# GLAC Net: GLocal Attention Cascading Networks for Multi-image Cued Story Generation

**Taehyeong Kim[1], Min-Oh Heo[2], Seonil Son[2], Kyoung-Wha Park[3], Byoung-Tak Zhang[1,2,3]**

1 Program in Cognitive Science, 2 Dept. of Computer Sci. and Eng., 3 Program in Neuroscience

Seoul National University

Seoul 08826 South Korea

{thkim,moheo,sison,kwpark,btzhang}@bi.snu.ac.kr

## Abstract

The task of multi-image cued story generation, such as visual storytelling dataset (VIST) challenge, is to compose multiple coherent sentences from a given sequence of images. The main difficulty is how to generate image-specific sentences within the context of overall images. Here we propose a deep learning network model, GLAC Net, that generates visual stories by combining global-local (glocal) attention and context cascading mechanisms. The model incorporates two levels of attention, i.e., overall encoding level and image feature level, to construct image-dependent sentences. While standard attention configuration needs a large number of parameters, the GLAC Net implements them in a very simple way via hard connections from the outputs of encoders or image features onto the sentence generators. The coherency of the generated story is further improved by conveying (cascading) the information of the previous sentence to the next sentence serially. We evaluate the performance of the GLAC Net on the visual storytelling dataset (VIST) and achieve very competitive results compared to the state-of-the-art techniques.

## 1 Introduction

Deep learning have brought about break-throughs in processing image, video, speech and audio (LeCun et al., 2015). The field of natural language processing has been also interested in deep learning, e.g., sentence classification (Kim, 2014; Iyyer et al., 2015), language modeling (Bengio et al., 2003; Mikolov et al., 2013), machine translation (Sutskever et al., 2014; Bahdanau et al., 2014; Wu et al., 2016), and question answering (Hermann et al., 2015). Naturally, bridging images and texts by deep learning has been following (Belz et al., 2018) such as image captioning (Vinyals et al., 2015; Xu et al., 2015; Karpathy and

Fei-Fei, 2017), visual question answering (Antol et al., 2015; Kim et al., 2016), and image generation from caption (Reed et al., 2016; Zhang et al., 2017).

The task of multi-image cued story generation is one of interesting visual-linguistic challenges to generate story of multiple coherent sentences from a given sequence of images. The main difficulty is how to generate image-specific sentences within the context of overall images. Additionally, it is harder than object recognition or image captioning since it needs fine-grained object recognition and context understanding in the images. Recently, visual storytelling dataset (VIST) was released for the task of multi-image cued story generation, which is composed of five-sentence stories, descriptions and the corresponding sequences of five images (Huang et al., 2016).

Here we propose a deep learning network model that generates visual stories by combining global-local (glocal) attention and context cascading mechanisms. To focus on the image-specific appropriateness of them, we develop two levels of attention, i.e., overall encoding level (global) and image feature level (local). In the image sequence encoders, the global context of the storyline is encoded using bi-directional LSTMs on features of five images, we give attention on the context (*global* attention). Additionally, we give *local* attention to image features directly. Then both of them are combined and sent to RNN-based sentence generators. While standard attention configuration needs a large number of parameters, we implement them in a very simple way via hard connections from the outputs of encoders or image features onto the sentence generators. To improve further the coherency of the generated stories, we design to convey the last hidden vector in the sentence generator to the next sentence generator as an initial hidden vector.

This paper is organized as follows. Section 2 presents related works to positioning. In Section 3 we show briefly dataset, section 4 explains the proposed models. Section 5 shows their experimental results. Finally, section 6 draws the conclusion.

## 2 Related Work

**Text Comprehension** As similar works without visual cues, there are text comprehension tasks such as bAbI tasks (Weston et al., 2015), SQuAD (Rajpurkar et al., 2016) and Story Cloze Test (Mostafazadeh et al., 2016). They have been widely used to benchmark new algorithms on document comprehension or story understanding.

**Visually-grounded Comprehension** Since AlexNet (Krizhevsky et al., 2012) as a milestone, object recognition/detection methods have grown explosively and outperformed human ability to capture objects in accuracy aspect (Geirhos et al., 2017). As used in our model, those visual features turned out to be also available as general features for scene description (Johnson et al., 2016; Karpathy and Fei-Fei, 2017), image captioning with attention (Xu et al., 2015), and image/video question answering about the stories (Tapaswi et al., 2016; Kim et al., 2017).

**Story Generation from Images** In the first work for image cued sentence generation (Farhadi et al., 2010), the triplet - <object, action, scene> was predicted for an input image using MRF, and used for searching or generating with templates. In the deep learning era, Jain et al. (2017) utilized the VIST dataset to translate description-to-story without images. Liu et al. (2017) developed semantic embedding of the image features on the bi-directional recurrent architecture to generate a relevant story to the pictures.

