

Leveraging Textual Information for Visual Story Generation

Merve Gül Kantarcı and Muhammed Yasin Adıyaman

Boğaziçi University

Department of Computer Engineering

34342 Bebek, Istanbul, Turkey

{adiyamanmyasin, mervegulkantarci}@gmail.com

Abstract

This paper describes a visual story generation system which uses image annotations as an auxiliary input. To fuse the image features with the textual ones, we propose the middle-fusion method: The first two neural paths are responsible for extracting visual and textual features by mapping visual and textual inputs into feature spaces and then before decoding they are concatenated to form a single representation of an image. For extracting visual features we used off-the-shelf pretrained ResNet152 architecture with a cascaded BiLSTM to leverage temporal dependencies between images. For textual feature extraction we used again off-the-shelf BERT model. Then, the extracted features are concatenated and fed to the final story generation model which is realized as a 2-layer LSTM network. The results show that combining the textual and image features is promising and can be further investigated in many research directions.

1 Introduction

Story-telling is one of the most challenging tasks in natural language processing (NLP) field since it is unstructured, open-ended, and highly objective task. Thus, in this context, the main challenges in story-telling become finding a good architecture, generating reasonable and diverse outputs, and evaluating the results fairly. Generally, story generation models are expected to generate stories conditioned on some given context rather than just random generation. The input context can be textual or visual or both. In our study we conditioned the generation model on both visual and textual input contexts utilizing 2-stage encoder-decoder architecture.

Since stories contain rich information coming from commonsense knowledge, the information gap between the input and output could be high, which results in non-coherent or irrelevant stories to the input context. To reduce the information gap, a possible option is to insert a squeezed intermediate layer which avoid divergence from the context. However, there is a trade-off between diversity of the generated stories and quality of them (like GAN networks). To alleviate this issue, our intermediate layer acts like a common feature space to concatenate visual and textual features, rather than readable but restrictive outlines.

In our model, both textual and visual inputs are obtained in vector representations. Then, the generated features are fed to a cascaded generation model which is conditioned on the concatenation of textual and visual features. We used VIST dataset (Huang et al., 2016) for training which consists of image sequences and human-generated short stories related to them. We try to generate 5-sentence short stories from an image sequence, length of 5, and for each image we use the images' description annotation as an auxiliary textual input which is also readily available in the dataset.

Following the VIST 2018 challenge¹, we evaluate our model with automatic evaluation metric METEOR. Although we could not convey the any ablation study to understand the dynamics of the model better, by looking the generated stories we infer that our model is able to merge the visual and textual features successfully.

