Children Stories Generator From Hand Drawings

Suhong Kim
Department of Computing Science
Simon Fraser University
Burnaby, BC, Canada
suhongk@sfu.ca

Mohammad Mazraeh
Department of Computing Science
Simon Fraser University
Burnaby, BC, Canada
mmazraeh@sfu.ca

Sara Jalili
Department of Computing Science
Simon Fraser University
Burnaby, BC, Canada
sja100@sfu.ca

Abstract

Children sometimes express their feelings and thoughts through their drawings. Adults can make children stories for them but it would be awesome if we could capture their drawings and interpret them. For this purpose, we designed the children stories generator based on the latest model for generating multi-sentences stories called "Neural Storyteller". In the approach section, we introduce our designed model, and describe the main implementation concepts such as Image Captioning and Skip-Thought model. Next, we show the testing outputs of our experiments along with our conclusion.

1 Introduction

In the past few years, several state-of-the-art models have been released for image captioning. However, for generating stories - specially for a specific style - more tools are required. The latest released model for generating multi-sentences stories is neural-storyteller[1, 2]. This model is trained on MS COCO dataset; it gets an image as an input, extracts the features with a CNN, produces several captions using an RNN and choose the most relevant one to generate a romance style story using the following methods: Image-sentence embeddings, Skip-thought vectors, conditional neural language models, and style shifting.

In this project we tried to get the use of this model to design a pipleine to generate children style stories, not from real images, but from children's raw drawings. For this purpose, the following actions were required:

- Converting real images to sketch form using Pencil Sketch filter.
- Training the model to generate children style stories.

In the following sections, all the mentioned topics above will be explained and further experiments will be discussed.

2 Approach

In this project, we designed the below pipeline based on the neural story teller model [1]. This model requires two training stages: Image Captioning (Encoder) and Story Telling (Decoder).

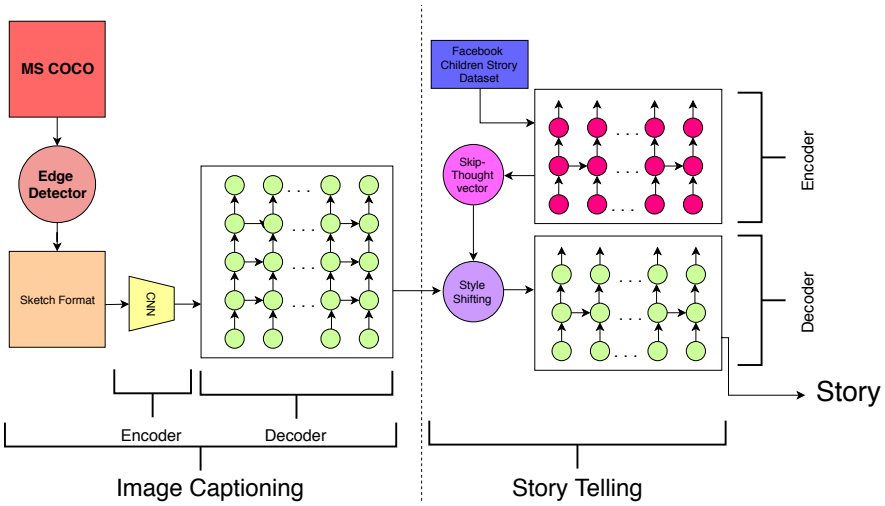


Figure 1: The proposed model structure for children stories generator

2.1 Image Captioning

The first part of the model is used to generate image-sentence embedding, which is the output of the CNN + RNN structure. The basic form of this model follows the same concept of the Show and Tell model [3], which introduced the following formula,

$$\theta^* = \arg\max_{\theta} \sum_{(I,S)} \log p(S|I;\theta)$$

where the θ is the parameters, I is an image input, and S is a ground truth sentence input. By applying the chain rule, the joint probability over the previous sentences can be represented as follow,

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \ldots, S_{t-1})$$

where N is the length of the sample at that point. To optimize the sum of the log probabilities, we applied the Convolutional Neural Network (VGG19) for the representation of an input image, and Recurrent Neural Network based on GRU for the long-term memory of the previous sentences [4]. Also, we used the Microsoft COCO dataset in order to obtain a pair of image and sentence as an input. Since our application requires hand drawing inputs, the online data augmentation method was used with the edge detector and Sketch-form filter while training.