## 3 VIST Dataset

VIST dataset is a dataset of story-like image sequences paired with: (1) descriptions for each image in isolation (DII) ($\sim 80\%$ only), and (2) descriptions to form a narrative over an image sequence (images/sentences aligned each) (SIS) as shown in Figure 1. It consists of 50,200 sequences (stories) using 209,651 images (train: 40,155, validation: 4,990, test: 5,055).

## 4 Approaches to Story Generation

The main difficulty of the multi-image cued story generation is to keep the overall context of



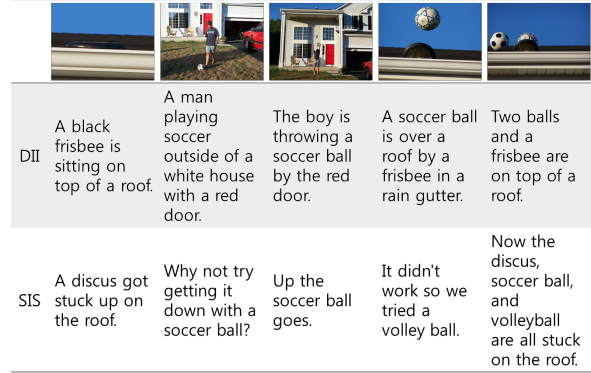| | | | | | |
|---|---|---|---|---|---|
| DII | A black frisbee is sitting on top of a roof. | A man playing soccer outside of a white house with a red door. | The boy is throwing a soccer ball by the red door. | A soccer ball is over a roof by a frisbee in a rain gutter. | Two balls and a frisbee are on top of a roof. |
| SIS | A discus got stuck up on the roof. | Why not try getting it down with a soccer ball? | Up the soccer ball goes. | It didn't work so we tried a volley ball. | Now the discus, soccer ball, and volleyball are all stuck on the roof. |

Figure 1: A VIST dataset example. DII: Descriptions of images in isolation. SIS: Stories of images in sequence.

the story while generating a well-aligned sentence for each image. To tackle it, we introduce 2 key ideas: (1) utilizing two-level attention mechanism in the encoder part, and (2) conveying the hidden state to the next sentence generator.

### 4.1 Two-levels of Attention

Soft attention mechanism (Bahdanau et al., 2014) utilizes additional weights on the inter-related outputs of the nodes, which improves the performance of the basic encoder-decoder model in machine translation. In the task of story generation from image sequences, however, each sentence should be visually grounded on not only each image but also overall context. To represent these relationship, we design to deliver the two-channel information to each decoder (1) from low-level image features, and (2) from high-level encoded features together. We implement them via a simple hard attention mechanism that focuses sequentially on the encoder output when generating story-like sentences.

In the sequence to sequence configuration, we can choose one of the followings as encoder of image sequences: concatenation of image features (for short sequences with the same length), uni-directional RNN, and bi-directional RNN. We choose bi-RNN because it is better for aggregated representations. For image specificity, we feed both bi-RNN outputs and image-specific features to decoders. The outputs of bi-RNN include overall information of the sequence (global). On the other hand, image-specific features are constrained only on the image (local). The glocal vectors are obtained by concatenating image-specific features and bi-RNN outputs. Including glocal vector as decoder inputs can be seen as
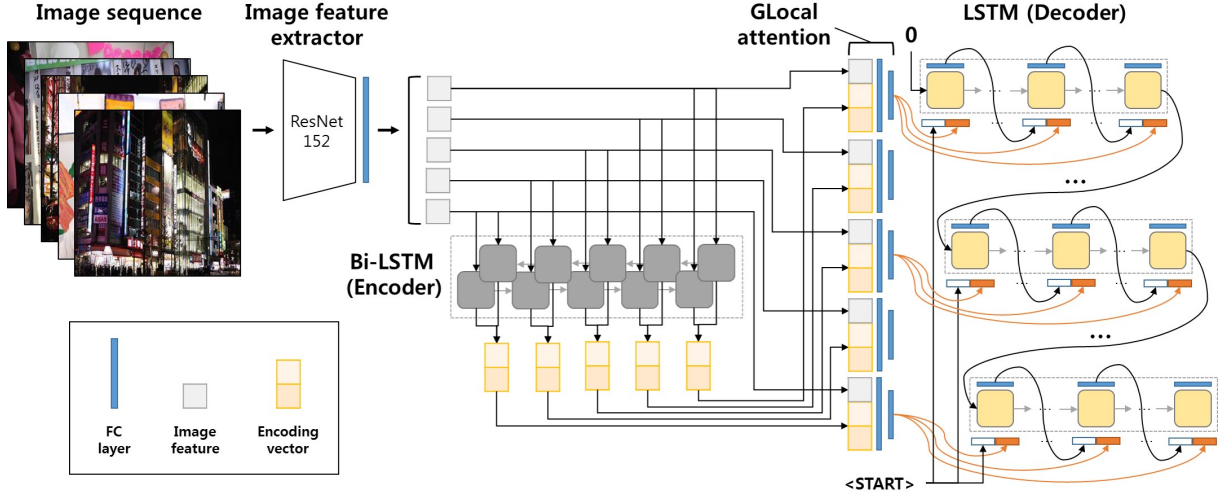
Figure 2: The global-local attention cascading (GLAC) network model for visual story generation. Note: activation function (ReLU), dropout, batch normalization, and softmax layer are omitted for readability.