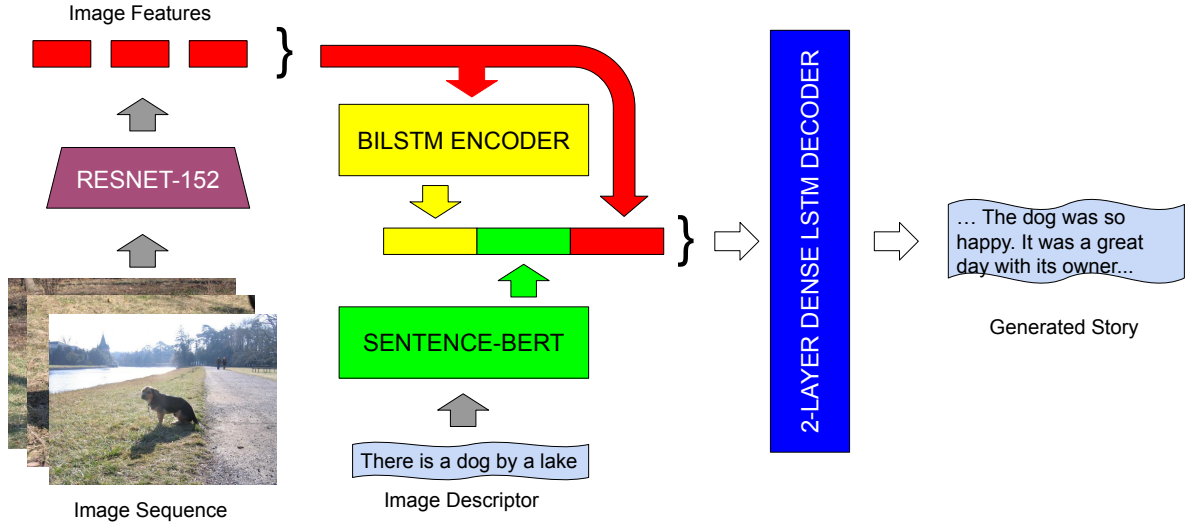


Figure 1: Our proposed model for story generation which consists of (a) RESNET-152 for image raw feature extraction; (b) BiLSTM encoder for leveraging temporal dependencies between images; (d) Sentence-BERT for extracting textual features from image annotations; and (d) 2-layer dense LSTM decoder for story generation conditioned upon the concatenation of textual and visual features.

2 System Description

The model vocabulary is created using only story sentences in the training set. Since we make use of descriptions at only sentence embedding level, we did not consider using the whole vocabulary fair and useful. 4 special tokens are added to vocabulary for special purposes: <pad> token to make each target story sentence in a sequence to have the same length, <start> token to denote start of a sentence, <end> token to denote end of a sentence, <unk> token for out-of-vocabulary words. Word frequency threshold to include in vocabulary is defined to be 10.

To leverage the image descriptions and make them involved in story generation model we should map them into a some intermediate layer. There are several methods for information fusion in neural networks. We can categorize the possible solutions into early-fusion, middle-fusion and late-fusion methods. In early fusion, one of the input space is mapped to another by some feature mapping network before further processing. Then, the converted features are combined fed to the subsequent layers together. In this case, since we enforce mapping between different types of data it is possible to lose information during mapping. On the other hand, in the late fusion there are multiple paths for generating different outputs. Then, the generated outputs are compared and the best one is selected or some average output is generated. However, in this case it is not possible to leverage one input data when processing the other one. As a compromise between the two methods we can meet them in a some intermediate layer, which is called middle-fusion. We choose this method in our study. Because, it offers soft transition for data sources and leverage one for the other when processing them.

Overall architecture can be explained in 2 parts: encoder and decoder. Encoder consists of 3 main modules to extract images, extract sentence embeddings, extract temporal dependencies between images. After concatenating the outputs of those encoder modules an attention layer is applied also. Decoder takes the output of the attention layer to generate the actual story corresponding given text and image data.

The proposed architecture is very similar to GLACNet architecture, hence we implement our model using their publicly available GitHub repository ².

¹Details can be found at <http://visionandlanguage.net/workshop2018/>

²<https://github.com/tkim-snu/GLACNet>

2.1 Image Encoder

Following the literature, ResNet-152 (He et al., 2015) is selected as image feature extractor. The pre-trained ResNet-152 model is available on many frameworks such as Tensorflow, Pytorch etc. Since we implemented the end-to-end model in PyTorch, we used PyTorch here as well. From the last hidden layer of the ResNet the raw image features are obtained. The ResNet model is used as pretrained and not optimized during training. For each image, 1x2048 sized constant vectors are obtained. This part is encoded using a simple FC layer into a more dense representation to align dimensions with sentence encodings.

2.2 Global Feature Encoder

Following the recent works in the literature, biLSTM is used to extract temporal relations between images. This can be considered as a way to enhance the single image vectors to obtain globally consistent vectors. The intuition behind using biLSTM is to infer the context both using the forward and backward context.

2.3 Sentence Encoder

There are several options to encode image descriptions including RNN, biLSTM or transformer. Among them, transformers shows SOTA performance in many NLP tasks including sentence encoding task. Therefore, we used an off-the-shelf pretrained BERT sentence encoder, which is called Sentence-BERT (Reimers and Gurevych, 2019). The base model was trained for natural language inference task which we consider as a perfect match for our case and a potentially good example for transfer learning. What we try to achieve using image descriptions is an inference to source images as an auxiliary information source. The primary source of the prompt is still image sequence, however our hypothesis is that using an auxiliary text description improves the quality of generation. This part is what distinguishes our work from the literature. To the best of our knowledge, we are the first use both descriptions and images together to construct a story.

