Storytelling from an Image Stream Using Scene Graphs

Ruize Wang,[1] Zhongyu Wei,[2,4*] Piji Li,[5] Qi Zhang,[3] Xuanjing Huang[3]

[1]Academy for Engineering and Technology, Fudan University, China

[2]School of Data Science, Fudan University, China

[3]School of Computer Science, Fudan University, China

[4]Research Institute of Intelligent and Complex Systems, Fudan University, China

[5]Tencent AI Lab, China

AAAI 2020

Outlines

- Introduction and Motivation
- Method
- Experiments and Analysis
- Conclusion

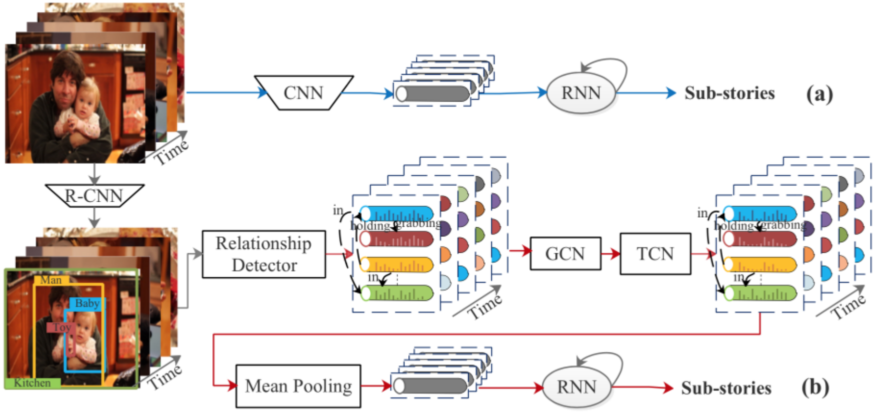
Introduction and Motivation

Introduction

- **What is *Visual Storytelling*?**
  - For most people, showing them images and asking them to compose a reasonable story about the images is not a difficult task, while it's still challenging for the machine.
  - This task can be formalized as *Visual Storytelling*, which aims at generating a story for an image stream.

Introduction

- **Previous work:**
  - Previous methods for visual storytelling employ encoder-decoder structure to translate images to sentences directly, with CNN-based models for representing images as the extracted high-level features and RNN-based models for text generation, and trained with MLE or RL.
- **Drawbacks:**
  - Not intuitive to represent all the visual information of the image with an abstract high-level feature, and this also hurts the interpretability and reasoning ability of the model.



Motivation

- **Recall:**
  - When humans telling stories for an image sequence: ① recognize the objects in each image ② reason about their visual relationships ③abstract the content into a scene ④ observe the images in order and reason the relationship among images.
- **Main Motivation:**
  - Translating each image into a graph-based semantic representation, i.e., scene graph, and reasoning the relationships on scene graphs at two levels, i.e., within-image and cross-images levels, would benefit representing and describing images.