an 'hard' attention mechanism, which emphasizes both image-specific and overall information.

## 4.2 Model Details

Figure 2 illustrates deep learning architecture for the story generation proposed in this paper. (1) the features of each image are extracted using the ResNet-152 (He et al., 2015). (2) The extracted features are sequentially fed into the bi-LSTM so that the context of the images can be evenly reflected in the entire story. The glocal vectors made up of bi-LSTM outputs and image-specific features go through the fully connected layers. After that, it is concatenated to the word tokens in order to be used as inputs to the decoder. Note that one glocal vector is used until the decoder meets an '<END>' token which denotes the end of sentence; five glocal vectors for each image works the same as described above. The cascading mechanism conveys the hidden state (context) of the previous sentence to the next sentence. The hidden state of the LSTM is initialized to zeros only at the beginning of the first sentence of the story for maintaining story context.

As a simple heuristics to avoid duplicates in the resulting sentence, we sample words one hundred times from the word probability distribution of the LSTM output, and choose the most frequent word from the sampled pool. This reduces the number of repetitive expressions and improve the diversity of the generated sentences. On the process of generating sentences of the story, We also count the selected words. The selection probabilities of the words are decreased according to the frequency of each word as Equation 1, and normalized.

$$\hat{p}(word) = p(word) \times \frac{1}{1 + k \cdot count_{word}} \quad (1)$$

where $k$ is a constant for sensitivity.

To build grammatically correct sentences, the probabilities of some function words such as prepositions and pronouns are not changed regardless of the frequency of occurrence.

## 4.3 Network Training

The training images are resized to $256 \times 256$ before training, and then augmented with random cropping of $224 \times 224$ accompanied by a horizontal flip at the training time. Pixels are normalized to [0,1]. Learning rate and weight decay are set to 0.001 and 1e-5 respectively and optimized with Adam optimizer. Each word is embedded into a vector of 256 dimensions, and the LSTM is trained with teacher forcing manner. Also, we apply batch normalization and dropout layers to prevent overfitting and improve the performance. Batch size are set to 64, and the training data is reshuffled at every epoch.

## 5 Experiments and Discussion

### 5.1 Experiment Settings

In order to evaluate the effects of the GLAC Net, we performed an ablation study as follows. The first model is a simple LSTM Seq2Seq network. In the second model, we remove context cascading from the full GLAC Net architecture.

| Used Attention | Valid Perplexity | Test Perplexity | METEOR Score |
|---|---|---|---|
| Baselines (Beam=10) | - | - | 0.2313 |
| Baselines (Greedy) | - | - | 0.2776 |
| Baselines (-Dups) | - | - | 0.3011 |
| Baselines (+Grounded) | - | - | 0.3142 |
| LSTM Seq2Seq | 21.89 | 22.18 | 0.2721 |
| GLAC Net (-Cascading) | 20.24 | 20.54 | **0.3063** |
| GLAC Net (-Global) | 18.32 | 18.47 | 0.2913 |
| GLAC Net (-Local) | 18.21 | 18.33 | 0.2996 |
| GLAC Net (-Count) | 18.13 | 18.28 | 0.2823 |
| GLAC Net | **18.13** | **18.28** | 0.3014 |

Table 1: Results from experiment settings. Baselines are reported in (Huang et al., 2016).

The third and fourth models are for testing the effect of global and local attention respectively. In the fifth, we remove the post processing routines to avoid word duplication when generating sentences. The last is the complete GLAC Net model.

## 5.2 Results and Discussion

We evaluate trained models on the VIST dataset. The evaluation criteria are perplexity and METEOR metric and the results are shown in Table 1. Compared with the performance of baselines (Huang et al., 2016), the GLAC Net is also competitive without beam search methods. From the results of 'GLAC Net (-Count)' and 'Baselines (-Dups)' in Table 1, the heuristics are helpful to reduce redundant sentences and improve the METEOR score. Compared to LSTM Seq2Seq models, GLAC Net-based model shows better performance in general. Although the differences are not much significant between the GLAC Net experiment settings, the complete GLAC Net shows the best overall performance.

