

Contextualize, Show and Tell: A Neural Visual Storyteller

Diana Gonzalez-Rico and Gibran Fuentes-Pineda

dianaglzrico@gmail.com and gibranfp@unam.mx

Institute for Research in Applied Mathematics and Systems (IIMAS)
Universidad Nacional Autónoma de México (UNAM)

Abstract

In this paper, we present a neural model for generating short stories from image sequences, which extends the image description model by Vinyals et al. (Vinyals et al., 2015). This extension relies on an encoder LSTM to compute a context vector of each story from the image sequence. This context vector is used as the first state of multiple independent decoder LSTMs, each of which generates the portion of the story corresponding to each image in the sequence by taking the image embedding as the first input. Our model showed competitive results with the METEOR metric and human ratings in the internal track of the Visual Storytelling Challenge 2018.

1 Introduction

Over the past few years, generating text from images and videos has gained a lot of attention in the Computer Vision and Natural Language Processing communities and several related tasks have been proposed, such as image labeling, image and video description and visual question answering. In particular, prominent results have been achieved in image description with various deep neural network architectures, e.g. (Lin et al., 2014), (Xu et al., 2015), (Karpathy and FeiFei, 2015), (Vinyals et al., 2015). However, the need of generating more narrative texts from images which may reflect experiences, rather than just listing objects and their attributes, has given rise to tasks such as visual storytelling (Huang et al., 2016). This task is about generating a story from a sequence of images. Figure 1 shows the difference between descriptions of images in isolation and stories for images in sequence.

In this paper, we describe the deep neural network architecture we used for the Visual Storytelling Challenge 2018. The problem to solve in this challenge can be stated as follows: *Given a*



Figure 1: Examples of stories for images in sequence (above) and image descriptions in isolation (below) from the VIST dataset (Huang et al., 2016).

sequence of 5 images, the system should output a story related to the content and events in the images. Our architecture is an extension of the image description architecture presented by (Vinyals et al., 2015). We submitted the generated stories to the internal track of the **Visual Storytelling (VIST)** Challenge, which were evaluated using the METEOR metric (Banerjee and Lavie, 2005) as well as human ratings.

2 Previous work

The work by (Park and Kim., 2015) presented probably the first system for generating stories from an album of images. This early approach involved the use of the NYC and Disney datasets mined from blog posts by the authors.

The visual storytelling task and dataset were introduced by (Huang et al., 2016). This was the first dataset specifically created for visual storytelling. They proposed a baseline approach which consists of a sequence to sequence model, where the encoder takes the sequence of images as input and the decoder takes the last state of the encoder as its first state to generate the story. Since this model produces stories with generic phrases, they

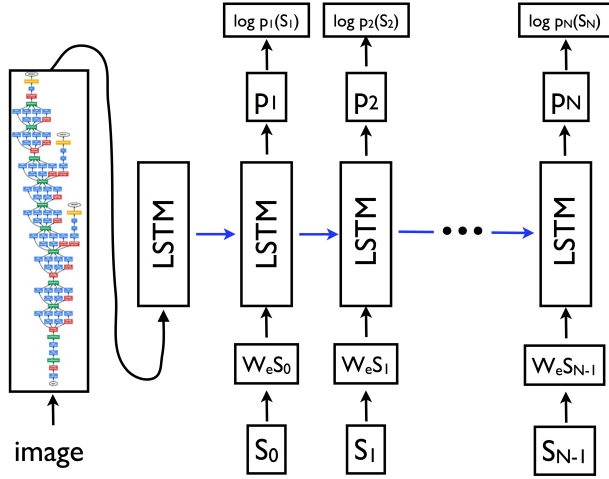


Figure 2: Show and Tell architecture. Image reproduced from (Vinyals et al., 2015).

used decode-time heuristics to improve the generated stories.

(Yu et al., 2017) presented a multi-task model that performs album summarization and story generation. Even though the model achieved state-of-the-art scores on the VIST dataset with different metrics, some of the sample stories presented in the paper are incoherent.

3 Model

Our model extends the image description model by (Vinyals et al., 2015), which consists of an encoder-decoder architecture. The encoder is a Convolutional Neural Network (CNN) and the decoder is a Long Short-Term Memory (LSTM) network, as presented in Figure 2. The image is passed through the encoder generating the image representation that is used by the decoder to know the content of the image and generate the description word by word. In the following, we describe how we extended this model for the visual storytelling task.