2.4 Decoder

First part of the decoder learns word embeddings. The output of encoder is concatenated with learned word embeddings of the target text. Embedding layer is followed by a drop-out layer and by a 2-layer dense LSTM. Afterwards, another dropout layer is applied. Final layer is a fully-connected layer followed by softmax layer to learn word distributions. Dropout layers have a key role in this architecture since the data size is not large compared to the size of the network.

3 Experimental Setup

3.1 Dataset

We use the de facto dataset to train and evaluate the network: Visual Storytelling Dataset (VIST). The datasets consist around 50000 stories with 230000 images. However some images are removed by the owners or does not include descriptions are stories. Once they are eliminated, we have around 40000 stories. Because of these eliminations, not all stories have 5 image sequence with corresponding 5 story and description sentences. This needs to be fixed since the network is not scalable to have variable length sequences. One setting would be padding images, sentences and descriptions to obtain strictly 5 image sequence. We remove sequences with less than 5 image-story-description pairs to work with a cleaner dataset compared to padding case. We use the train/val/test splits that is released with VIST dataset. The ratio is around 80%, 10% and 10% respectively. After cleaning we observe that ratios are similar, hence cleaning process does not disrupt the common splits in the literature that are used with this dataset.

In the resulting dataset, there are 19096 stories with 30254 unique images in train set, 2030 stories with 3043 unique images in validation set and 2594 stories with 3874 unique images in test set. Since the images are originally very large and have varying sizes, we resize all of them in to 256x256 pixels before training the network.






	1	2	3	4	5
					
Desc-in-Isolation	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.
Desc-in-Sequence	A roof top with a black frisbee laying on the top of the edge of it.	A man is standing in the grass in front of the house kicking a soccer ball.	A man is in the front of the house throwing a soccer ball up.	A blue and white soccer ball and black Frisbee are on the edge of the roof top.	Two soccer balls and a Frisbee are sitting on top of the roof top.
Story-in-Sequence	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 2: VIST Dataset: An example image-story-description sequences (Huang et al., 2016)

3.2 Implementation Details

The network is implemented using PyTorch framework. The final model is trained on Google Colab environment. Each epoch takes around 15 minutes. In order to reduce the memory usage for early experiments, Sentence-BERT and ResNet models are used to extract features prior to training and the vectors only are included during training or inference.

Sentence-Bert and ResNet-152 model is used as pretrained and does not fine-tuned. All the other parts of the network is trained in end-to-end fashion. Cross entropy loss is optimized using Adam optimizer. Initial learning rate is determined as 0.001 and weight decay is determined as 0.00001. Batch size is 32. The training set is shuffled in each epoch. If validation loss does not decrease in 5 consecutive epochs, the model training stops. With this setting the model is trained for approximately 40 epochs.

4 Results and Discussion

4.1 Quantitative Results

We compare our results with the results (obtained from papers) of some of the models from the literature using METEOR score. METEOR score is obtained using official evaluator of VIST 2018 challenge.

Model	Loss	Perplexity	METEOR	Dataset Size
AREL (Wang et al., 2018)	-	-	0.35	~100%
GLACNet (Kim et al., 2018)	-	18.28	0.30	~100%
CAMT (Aljawy et al., 2021)	-	15.89	0.34	~100%
Ours	3.04	20.97	0.30	~50%

Table 1: Evaluation results in comparison with the state-of-the-art baseline models.

Table 1 clearly shows that state-of-the-art results are not obtained. However, one interesting outcome of the results are it shows very close performance (see METEOR score) to GLACNet which we based our model on top of that. Another interesting point is that our model uses around half of the dataset to obtain the presented results. The size of dataset has a key role in the quality of text generation models. Hence we still find the results promising from this perspective. In order to further investigate the effect, models with the implementation results can be run again to see the results with the reduced dataset. However, we are not able to make this experiment currently due to time and resource limitations. GLACNet provides very good results with VIST dataset, however in more recent works, better performing architectures are presented. We consider our model as a plug-in in terms of its implementation efficiency. Hence we suggest that with our proposed method, this can be applied to similar architectures with minimal effort to see the efficiency of text and image fusion models.