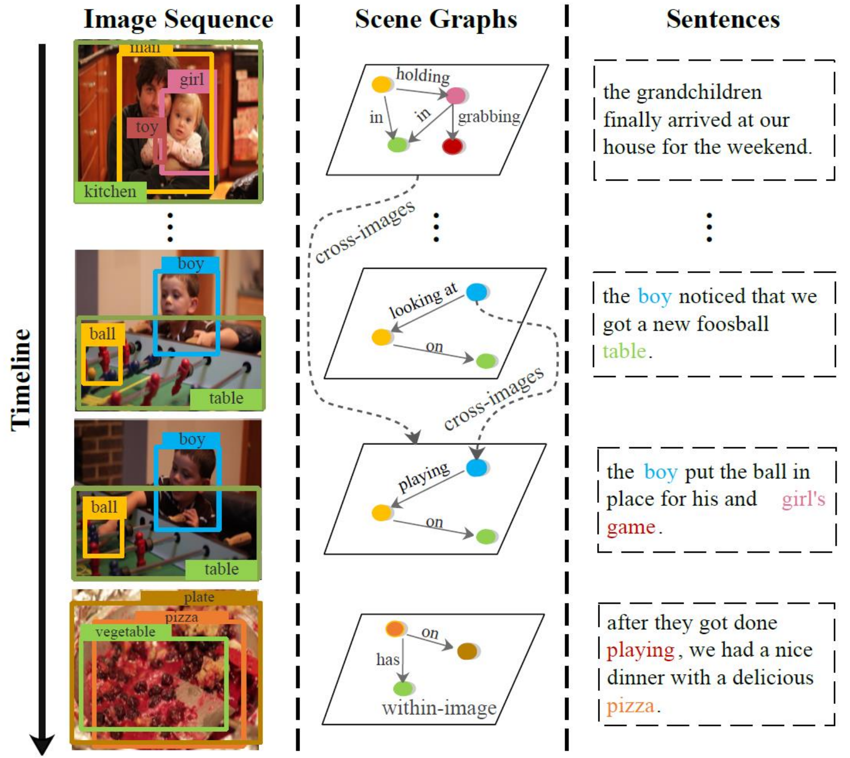
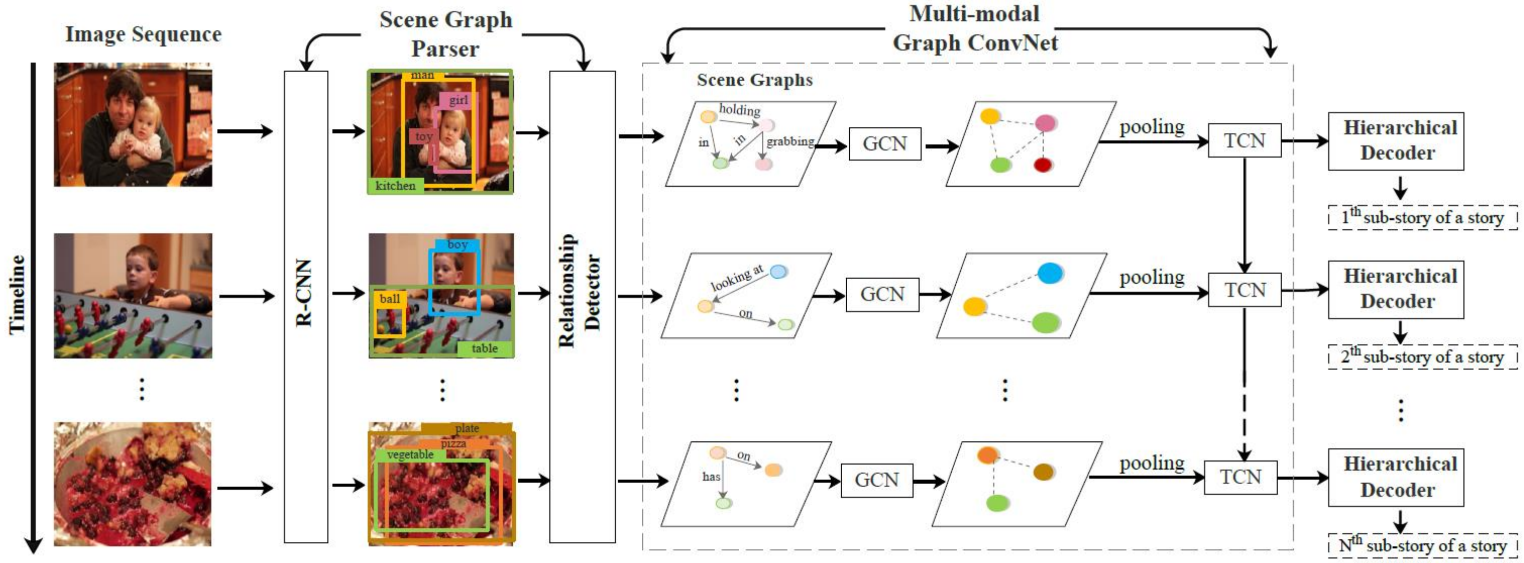


