Graph Proposal Network
  - To identify meaningful sub-graphs that are likely to capture major scene components.
  - A Graph Convolutional Network (GCN) aggregates information from nearby nodes and edges using visual and text features projected into a common sub-space.
  - The final layer node embeddings are pooled for each sub-graph and fed into a scoring function which is used to propose important sub-graphs.

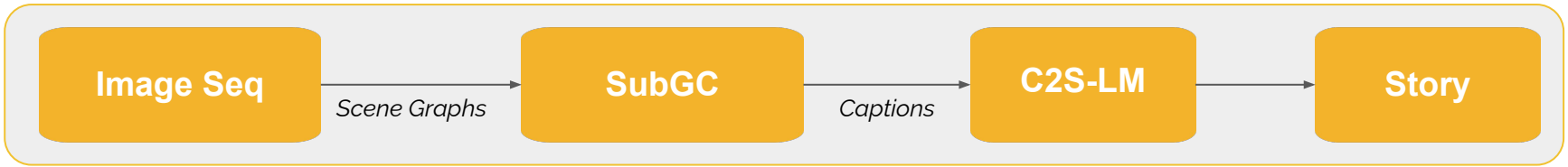
Sub-GC Details

Graph Convolution Network Training in Sub-GC (continued)

- Decoding Sentences from Sub-graphs
  - Two LSTM models - Attention LSTM and Language LSTM.
  - The Attention LSTM computes scores for all nodes in a particular subgraph considering the textual embeddings, the node embeddings and the pooled subgraph embeddings.
  - The Language LSTM takes as input the hidden state of the attention LSTM and the attention re-weighted sub-graph features to generate text.
- Training and Inference
  - Trained using two loss components - a binary cross-entropy Loss for the sub-graph proposal network, and a multi-way cross-entropy loss for the attention-based LSTM model (language modelling loss).
  - During inference, greedy non maximal suppression is used to remove redundant sub-graphs having high IoU.

Our Idea

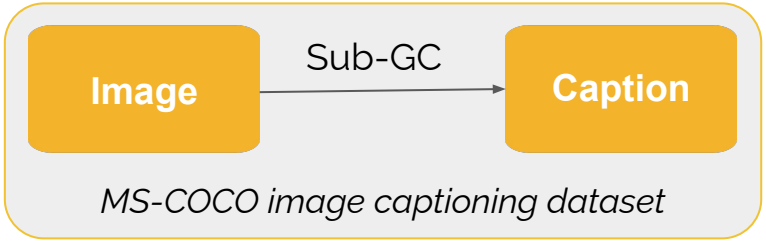
- Combine subgraph-based caption generation with a Seq2Seq language model (C2S-LM) to produce stories from image sequences indirectly.
  - **Simple** - Breaks down the task into smaller pieces that are trained independently (scene-graph to captions; captions to story).
  - **Flexible** - Doesn't require matched image sequence – story data.
  - **Cheap** - Less computation involved in training.

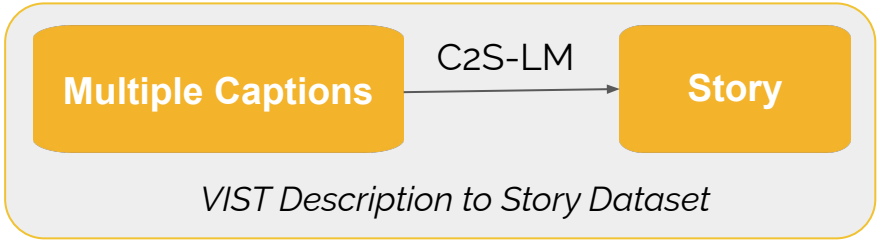


Brief Overview of our proposed method

Training

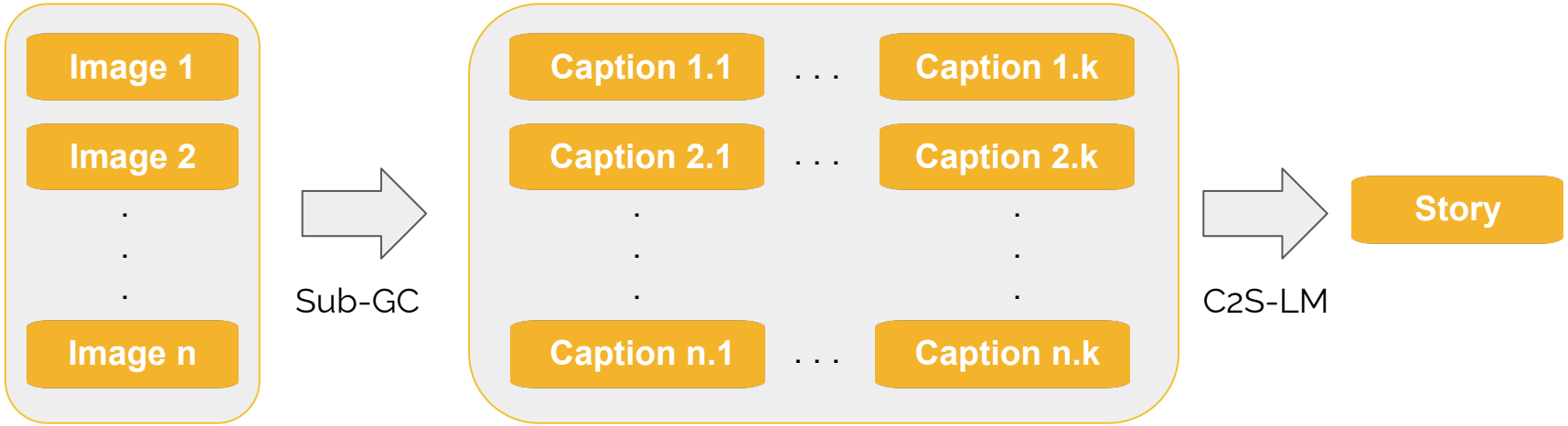
- The Subgraph-based image captioning model (Sub-GC) is trained using the MRNN split of the MS-COCO image captioning dataset.
  - Uses the training process as described in the Sub-GC slides.
- The Caption to Story Language Model (C2S-LM) is trained using the VIST dataset, which contain images extracted from Flickr.
  - Pretrained T5 model is finetuned using an appropriate prompt tuning methods.
  - Validation score and ROUGE are used for selecting best checkpoint.