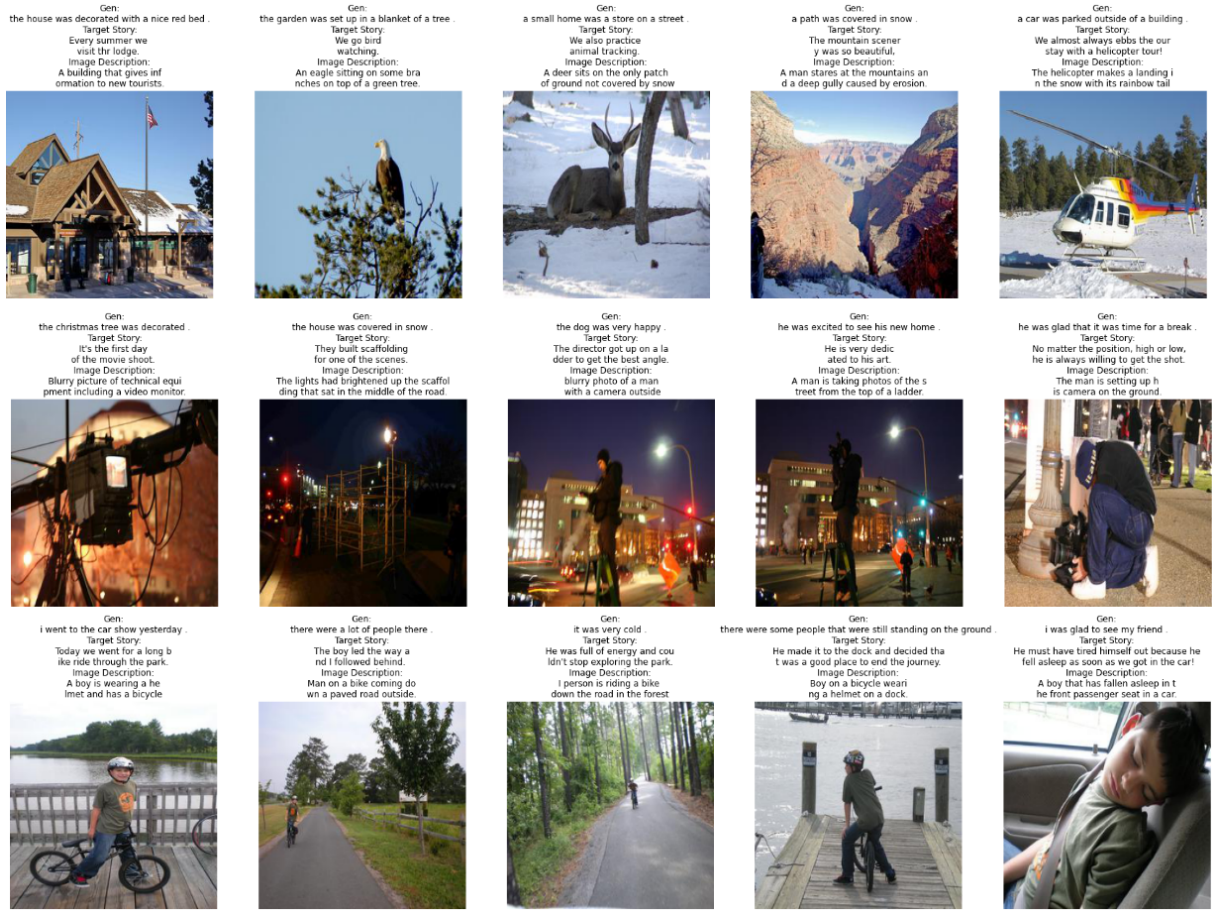


Figure 3: An example of the results

4.2 Qualitative Results

Figure 3 shows some of the test results to further evaluate the network. We conclude that even if the success of the model compared to baselines are not very distinct, the model learns some sentences that is likely to score well. For example, we see that the model tends to generate sentences like "the family was having fun." or "it was nice outside" more likely compared to other sentences. This shows that the model is not very good at generating diverse outputs. Another problem is that it does not condition well on the input. Any type of vehicle, e.g. a helicopter, a bicycle, a ship, tends to be defined as a "car". The both issues mentioned here of course might be caused by a bias in data as well. However, we can clearly state that in training these two issues should be definitely addressed to further improve the performance.