Figure 1: A scene graph based example for visual storytelling

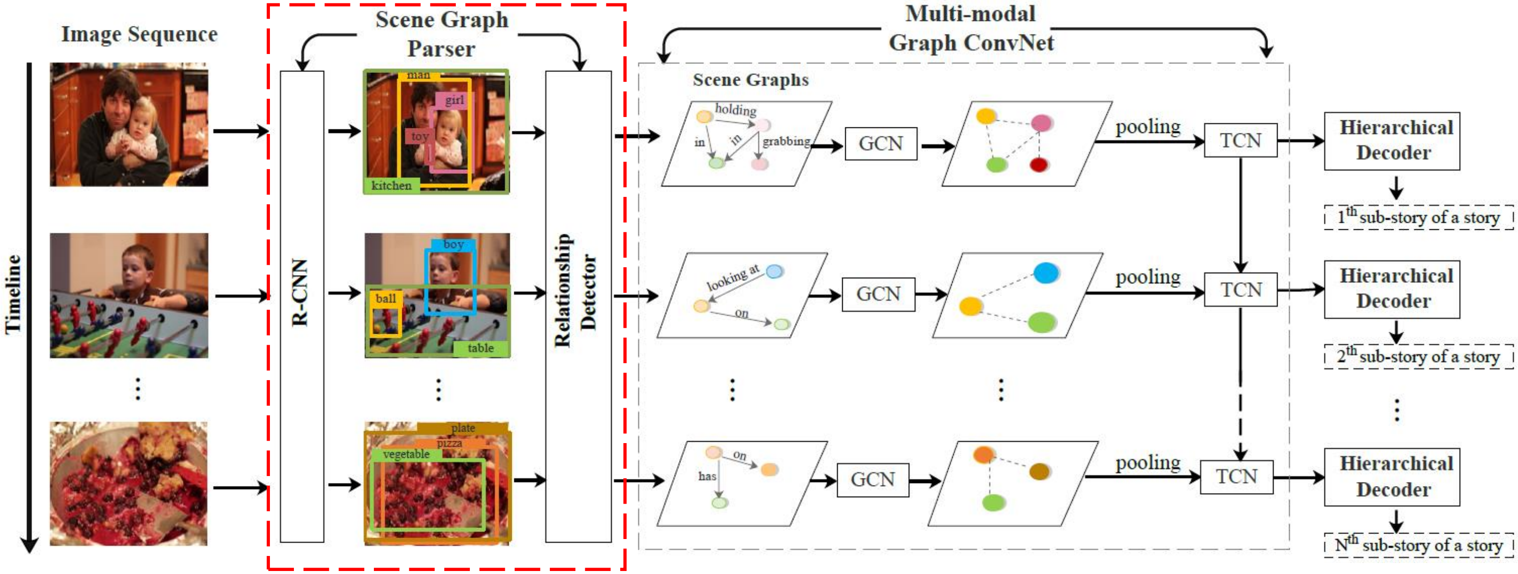
Method

Method - Overview



- Propose a novel *graph-based* architecture for modeling the two-level (within-image and cross-image) relationships through *Multi-modal Graph ConvNet* on scene graphs.
  - **Input:** an image sequence stream $I = \{I_1, \ldots, I_N\}$ ---> scene graphs $G = \{G_1, \ldots, G_N\}$
  - **Output:** a story $y = \{y_1, \ldots, y_N\}$, where $y_n = \{w_1, \ldots, w_T\}$