3.1 Encoder

The model’s first component is a Recurrent Neural Network (RNN), more precisely an LSTM that summarizes the sequence of images. At every timestep t the network takes as input an image I_i where $i \in \{1, 2, 3, 4, 5\}$ from the sequence. At time $t = 5$, the LSTM has encoded the 5 images and provides the sequence’s context through its last hidden state denoted by $h_e^{(t)}$. The representation of the images was obtained through Inception V3.

3.2 Decoder

The decoder is the second LSTM network that uses the information obtained from the encoder to generate the sequence’s story. The first input x_0 to the decoder is the image for which the text is being generated. The last hidden state from the encoder $h_e^{(t)}$ is used to initialize the first hidden state of the decoder $h_d^{(0)}$. With this strategy, we provide the decoder with the context of the whole sequence and the content of the current image (i.e. global and local information) to generate the corresponding text that will contribute to the overall story.

Our model contains five independent decoders, one for each image in the sequence. All the 5 decoders use the last hidden state of the encoder (i.e. the context) as its first hidden state and take the corresponding image embedding as its first input. In this way, the first decoder generates the sequence of words for the first image in the sequence, the second decoder for the second image in the sequence, and so on. This allows each decoder to learn a specific language model for each position of the sequence. For instance, the first decoder will learn the opening sentences of the story while the last decoder the closing sentences. The word embeddings were computed using word2vec (Mikolov et al., 2013).

Our proposed architecture is presented in Figure 3. For each image in the sequence, we obtain its representation $\{e(I_1), \dots, e(I_5)\}$ using Inception v3. The encoder takes the images in order, one at every timestep t . At time $t = 5$, we obtain the context vector through $h_e^{(t)}$ (represented by \mathbf{Z}). This vector is used to initialize each decoder’s hidden state while the first input to each decoder is its corresponding image embedding $e(I_i)$. Each decoder generates a sequence of words $\{p_1, \dots, p_n\}$ for each image in the sequence. The final story is the concatenation of the output of the 5 decoders.

4 Evaluation

4.1 Methodology

The generated stories were evaluated using both automatic metrics and human ratings. The automatic evaluation was performed by computing the METEOR metric (Banerjee and Lavie, 2005) on a public test set and a hidden test set. The former is a set of 1,938 image sequences and stories taken from the test set of the VIST dataset (Huang et al.,

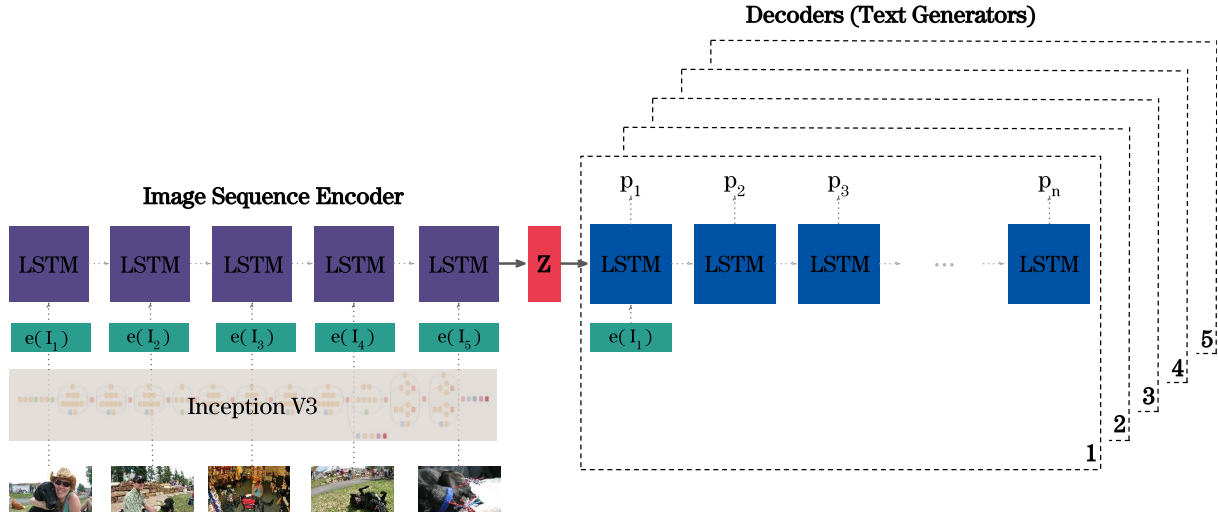


Figure 3: Proposed sequence to sequence architecture.

2016). The latter consists of new stories generated by humans from a subset of image sequences of the public test set.