4.3 Discussion

Although the evaluation results show that the SOTA results are ahead of the proposed model in terms of METEOR scores and also in terms of qualitative analysis, it could be because of using less amount of data in training phase with low number of epochs due to limitation over time and computational resources. Another problem we want to mention is that the whole network is quite large and it might not be able to scale with this reduced data size. An ablation study to experiment with smaller networks is needed at this point.

Since the output of the qualitative experiments show that conditioning on the prompt is still quite weak and insufficient, fine-tuning sentence encoder is new research direction. The model is not very good at object detection at image prompts, FasterRCNN (Ren et al., 2015) is a good option as an image encoder to solve this problem as it is used in (Hsu et al., 2020).

One fair objection to this models would be its increased number of prompts. The reviewed literature

uses only image data hence making easier of the practical usage. We stand by this opinion but for this we suggest that experimenting with ready image captioning models. With the recent advances, image captioning tools are quite successful. We suggest that when the improvement in visual storytelling is potential, further promising, using off-the-shell image captioning networks as a side model would solve the need of descriptions.

The experiments should be further extended in the direction of modifying the SOTA architectures with the auxiliary textual data as it is proposed in this paper. We show its efficiency using one of the baselines, GLACNet, by presenting METEOR score of the network trained with half of the data. However, we leave modifying more architectures and using image captioning models instead of ground truth descriptions, to future work.

5 Conclusion

In this paper, we proposed a 2-stage visual storytelling model which uses image description as an auxiliary information source to generate context-aware short stories. The proposed model utilizes off-the-shell pretrained RESNet-152 network as image encoder which gives raw image features. To further enhance the single image features, the raw features are then fed to a biLSTM network which is an effective structure to leverage temporal dependencies between image sequences. For encoding images' description annotations we used again off-the-shell Sentence-BERT encoder. After obtaining image and textual features we combine them by concatenation and feed them into a 2-layer LSTM network which outputs 5-sentence short stories at the output. The proposed model is evaluated using one of the common automatic evaluation metric, METEOR, using the official code of the 2018 VIST challenge and results are presented in comparison with the baseline state-of-the-art methods. We show that using both textual and image data offers promising results and is an interesting path for visual storytelling community.

References

- Zainy M. Malakan Aljawy, Nayyer Aafaq, Ghulam Mubashar Hassan, and Ajmal Mian. 2021. Contextualise, attend, modulate and tell: Visual storytelling. In Giovanni Maria Farinella, Petia Radeva, Jose Braz, and Kadi Bouatouch, editors, *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, volume 5, pages 196–205. Scitepress, February. 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP ; Conference date: 08-02-2021 Through 10-02-2021.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.
- Chao-Chun Hsu, Zi-Yuan Chen, Chi-Yang Hsu, Chih-Chia Li, Tzu-Yuan Lin, Ting-Hao Kenneth Huang, and Lun-Wei Ku. 2020. Knowledge-enriched visual storytelling. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7952–7960. AAAI Press.
- Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California, June. Association for Computational Linguistics.
- Taehyeong Kim, Min-Oh Heo, Seonil Son, Kyoung-Wha Park, and Byoung-Tak Zhang. 2018. GLAC net: Global attention cascading networks for multi-image cued story generation. *CoRR*, abs/1805.10973.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Xin Wang, Wenhui Chen, Yuan-Fang Wang, and William Yang Wang. 2018. No metrics are perfect: Adversarial reward learning for visual storytelling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 899–909, Melbourne, Australia, July. Association for Computational Linguistics.