Incorporating Textual Evidence in Visual Storytelling

Tianyi Li **Sujian Li**
MOE Key Lab of Computational Linguistics, School of EECS, Peking University
Peng Cheng Laboratory, Shenzhen, China
{litianyi01,lisujian}@pku.edu.cn

Abstract

Previous work on visual storytelling mainly focused on exploring image sequence as evidence for storytelling and neglected textual evidence for guiding story generation. Motivated by human storytelling process which recalls stories for familiar images, we exploit textual evidence from similar images to help generate coherent and meaningful stories. To pick the images which may provide textual experience, we propose a two-step ranking method based on image object recognition techniques. To utilize textual information, we design an extended Seq2Seq model with two-channel encoder and attention. Experiments on the VIST dataset show that our method outperforms state-of-the-art baseline models without heavy engineering.

1 Introduction

Multi-image visual storytelling is extended from a long trend of research in image captioning and has attracted considerable attention in recent years.

To generate the stories, previous work employed a Seq2Seq framework, using image encoder to encode the image sequences and sentence decoder to generate stories from encoded image sequences. Most of the researches (Smilevski et al., 2018; Kim et al., 2018; Gonzalez-Rico and Pineda, 2018; Wang et al., 2018b; Huang et al., 2018; Yu et al., 2017) focused on improving the decoder, and took simple concatenation or an LSTM as encoder. With such design, only images are utilized as input in generating the stories.

However, through our observations, the images alone are inadequate for visual storytelling. Storytelling is creative and diversified, so background knowledge is often required to convert a few images to a complete story. However, extracting such background knowledge is very difficult, especially with limited data.

To alleviate such drawback, it is important to take previous experience of story-writing into account. Imagining when a person starts to tell stories from images, he/she may not understand the implications in those images and fail to write a proper story. However, if he/she had heard others telling stories, he/she may be able to tell a story from the stories of similar image sequences he/she previously heard. Motivated by such process, we propose to utilize the large corpus as an inventory and improve the visual storytelling model by including stories from similar image sequences in corpus as input to strengthen the encoder design.

On building such models, two major problems need to be solved: (1) how to measure the relatedness of stories from the image sequence pair; (2) how to incorporate the textual information into the model so as to fully exploit it for storytelling.

To handle the first problem of picking the most relevant stories, we propose a two-step ranking method for their image sequences. We first filter out the 'dissimilar' images with object co-occurrence, and then sort the remaining candidates with feature vectors. For the second problem of incorporating textual information, we design an enhanced Seq2Seq model with two-channel encoder, one for visual input and the other for textual input.

We conduct experiments on the VIST dataset (Huang et al., 2016), a widely used multi-image visual storytelling dataset. We show that with textual evidence, our model outperforms our baselines and state-of-the-art models.

2 Method

Our method is based on the Seq2Seq framework, composed of a two-channel encoder and a RNN-based decoder. The whole architecture of our method is shown in Figure 1.

In the two-channel encoder, one channel en-

Proceedings of the 1st Workshop on Discourse Structure in Neural NLG, pages 13–17

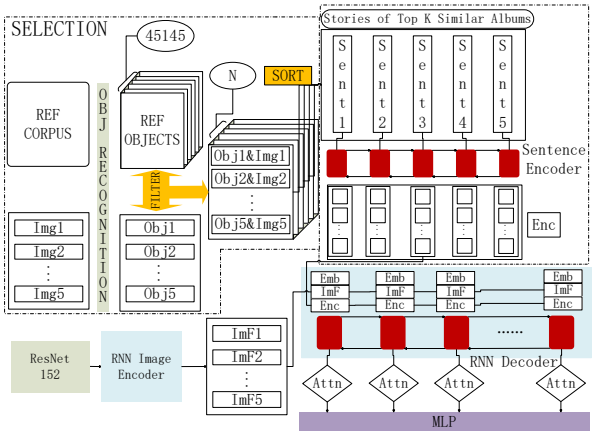

Figure 1: Overall architecture of our proposed method.

codes visual evidence from the image sequence and the other encodes textual evidence from relevant stories. In the decoder, we adopt another RNN model to generate stories from the two encoder outputs. To integrate the two types of information, we use Luong attention (2015) to dynamically attend to the stories. There are also other modifications, as further explained in 2.1.

To collect the textual evidence for encoder input, we design a selection method described in Section 2.2 to get stories from the most similar images.

2.1 Visual Storytelling Framework

Most previous works on visual storytelling followed the Seq2Seq framework, taking image recognition models such as ResNet (He et al., 2015) or Inception (Szegedy et al., 2016) to extract image features, feeding them into a story-level RNN encoder, bringing encoder output to the sentence-level decoder throughout the generation of the corresponding sentence.