Human ratings of the stories were collected from crowd workers in Amazon Mechanical Turk. Only 200 stories were selected from the hidden test set for this evaluation. The crowd workers evaluated each story on a Likert scale with respect to 6 aspects: **a)** the story is focused, **b)** the story has good structure and coherence, **c)** would you share this story, **d)** do you think this story was written by a human, **e)** the story is visually grounded and **f)** the story is detailed. The crowd workers were also asked to evaluate stories generated by humans for comparison purposes.

4.2 Results

Table 1 shows the METEOR scores by our model in the public and hidden test set of the Visual Storytelling Challenge 2018. Table 2 presents results of the human evaluation. Our model achieved competitive METEOR scores in both test sets and performed well in the human evaluation.

Public Test Set	Hidden Test Set
.3088	.3100

Table 1: Automatic evaluation of stories generated by our visual storyteller using the METEOR metric.

An evaluation over the complete VIST test set was also performed and the results are shown in Table 3¹.

¹We used the code available at github.com/lichengunc

Our model obtained the highest scores with the METEOR and BLEU-3 metrics but lagged behind the model by (Yu et al., 2017) with the ROUGE and CIDEr metrics.

Figure 4 shows some sample stories generated by our model from the public test set of the Visual Storytelling Challenge 2018. Although some of the generated stories are grammatically correct and coherent, they tend to contain repetitive phrases or ideas. We can also observe that some stories are not nearly related to the actual content of the images or include generic phrases like *This is a picture of a store.* These limitations of our model reflected on the ratings of the **visually grounded** and **detailed** aspects of the human evaluation.

5 Conclusions and Future Work

Our visual storyteller incorporates a context encoder and multiple independent decoders to the image description architecture by (Vinyals et al., 2015) to generate stories from image sequences. Having an independent decoder for each position of the image sequence, allowed our visual storyteller to build more specific language models using the context vector as its first state and the image embedding as its first input. In the internal track of the Visual Storytelling Challenge 2018, we obtained competitive METEOR scores in both the public and hidden test sets and performed well in the human evaluation.

In the future, we plan to explore the use of an attention mechanism or a bidirectional LSTM to

	a)	b)	c)	d)	e)	f)	Total score
Ours	3.347	3.278	2.871	3.222	2.886	2.893	18.498
Human	4.025	3.975	3.772	4.003	3.965	3.857	23.596

Table 2: Human evaluation of stories generated by our visual storyteller, compared to stories generated by humans.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr
Huang et al.	-	-	-	-	31.4	-	-
Yu et al.	-	-	21.0	-	34.1	29.5	7.5
Ours	60.1	36.5	21.1	12.7	34.4	29.2	5.1

Table 3: Automatic evaluation on the VIST dataset. A comparison between the baseline (Huang et al., 2016), (Yu et al., 2017) and ours.



cope with repetitive phrases within the same story.






6 Acknowledgments

This research is supported by the PAPIIT-UNAM research grant IA104016. Diana González Rico is supported by CONACYT.

References

- Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.
- Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, and Dhruv Batra et al. 2016. Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*.
- Andrej Karpathy and Li FeiFei. 2015. Deep visual-semantic alignments for generating image descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 740755.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision - ECCV*, pages 740–755.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.
- Cesc C Park and Gunhee Kim. 2015. Expressing an image stream with a sequence of natural sentences. *Advances in Neural Information Processing Systems*, pages 73–81.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: a neural image caption generator. In *Computer Vision and Pattern Recognition*.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*.
- Licheng Yu, Mohit Bansal, and Tamara Berg. 2017. Hierarchically-attentive rnn for album summarization and storytelling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 966–971.

					
Ours	Today was graduation day.	The students were excited.	My parents were so happy.	He was very happy to be graduating.	Everyone was so proud of him.
Human	Today was graduation, and Schyler was extremely happy.	However he was nervous about what the future would bring.	His parents assured him that he would do well in life. That helped a little.	Of course when Benny the Squirrel gave him life advice his whole demeanor turned happily.	Schyler is now ready for life, after the first big chapter ending high school.

					
Ours	We went to a fancy restaurant.	I made a lot of food for everyone.	The food was delicious.	The main course was presented beautifully.	The dessert was even better.
Human	Five Star restaurants are cool.	The only problem is that sometimes style takes the place of quantity.	And you get left with a disappointing plate.	But that doesn't happen if you know what to order.	And then you can think about desert.

					
Ours	I went to the art gallery yesterday.	There were a lot of people there.	There was also a fortune teller.	This is a picture of a store.	I had a great time.
Human	Almost every dad likes beer.	Father's Day is a great day for Dads.	There was free beer for dad!	The pub took reservations early.	Lots of people took their fathers out for free beer.

Figure 4: Sample stories generated by our visual storyteller, compared to generated by humans.