2.2 Story Telling

The second part of the model in the figure 1 requires the pre-training only on the text dataset to learn the style of the writings. Based on the original paper[1], we applied the Skip-Thought model, which learns the fixed length representations of sentences in any Natural Language without any labeled data or supervised learning. The only supervision or training signal that Skip-Thoughts uses is the ordering of sentences in a natural language corpus. To generate the relevant stories given an image input, this model is composed of three parts: Skip-Thought model (Encoder), Style Shifting, and Conditional Neural Language model (Decoder).

2.2.1 Skip-Thought Model (Encoder)

In the encoder-decoder framework of skip-thoughts model, the encoder is responsible to map words into a sentence vector and the decoder is used to generate the surrounding context. In the Encoder of the Skip-Thought Model, the hidden states encode the full sentence s_i which can be represented by its words $w_i^1, w_i^2, ..., w_i^N$. The decoder is a neural language model which conditions on the encoder output h_i. At the training stage, it uses two decoders; one for the next sentence s_{i+1} and the other for the previous sentence s_{i-1}. Also, those decoders has three bias matrices C_z, C_r and C for the update gate, the reset gate and the hidden state respectively. All the parameters for two decoders are learned separately with an exception of vocabulary matrix V which is the weight matrix connecting the decoder's hidden state for computing a distribution over words. Considering the (s_{i_1}, s_i, s_{i+1}), the objective function is sum of the log-probabilities for the forward and backward sentences conditioned on the encoder representation [1],

$$\sum_t logP(w_{i+1}^t | w_{i+1}^{<t}, h_i) + \sum_t logP(w_{i-1}^t | w_{i-1}^{<t}, h_i)$$

After training, we only keep the Encoder part for getting the skip-thought vectors for the Style Shifting.

2.2.2 Style Shifting

With Skip-Thought Vector's properties, the style shifting function F can be defined as below,

$$F(x) = x - c + b$$

Where x is image caption vector, c is "caption style" vector, and a b is "book style" vector. Both caption style and book style vectors are pre-trained values with GRU encoder. This function can be interpreted as we want to keep the main idea from the image caption vector x but convert the style of writing to that of a book. Thus, this style shifting can be a role of bridge between two datasets. The style vector for a dataset is simply the mean of all Skip-thought vectors for all sentences in that dataset.

2.2.3 Conditional Neural Language Model (Decoder)

Since one of the two decoders in Skip thought model can generate the related sentences given an input sentence, we can use the one of two decoders as our neural language model. By concatenating the generated sentence to the previous results, this decoder can write one paragraph of stories given an image and style vector.

2.2.4 Vocabulary Expansion

The vocabulary expansion is a technique used in [1] to reduce the size of vocabulary dictionary. Using this technique allows to use the words which do not exists in the dictionary. Let's denote the big dictionary with D_{big} and the smaller dictionary as D_{small}. Main task of linear vocabulary expansion is to learn a matrix W which minimizes the error of $V_{D_{small}} = WV_{D_{big}}$.

3 Experiments

3.1 Generating Hand-drawing image captioning dataset

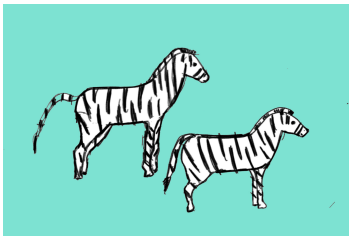
Due to the lack of the dataset for hand drawing images, we came up with the idea to convert the MS COCO dataset images to sketches using edge detector and the Pencil Sketch filters, utilize them for training and then have our hand drawing images for testing. Before training the model, we tested it with sketches and our drawing to see if it is already able to give reasonable results. Below are some tested images. As shown, the original model is somehow able to detect some sketches and hand drawings, but it gave poor results for the most parts.

Two zebras back to back grazing on plants in middle of desert.

The poster has the picture of a man with words around it.

The cake is prepared and ready to be eaten.

Figure 2: Generated captions for sketch version of the real images



A group of zebras walking to the right of the image .

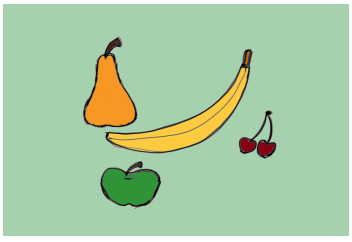
A large picture of a evil dictator on the side of a wall .

Two apples with faced painted on them one impaled by a knife .

Figure 3: Generated captions for drawing version of the real images

Various clocks and a picture frame hanging on a tree .

Someone has been writing on the wall above the clock .



A collage with several artistic pictures on it .