Inference

- Availability of multiple captions for each image ensures that we can generate diverse set of stories considering different set of captions.
- The best set of captions is determined using the sGPN scores.



Results

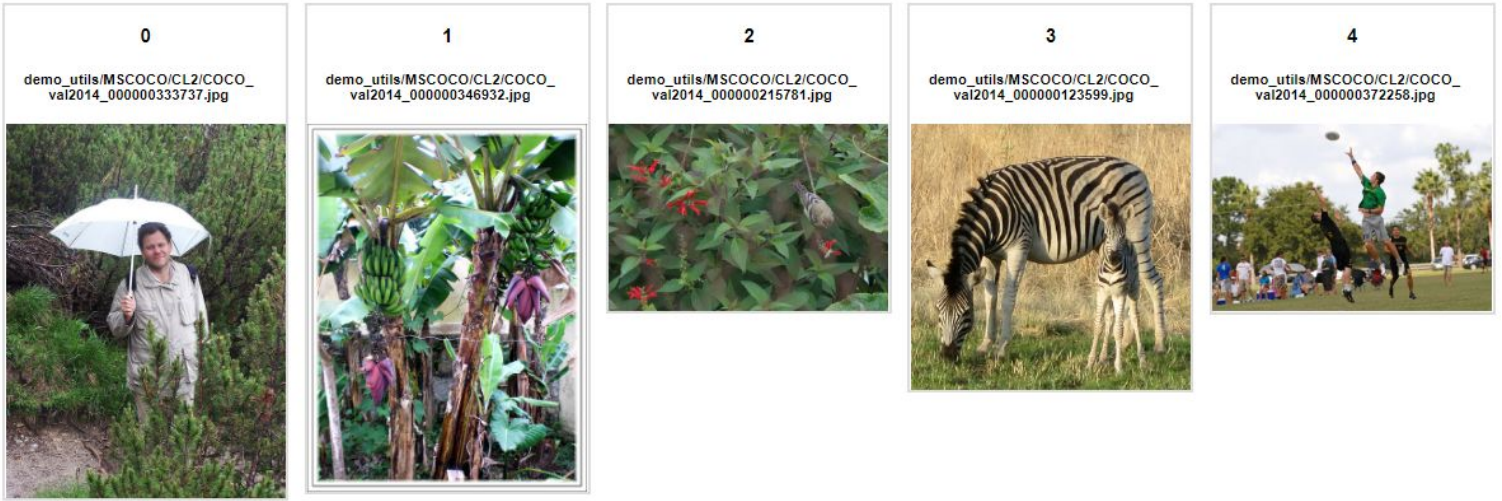
We perform subjective evaluations only for two cases.

1. Proper image sequences sampled from the VIST dataset.

Generated Story: The group of friends gathered for the dance. The speaker was very funny. The group danced for a while. Then they all gathered for a group photo.

Results (continued)

2. Seemingly random sequence of images sampled from MS-COCO



Generated Story: We went to the park to see the animals. We saw a lot of bananas. We saw a bird. We saw a zebra. We played frisbee..

Demo

Demo available at -

https://github.com/Debjoy10/Sub-GStory/blob/master/StoryGen/demo.ipynb

1. Clone repository
2. Download dependencies
3. Download pretrained models
4. Run ipynb

Conclusion & Further Improvements

1. Issue: Stories generated for the randomly sampled images are plain.
   Solution: The generation variety can be improved by employing training objectives which encourages more variety in generation.
2. Issue: Some information in image can disappear during caption generation.
   Solution: Instead of using captions for the story generation process, we can use the sub-graph embeddings directly. It is a costlier process, but will give us better, more diverse generations.

- In the future, we plan on incorporating these changes and hopefully improving the quality of the generated stories.

References

[1] Hong, Xudong, et al. "Diverse and Relevant Visual Storytelling with Scene Graph Embeddings." Proceedings of the 24th Conference on Computational Natural Language Learning. 2020.

[2] Zhong, Yiwu, et al. "Comprehensive image captioning via scene graph decomposition." European Conference on Computer Vision. Springer, Cham, 2020.

[3] Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." J. Mach. Learn. Res. 21.140 (2020): 1-67.