We base our model on this framework with two key modifications: first, we design a text encoder to model the most similar stories which may provide evidence for story generation; second, we adopt the Luong attention Luong et al. (2015) mechanism on the textual side of encoded input to better utilize its information.

Text Encoder We use an RNN encoder to model the textual inputs. For each story, we feed its 5 sentences into the RNN one by one, retaining the hidden state across sentences. We take the RNN output of every step through the fully connected layers as encoder output.

Joint Decoder Different from previous methods, our decoder depend on both image and text encoder. The incorporation of the two encoders is the key problem. Here we adopt two approaches to solve this problem. First, we use the concatenation of the image encoder output, the embedding of last word and the last hidden states of sentence encoder as the input of the decoder. Second, we design a Luong attention layer in decoder to attend to sentence encoder outputs. Formally, the concatenation decoder can be denoted as:

$$s_t^i = DEC(s_{t-1}^i, [emb_{t-1}^i, sent^i_{len_{sent^i}}, img^i]) \quad (1)$$

and the downstream attention mechanism can be denoted as:

$$weights_t = s_t^i \cdot sent^i \quad (2)$$

$$C_t = Softmax(weight_t) \cdot sent^i \quad (3)$$

$$\pi_\beta(w_t^i | w_{1:t-1}^i) = softmax(W_c \cdot [C_t, s_t^i] + b_c) \quad (4)$$

where DEC is decoder RNN, s_t^i is RNN output for image i at step t, emb is word embedding, img and $sent$ are image and sentence encoder output, W_c and b_c are appended linear matrix and bias.

To be noticed, in our model, both decoder RNN and image encoder are generic and not limited to one particular design. The image encoder can be of arbitrary architecture as long as it generates a vector for each image, and the decoder RNN can also be designed flexibly as long as it takes a vector as input and outputs another vector at each step.

Specifically, we implemented these modifications on two popular systems: GLACNet (2018), the group with best human evaluation scores in Visual Storytelling Challenge NAACL 2018, wwho use residual encoder to generate GLOCAL vectors; XE-ss, a baseline model of Wang et al. (2018b), who proposed to improve performance with reinforcement model (AREL). We call our two models GLAC-TG and XE-TG. (see section 3.1 for details).

2.2 Textual Evidence Selection

To provide strong textual evidence for story generation, we aim to select stories which are most similar to the expected story for the given sequence of images.

With the assumption that similar images usually have similar stories, we take stories of similar im-

ages as similar stories. While it's most straightforward to choose the image with the most similar feature vector, it's shown through experiments 2 that comparing each pair of feature vectors for a large image corpus would be computationally expensive and suffer severely from false positives. Therefore, we propose to employ a two-step filter-and-sort method to pick out the most similar stories.

2.2.1 Filter

In the filter step, we use object co-occurrence to discriminate 'roughly similar' image sequences from 'dissimilar' ones. Here we filter by image object information because it conforms with the intuition that images with similar objects describe relevant events. It is also because object information has been widely used in image captioning as helpful information on images. (Mishra and Liwicki, 2019; Liu et al., 2018; Jiang et al., 2018; Anderson et al., 2017; Yin and Ordonez, 2017; Wang et al., 2018a).

We first get the types and numbers of objects in each image using an object recognition model, and then we measure image similarity with a categorical criterion and a numerical criterion. Formally, O_a and O_b are the set of objects present in image a and b respectively, c_x^k is the count of occurrence for object k in image x. The categorical criterion concerns the types of common objects, namely $score_{cat} = \frac{|O_a \cap O_b|}{\sqrt{|O_a||O_b|}}$; the numerical criterion concerns the differences in times of occurrence, namely $score_{num} = \frac{|O_a||O_b|}{|\Sigma_{k \in (O_a \cup O_b)}(c_a^k - c_b^k)^2|}$. Additionally, we set similarity scores to 0 when no objects are recognized in either image.

As mentioned above, we compare images in sequences. We measure the similarity between the sequences as the average score of its images. By filtering on the corpus and keeping only the image sequences scored on the top, we narrow down our candidate sequences to a modest size.

2.2.2 Sort

After obtaining a small set of roughly similar image sequences, we use feature vectors to rank similarity more precisely. Here we experiment on two approaches: a simple cosine similarity measure and a Bi-Linear model with Meteor score as gold annotation inspired by Cao et al. (2018). Empirically we find that Bi-Linear model shows no advantage against cosine similarity. Thus, we simply sort the roughly similar sequences with cosine similarity for downstream models.