Method – Scene Graph Parser

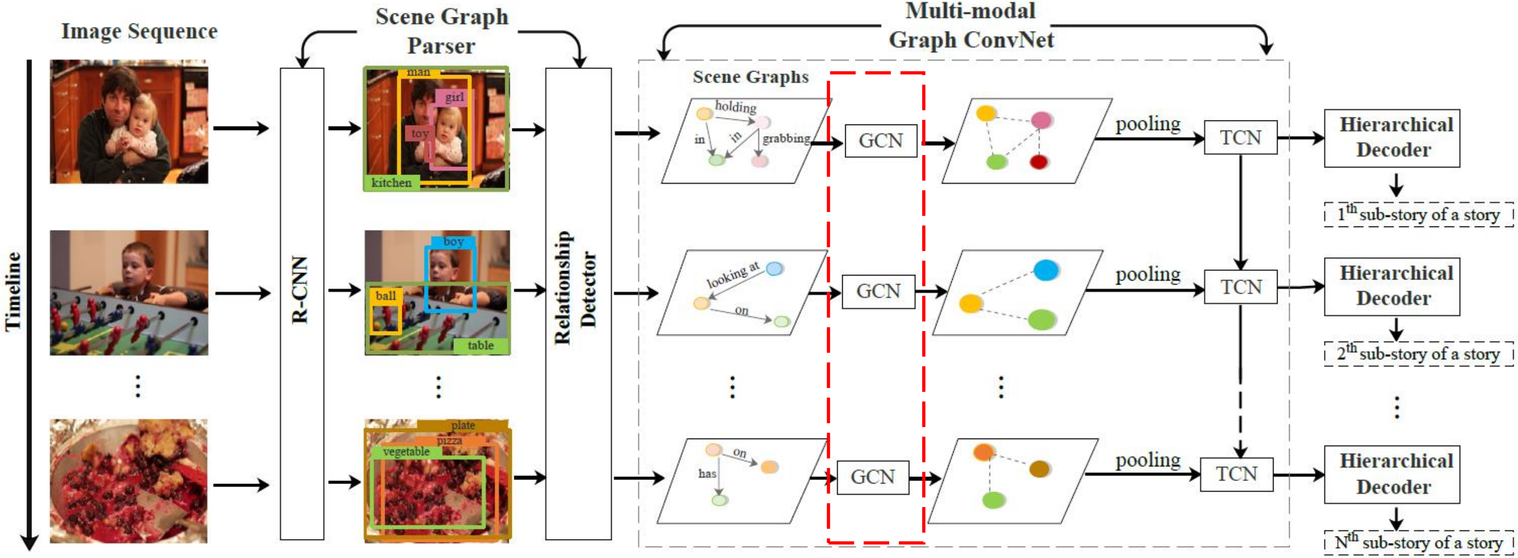


- Parse an image I_n to a scene graph $G_n=(V_n,E_n)$, where $V_n=\{v_{n,1},\ldots,v_{n,k}\}$ is a set of K detected objects with each image region and E_n is a set of directed edges denoting visual relationship between objects: <*subject-predicate-object*>, e.g., <*man-holding-boy*>.
- Object Detector: produce and classify objects using *Faster-RCNN*
- Relationship Detector: classify relationships between objects using *MOTIFS*

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In CVPR.

Method – Multi-modal Graph ConvNet



Within-image level: *Graph Convolution Network (GCN)*

- **Input:** region features and edge features $v_{n,i} \in \mathbb{R}^{D_v}, v_r \in \mathbb{R}^{D_r}$ **Output:** refined region features and edge features $v'_{n,i}$ v'_r
- Compute output features for all nodes and edges using three functions (*MLP*) g_s, g_p and g_o

Output edges vectors

$$v'_r = g_p(v_{n,i}, v_r, v_{n,j})$$

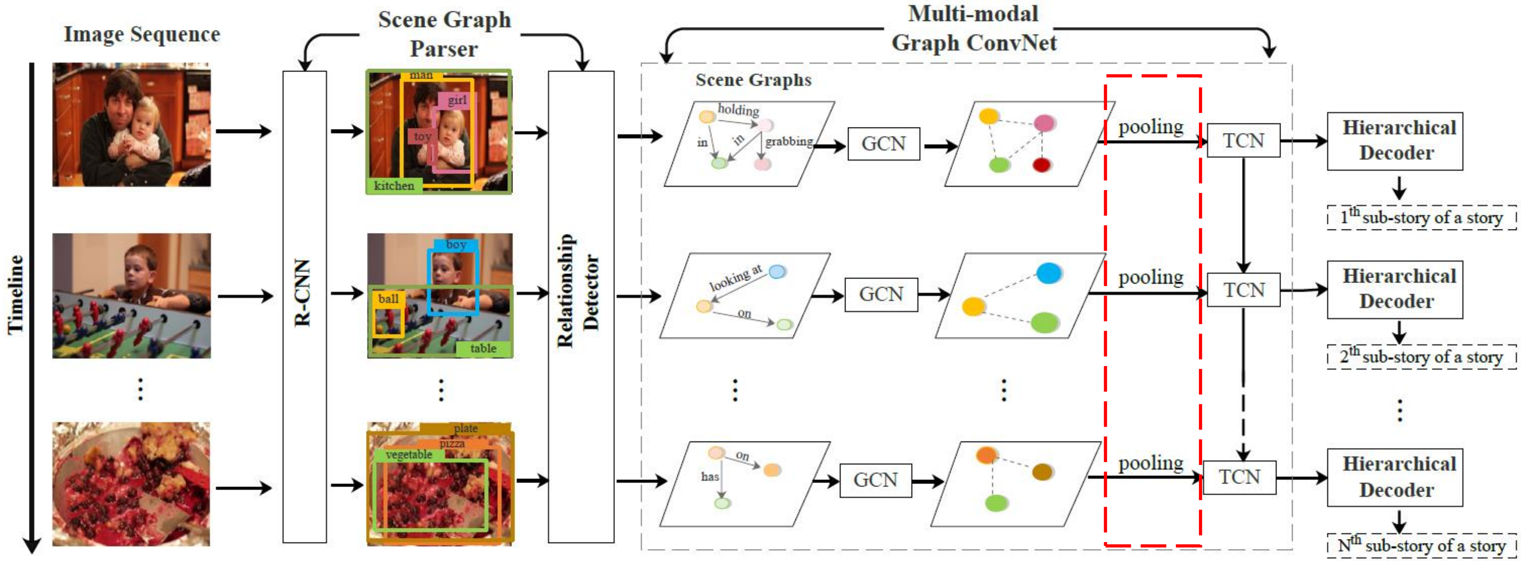
Output objects vectors

$$V^s_{n,i} = \{g_s(v_{n,i}, v_r, v_{n,j})\}$$
$$V^o_{n,i} = \{g_o(v_{n,j}, v_r, v_{n,i})\}$$