We also consider human evaluation criteria to choose better models (Mitchell et al., 2018). Figure 3 shows the examples of the generated stories with the test dataset. The context of successive images is well reflected, and the content of each image is properly described. However, the story development is slightly monotonous. Sometimes the content is different from the specific image or the sentence is a little awkward. It is observed that most generated sentences have simple structures.

## 6 Conclusion and Future Work

We proposed the GLAC Net that uses glocal attention and context cascading mechanisms to generate stories from a sequence of images. The model is designed to maintain the overall context
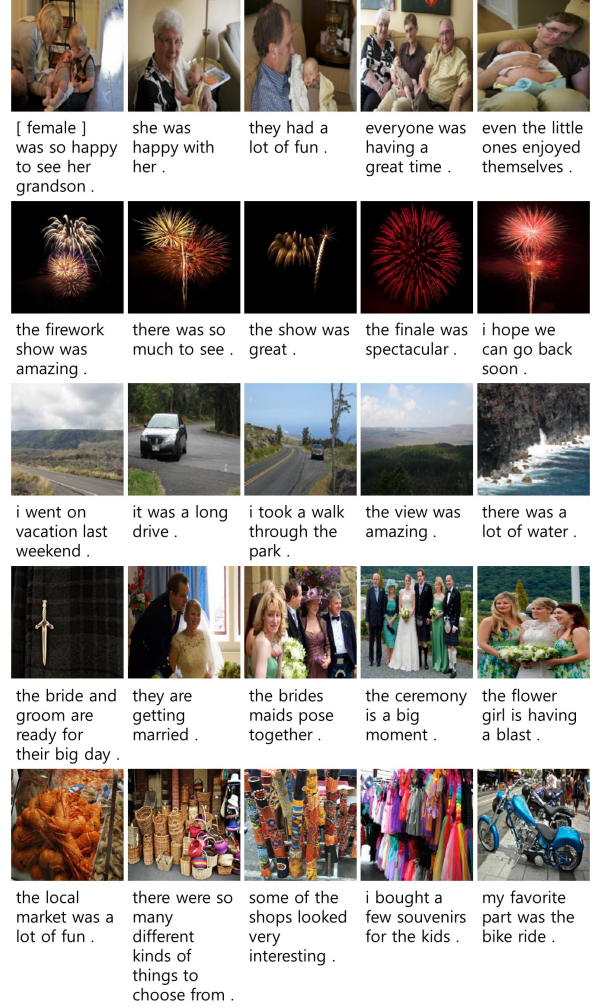


Figure 3: Samples of multi-image cued story generation results

of the story from the image sequence and to generate context-aware sentences for each image. In the experiment using the VIST dataset, the proposed model is proved to be effective and competitive in storytelling task according to the crowd-sourced human evaluation results with METEOR score ~0.3.

Although the experimental results are promising, visual storytelling task is remaining as a challenge. We are planning to extend and refine the GLAC architecture to further improvement of its performance. Also, generating various stories based on the purpose and theme from the same image sequence would be the following topic to be explored in the future works.

## Acknowledgments

# References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Anya Belz, Tamara L. Berg, and Licheng Yu. 2018. From image to language and back again. *Journal of Natural Language Engineering*, 24(3):325–362.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer.

Robert Geirhos, David H. J. Janssen, Heiko H. Schütt, Jonas Rauber, Matthias Bethge, and Felix A. Wichmann. 2017. Comparing deep neural networks against humans: object recognition when the signal gets weaker. *CoRR*, abs/1706.06969.

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

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

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)*.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1681–1691.

Parag Jain, Priyanka Agrawal, Abhijit Mishra, Mohak Sukhwani, Anirban Laha, and Karthik Sankaranarayanan. 2017. Story generation from sequence of independent short descriptions. *arXiv preprint arXiv:1707.05501*.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Andrej Karpathy and Li Fei-Fei. 2017. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676.

Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2016. Multimodal residual learning for visual qa. In *Advances in Neural Information Processing Systems*, pages 361–369.

Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. 2017. Deepstory: video story qa by deep embedded memory networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2016–2022. AAAI Press.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA. Curran Associates Inc.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436.

Yu Liu, Jianlong Fu, Tao Mei, and Chang Wen Chen. 2017. Let your photos talk: Generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks. AAAI Conference on Artificial Intelligence.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Margaret Mitchell, Ishan Misra, Ting-Hao K. Huang, and Frank Ferraro. 2018. Human evaluation criteria for visual storytelling challenge.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James F. Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *CoRR*, abs/1604.01696.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. MovieQA: Understanding Stories in Movies through Question-Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *CoRR*, abs/1502.05698.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France. PMLR.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE Int. Conf. Comput. Vision (ICCV)*, pages 5907–5915.