3 Experiments

3.1 Experiment Setup

Our experiment is built on VIST (Huang et al., 2016) dataset, which is organized in 5-image sequences annotated with 5-sentence complete stories. The dataset size is 40098 for train, 4988 for validation and 5050 for test.

In GLAC-TG, we use LSTM RNN model with hidden size 1024, embedding size 256 and learning rate 1×10^{-3}; in XE-TG. We use GRU RNN model with hidden size 512, embedding size 512 and learning rate 4×10^{-4}.

In both models, we use ResNet152 (He et al., 2015) pre-trained on ImageNet (Krizhevsky et al., 2012) as image features, and we use Bi-LSTM and Bidirectional GRU respectively for image encoder.

In both models, we keep the hyper-parameters from their baseline models unmodified. For loss function, we use cross-entropy averaged on the sentence lengths.

On textual evidence selection, we use all stories and image sequences in train and validation set as reference corpus, and a Fast RCNN (He et al., 2017; Abdulla, 2017) model pre-trained on COCO dataset (Lin et al., 2014) to detect objects from each image. Roughly similar stories are filtered with numerical criterion at 500 candidate size as it shows the best performance.

3.2 Results

Methods	R	C	M
Huang et al. (2016)	-	-	31.4
Yu et al. (2017)	29.5	7.5	34.1
Gonzalez-Rico and Pineda (2018)	29.2	5.1	34.4
Huang et al. (2018)	30.8	10.7	35.2
GLACNet(2018) (re-trained)	26.3	2.2	33.0
GLAC-TG-top1(ours)	**26.5**	**2.0**	**33.4**
XE-ss(2018b)	29.7	8.7	34.8
AREL(2018b)	29.9	8.4	35.2
XE-TG-top1(ours)	**30.0**	**8.7**	**35.5**
XE-TG-top3(ours)	29.6	8.3	35.4
XE-TG-top1-attn(ours)	**29.9**	**9.2**	**35.2**
XE-TG-top3-attn(ours)	29.4	9.2	35.0
XE-TG-only	29.1	7.7	34.8

Table 1: Performance of our method compared to existing visual storytelling models, R is ROUGE-L, C is CIDEr, M is METEOR (models we re-trained in same setting as original are listed in (re-trained) rows)

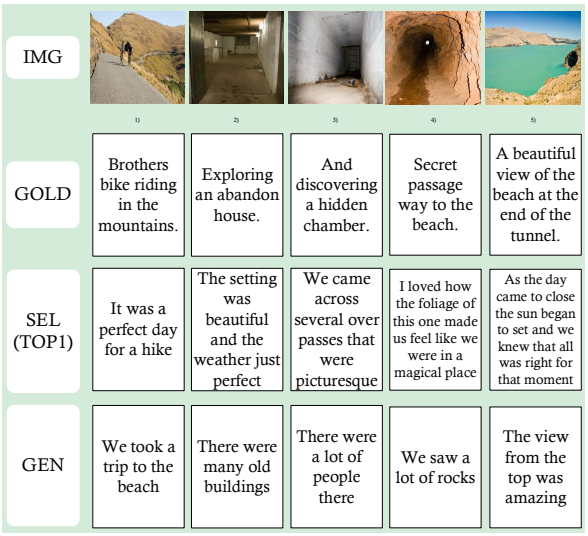

Figure 2: An example sequence of visual storytelling.

In Table 1, we compare our models with several strong baselines on three automatic evaluation metrics, ROUGE-L, CIDEr and METEOR. In the top block of Table 1, we present 4 previous baselines: 1) a standard Seq2Seq baseline model developed by Huang et al. (2016); 2) a hierarchically attentive model designed by Yu et al. (2017); 3) the Seq2Seq model with sentence-wise separate decoders by Gonzalez-Rico and Pineda (2018); 4) reinforcement learning with topic guided decoders by Huang et al. (2018). In the middle block, we present the GLACNet model Kim et al. (2018) and our improved GLAC-TG model. In the bottom block, we present our XE-TG models which are improved based on the XE-ss model in AREL framework (Wang et al., 2018b). For fair comparison, we evaluate all models with the open source evaluation code[1] (Yu et al., 2017).

Result shows that both our models outperform their corresponding baselines. Even using textual evidence only, our XE-TG-only model shows competitive performance compared to the baselines. Moreover, our XE-TG models using cross entropy loss outperformed state-of-the-art baselines with reinforcement learning techniques (Wang et al., 2018b; Huang et al., 2018). By using simple cross entropy loss, our models are also less costly to train, easier to tune and more stable when re-trained.