$$v'_{n,i} = h(V^s_{n,i} \cup V^o_{n,i})$$

h: average pooling

Method – Multi-modal Graph ConvNet

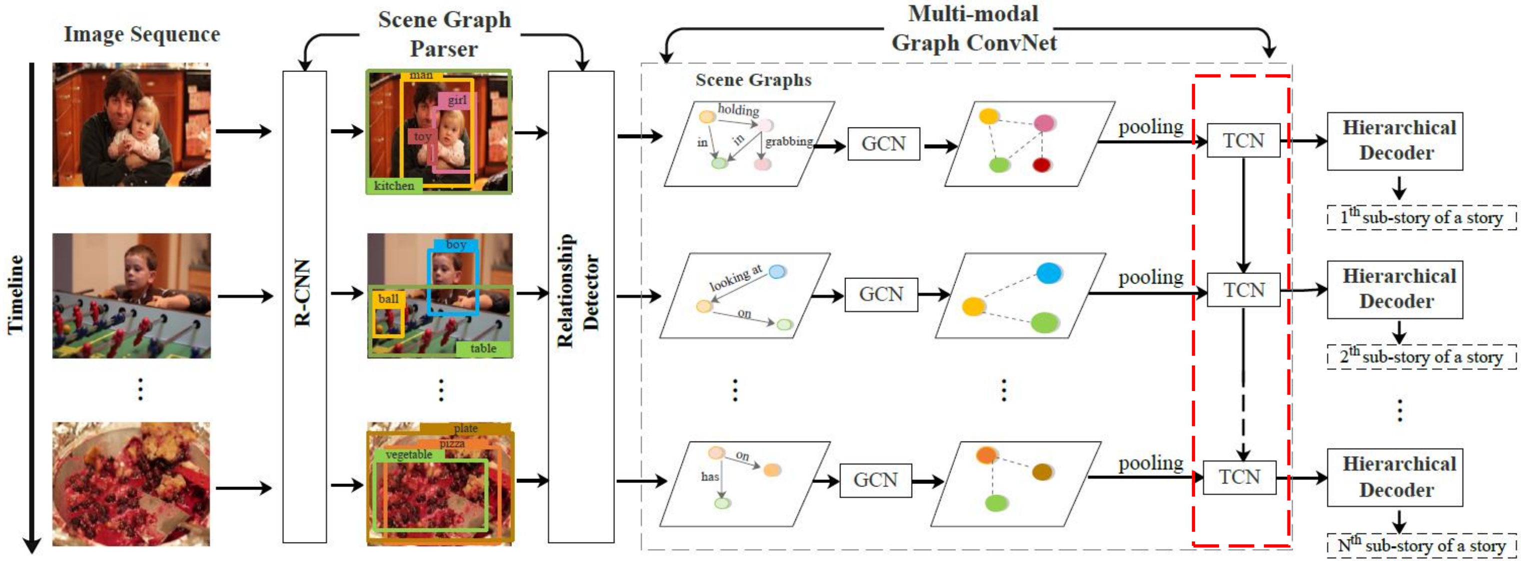


Cross-image level: *Temporal Convolution Network (TCN)*

Calculate and get single mean-pooled region vectors over K object regions $\{\mathbf{v}'_{n,i}\}_{i=1}^{K}$

$$\bar{\mathbf{v}}_n = \frac{1}{K}\sum_{i=1}^{K}\mathbf{v}'_{n,i}$$

Method – Multi-modal Graph ConvNet

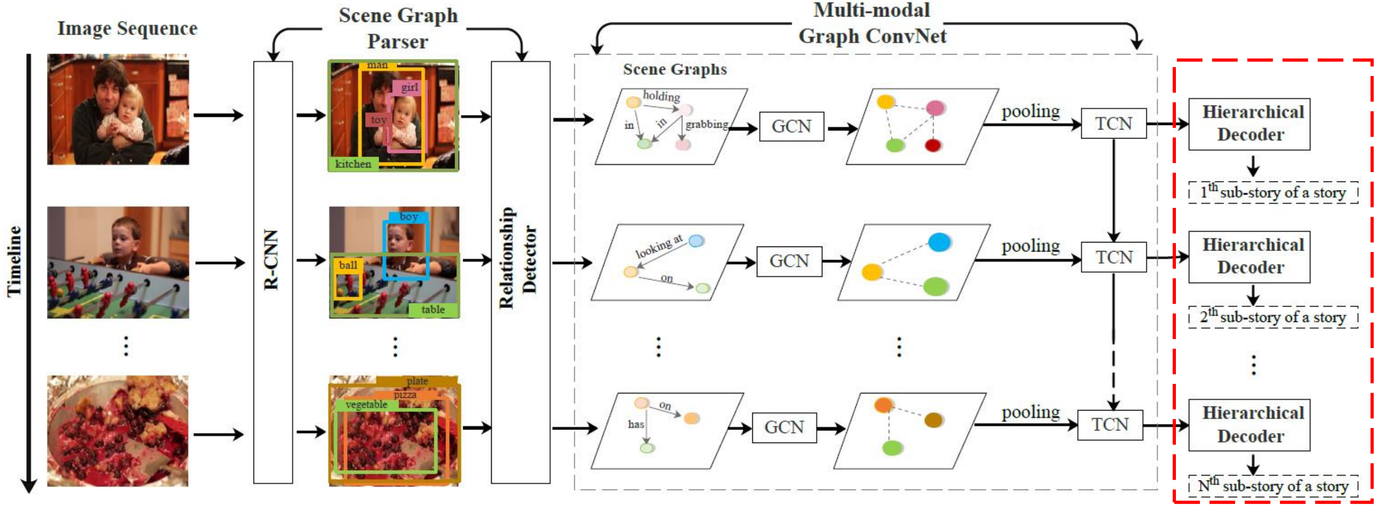


Cross-image level: *Temporal Convolution Network (TCN)*

- **Input:** region features $\bar{\mathbf{v}}_n$ **Output:** updated region features $\bar{\mathbf{v}}_n$

$$F(\bar{\mathbf{v}}_n) = \sum_{i=0}^{k-1} f(i) \cdot \bar{\mathbf{v}}_{n-d \cdot i} \qquad \bar{\mathbf{v}}_n = \mathrm{ReLU}(\bar{\mathbf{v}}_n + F(\bar{\mathbf{v}}_n))$$

Method – Hierarchical Decoder



$$h^1_{n,t} = \mathrm{GRU}(h^1_{n,t-1}, [W_s w_{n,t-1}, \bar{\mathbf{v}}_n, h^2_{n,t-1}])$$

$$Z = \tanh\left(\mathbf{W}_v \bar{\mathbf{v}}_n + \mathbf{W}_h \mathbf{h}^1_{n,t}\right)$$

$$a_{att} = \mathrm{softmax}(\mathbf{W}_z Z)$$

$$\hat{\mathbf{v}}_n = \bar{\mathbf{v}}_n a_{att}^T$$

$$h^2_{n,t} = \mathrm{GRU}\left(h^2_{n,t-1}, [w_{n,t-1}, \hat{\mathbf{v}}_n]\right)$$

$$p(w_{n,t}|w_{n,1:t-1}) = \mathrm{softmax}\left(\mathrm{MLP}(h^2_{n,t})\right)$$

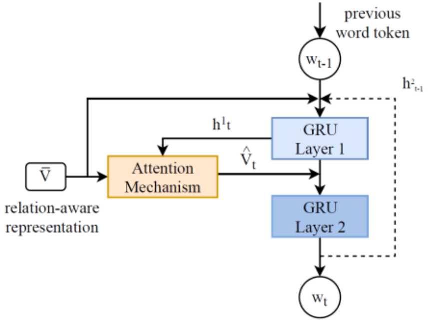


Figure 2: hierarchical story decoder

Method – Training and Inference

- **Training:**
  - Fix the parameters of our pre-trained scene graph parser (on VG dataset), and other components of our model are trained and evaluated on VIST dataset for visual storytelling task.
  - Adopt cross-entropy (MLE) loss for the training process .

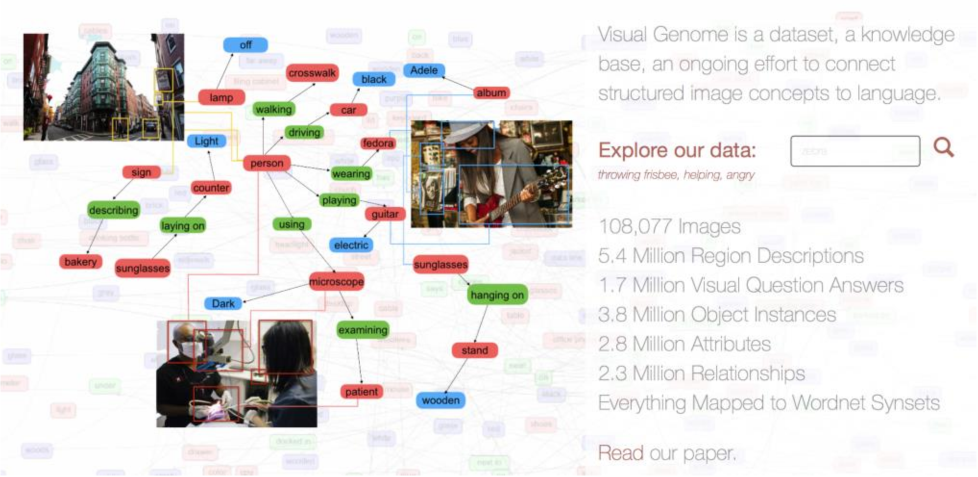
$$L(\theta) = -\sum_{t=1}^{T} log\left(p_\theta(y_t^*|y_1^*, ..., y_{t-1}^*)\right)$$

- **Inference:**
  - Adopt the beam search strategy to produce story.