We conduct a qualitative analysis on XE-TG-top1 model in Figure 2 as an example. It shows that the selected similar story shares the same topic of wilderness adventure with similar story-flows. The generated story also catches the essence of the image sequence, with basic details closely relevant. It shows that our textual evidence selection method is capable of selecting proper textual evidence, and our storytelling framework is capable of capturing the provided information and telling fluent and coherent stories.

3.3 Analysis on Textual Evidence Selection

In this section, we further explore the effectiveness of similar stories. We experimented on filtering candidate size 50, 100 and 500 with both categorical and numerical criteria, using sorting on the entire reference corpus for comparison and METEOR score as a metric of actual story similarity. In Table 2, we show that for all methods, the selected stories are significantly more similar to gold stories than randomly selected ones, and stories with higher rankings are generally better than those with lower rankings. Moreover, for both criteria, candidate size poses negligible effect.

On the other hand, neither sorting on full corpus nor sorting by bi-linear model shows competitive results compared to our approach.

M	categorical			numerical		
	50	100	500	50	100	500
1	24.8	24.8	25.0	24.9	24.7	24.5
2	24.9	24.8	24.7	24.4	24.5	24.6
3	24.6	24.5	24.6	24.6	24.6	24.5
4	24.5	24.9	24.8	24.5	24.5	24.3
5	24.8	24.6	24.6	24.5	24.5	24.5
rand	23.8					
full	23.28 (average on top 5)					
B-L	23.62 (average on top 5)					

Table 2: METEOR scores for top 1 to 5 similar stories regarding two criteria, B-L refers to Bi-Linear

4 Conclusion

In this paper, we show that textual evidence from similar image sequences contains rich information for visual storytelling, therefore it's capable of boosting storytelling performance. We propose a feasible two-step approach to extract textual evidence from a large corpus. We also design a two-channel encoder to incorporate textual and visual evidence into the Seq2Seq visual storytelling models and achieve state-of-the-art performance with-

[1] https://github.com/lichengunc/vist_eval

out heavy engineering.

Acknowledgments

We thank the anonymous reviewers for their helpful comments on this paper. This work was partially supported by National Natural Science Foundation of China (61572049 and 61876009).

References

Waleed Abdulla. 2017. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and VQA. *CoRR*, abs/1707.07998.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2018. Retrieve, rerank and rewrite: Soft template based neural summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 152–161, Melbourne, Australia. Association for Computational Linguistics.

Diana Gonzalez-Rico and Gibran Fuentes Pineda. 2018. Contextualize, show and tell: A neural visual storyteller. *CoRR*, abs/1806.00738.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. *CoRR*, abs/1703.06870.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Qiuyuan Huang, Zhe Gan, Asli Çelikyilmaz, Dapeng Oliver Wu, Jianfeng Wang, and Xiaodong He. 2018. Hierarchically structured reinforcement learning for topically coherent visual story generation. *CoRR*, abs/1805.08191.

Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*.

Wenhao Jiang, Lin Ma, Xinpeng Chen, Hanwang Zhang, and Wei Liu. 2018. Learning to guide decoding for image captioning. *CoRR*, abs/1804.00887.

Taehyeong Kim, Min-Oh Heo, Seonil Son, Kyoung-Wha Park, and Byoung-Tak Zhang. 2018. GLAC net: Glocal attention cascading networks for multi-image cued story generation. *CoRR*, abs/1805.10973.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Fenglin Liu, Xuancheng Ren, Yuanxin Liu, Houfeng Wang, and Xu Sun. 2018. simnet: Stepwise image-topic merging network for generating detailed and comprehensive image captions. *CoRR*, abs/1808.08732.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025.

Ashutosh Mishra and Marcus Liwicki. 2019. Using deep object features for image descriptions. *CoRR*, abs/1902.09969.

Marko Smilevski, Ilija Lalkovski, and Gjorgji Madjarov. 2018. Stories for images-in-sequence by using visual and narrative components. *CoRR*, abs/1805.05622.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Josiah Wang, Pranava Swaroop Madhyastha, and Lucia Specia. 2018a. Object counts! bringing explicit detections back into image captioning. *CoRR*, abs/1805.00314.

Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. 2018b. No metrics are perfect: Adversarial reward learning for visual storytelling. *CoRR*, abs/1804.09160.

Xuwang Yin and Vicente Ordonez. 2017. OBJ2TEXT: generating visually descriptive language from object layouts. *CoRR*, abs/1707.07102.

Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2017. Hierarchically-attentive RNN for album summarization and storytelling. *CoRR*, abs/1708.02977.