Experiments and Analysis

Experiment setup

- Dataset
  - Visual Genome (VG)
    - Comprises 108,077 images annotated with scene graphs, which containing 150 object classes and 50 relation classes.
    - The VG dataset is only used to train the relationship detector in our scene graph parser.
  - VIST
    - 40,098 for training, 4,988 for validation and 5,050 samples for testing, respectively.
    - Each sample (album) contains five images and a story with five sentences.
    - Used for training and evaluating our models (except the scene graph parser) on VIST.



https://visualgenome.org

Experiment – Quantitative Results

FUDAN UNIVERSITY 1905 復旦大學 FUDAN UNIVERSITY

- Comparing with state-of-the-art

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDEr	METEOR
seq2seq† (Huang et al. 2016)	–	–	–	3.5	–	6.8	31.4
BARNN† (Liu et al. 2017)	–	–	–	–	–	–	33.3
h-attn-rank† (Yu, Bansal, and Berg 2017)	–	–	21.0	–	29.5	7.5	34.1
HPSR† (Wang et al. 2019)	61.9	37.8	21.5	12.2	**31.2**	8.0	34.4
AREL* (Wang et al. 2018b)	63.7	39.0	23.1	14.0	29.6	9.5	35.0
HSRL* (Huang et al. 2019)	-	-	-	12.3	30.8	**10.7**	35.2
SGVST w/o GCN or TCN†	62.8	38.4	22.8	13.9	29.6	8.5	35.1
SGVST w/o GCN†	63.1	39.0	23.3	14.1	29.8	8.8	35.2
SGVST w/o TCN†	**65.4**	39.8	23.5	14.2	29.6	9.3	35.4
SGVST w/ single-dec†	64.5	39.7	23.5	14.4	29.7	9.4	35.5
SGVST w/o high-level-enc†	64.9	40.0	23.6	14.5	29.8	9.6	35.6
SGVST†	65.1	**40.1**	**23.8**	**14.7**	29.9	9.8	**35.8**

Table 1: Overall performance of story generation on VIST dataset for different models in terms of automatic metrics. * directly optimized with RL rewards, e.g., the CIDEr metric, † optimized with cross-entropy (MLE). Bolded numbers are the best performance in each category.

- The proposed SGVST model achieves superior performances over other state-of-the-art models optimized with MLE and RL in almost all metrics.
- Translating the image in to graph-based semantic representation, i.e., scene graph, can benefit representing images and high-quality story generation.

Experiment – Quantitative Results

- Comparing with ablations

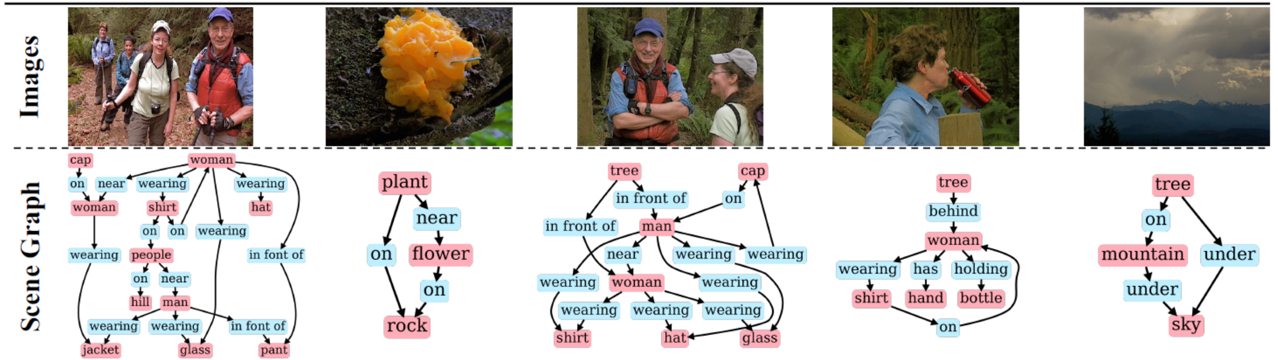
Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDEr	METEOR
seq2seq† (Huang et al. 2016)	–	–	–	3.5	–	6.8	31.4
BARNN† (Liu et al. 2017)	–	–	–	–	–	–	33.3
h-attn-rank† (Yu, Bansal, and Berg 2017)	–	–	21.0	–	29.5	7.5	34.1
HPSR† (Wang et al. 2019)	61.9	37.8	21.5	12.2	**31.2**	8.0	34.4
AREL* (Wang et al. 2018b)	63.7	39.0	23.1	14.0	29.6	9.5	35.0
HSRL* (Huang et al. 2019)	-	-	-	12.3	30.8	**10.7**	35.2
SGVST w/o GCN or TCN†	62.8	38.4	22.8	13.9	29.6	8.5	35.1
SGVST w/o GCN†	63.1	39.0	23.3	14.1	29.8	8.8	35.2
SGVST w/o TCN†	**65.4**	39.8	23.5	14.2	29.6	9.3	35.4
SGVST w/ single-dec†	64.5	39.7	23.5	14.4	29.7	9.4	35.5
SGVST w/o high-level-enc†	64.9	40.0	23.6	14.5	29.8	9.6	35.6
SGVST†	65.1	**40.1**	**23.8**	**14.7**	29.9	9.8	**35.8**

Table 1: Overall performance of story generation on VIST dataset for different models in terms of automatic metrics. * directly optimized with RL rewards, e.g., the CIDEr metric, † optimized with cross-entropy (MLE). Bolded numbers are the best performance in each category.

- *Multi-modal Graph ConvNet* module is the core component of our model since it equips the model with the capability of reasoning visual relationships through GCN on the within-image level and through TCN on the cross-images level.

Experiment – Qualitative Results

- Qualitative Examples



Story

(1) **Seq2seq:** we took a trip to the mountains . there were many different kinds of different kinds . we had a great time . he was a great time . it was a beautiful day .

(2) **AREL:** the family decided to take a trip. there were many different kinds of things to see . the family decided to go on a hike . i had a great time . at the end of the day , we were able to take a picture of the beautiful scenery .

(3) **SGVST:** we took a trip to the mountains this weekend . there were a lot of interesting plants to see . we had a great time . this woman was drinking water to relax . the view from the top was spectacular .

(4) **Ground-truth:** we went on a hike yesterday . there were a lot of strange plants there . i had a great time . we drank a lot of water while we were hiking . the view was spectacular .

- The story generated by SGVST is more coherent, informative and descriptive.

Experiment – Qualitative Results

- Human Evaluation - Pairwise Comparison

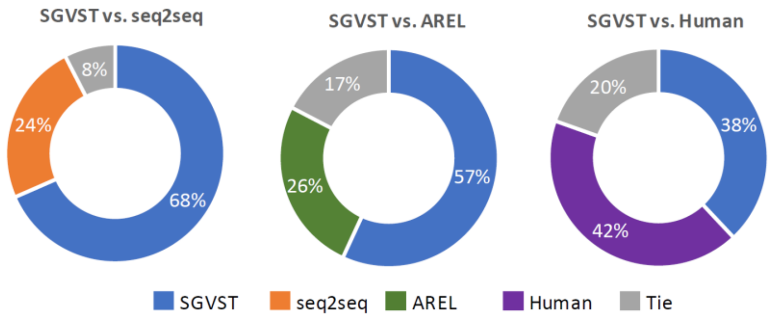


Figure 2: Each color represents the percentage of works who consider the story generated by the corresponding method is more human-like and descriptive. "Tie" in grey color indicates hard to tell.

- Human Evaluation - Human Rating

Methods	Focused	Coherent	Share	Human-like	Grounded	Detailed
seq2seq	2.30	2.33	2.12	2.22	2.30	2.30
AREL	3.51	3.53	3.37	3.43	3.31	3.39
SGVST	**3.97**	**4.01**	**3.91**	**3.99**	**4.02**	**4.07**
GT	4.37	4.40	4.21	4.38	4.32	4.39

Table 2: Human evaluation results. Workers on AMT rate the quality of the story by telling how much they Agree or Disagree with each question, on a scale of 1-5.

- The stories generated by SGVST are significantly better than stories generated by other machines, and achieve competitive performance compared with human.
- SGVST model outperforms in all six characteristics, which further proves the stories generated by our model are more informative and high-quality.

Conclusion

Conclusion

- Translating the image into graph-based semantic representation, i.e., scene graph, can benefit representing images and high-quality story generation.
- The proposed graph-based method (SGVST) can parse images to scene graphs, and reason the relationships on scene graphs on two levels, i.e., within-image and cross-images levels.
- Extensive experiments demonstrate that our method achieves state-of-the-art, and the stories generated by our method are more informative and fluent.
- The quality of scene graph generation limits the performance of our proposed method. The performance of our method can be further improved with better scene graph parser.

Thank you

Free to contact Ruize Wang (rzwang18@fudan.edu.cn) if you have any questions ^_^