Figure 4: wrong generated captions

3.2 Generating Children style stories

In this experiment, we tried to train the story-telling part of the model on a children story dataset published by Facebook [5] which contains 1.6GB of children stories. This dataset requires some pre-processing techniques such as removing non ASCII characters before training. We tried to train this model on the whole dataset, but it was not possible to complete during this project period due to some computational limitations. First, the original model required more than 2 weeks training as well as huge memory consumption during computation. Also, the environment setting for the neural storyteller model had a lot of dependency errors whenever we tried to modify some parts of the model. Although we tried to train the model on an small portion of the dataset, it didn't yield good results in terms of an story in children language.

Original: No man was looking up at her , taking in a deep breath . I have no idea what he meant to me , but I didn't know what else to do . In the past twenty-four hours I had become a woman in the office building , her cell phone number on the screen . It was one of the most important members of the club , and it gave him a sense that he would be able to tear his eyes open and disappear . In fact , Elizabeth 's face was a mask of steel . She listened intently , giving him the last thing I d heard in his life .

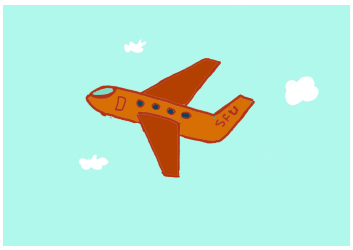
Sketch: He was finally showing a portrait of the man , and I winced at his words . For a brief moment , I tried to imagine what it would be like to make contact with Jesse , but he didn't look at me . I had no clue what she was going to do . In fact , she was pretty much the only other woman in the world . Her hair fell in a loose knot at the nape of her neck , and I held out my ID as well . I spent so much time searching for an image of Cole , the one who had to deny it .

Drawing: I had the picture taken on a wall of mirrors , and it seemed to be the most important thing in the world . Of course , I didn't know what to say . In spite of the fact that I had no idea what she was doing , I let my eyes go open and closed the door behind me . I took a deep breath and tried to recall the details of Gabriel 's murder , but he was also an artist . For a thousand years , I would have fallen asleep on either side of him . She looked up at the ceiling , a new reminder that I was bound to lose and deal with it .

Figure 5: output stories for all different versions of the same image

Original: Images commercial jet flew off the plane , and I gasped for air . By the time I reached the top of the plane , my heart was pounding so fast that I didn't want to leave . As far as he was concerned , he had to be the only female in the world to rescue her . In fact , I had never been more comfortable with the news . I felt like I could fly out of the sky just as long as the sun rose above us , sending flying waves crashing down into the sky . Like it s mine , Dr. Z and I were flying the full flight .

Sketch: We were flying rapidly to the breaking point , flying in the sky . Of course , I heard more of them as we stared at each other . The air in the sky seemed to be the most exciting thing on Earth , but I had no idea how long it was for me to stay alive . In fact , I felt as if he were the most precious thing in the world . Alyssa 's breath caught in her ear , and she was tired of flying planes and flying planes . I gave those orders , the speed coming out more quickly than ever before .



Drawing: We were flying to the bottom of the plane , and I gasped in surprise . As much as I hated the feeling , it had nothing to do with the power flowing from me , but I had no idea how long it would be for us to remain silent until the sun came up . The air in the air made me feel more and more powerful . By the time he pulled her into his arms , I couldn't help but watch . In fact , I was just pissed off that Dr. Whitney 's team had captured her in a very short period of time .

Figure 6: output stories for all different versions of the same image

4 Conclusion

We proposed our creative idea for the application of the Image Captioning. We designed our model in Figure 1 based on the previous work[1] and suggested the way to overcome the problems of getting proper dataset using data augmentation. Also, we tried to train to obtain better results from the new story dataset.

However, our base model has several limitations such as the loose connection between image and story, and unnatural English sentences. Thus, based on the recent paper "Semstyle" [6], we believe that we can improve the accuracy and performance of the model with the following implementation,

- Extract more relevant features from images with applying Inception V3 model on CNN.
- Generate more human style sentences with using attention based GRU.

Contributions

- **Suhong Kim** : Research, Experiment, Preparing and editing the poster and report
- **Mohammad Mazraeh** : Research, Experiment, Enviroment setting, Preparing resources
- **Sara Jalili** : Research, Experiment, Preparing and editing the poster and report

References

[1] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in *Advances in neural information processing systems*, pp. 3294–3302, 2015.

[2] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE international conference on computer vision*, pp. 19–27, 2015.

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

[4] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *International Conference on Machine Learning*, pp. 595–603, 2014.

[5] F. Hill, A. Bordes, S. Chopra, and J. Weston, "The goldilocks principle: Reading children's books with explicit memory representations," *arXiv preprint arXiv:1511.02301*, 2015.

[6] A. Mathews, L. Xie, and X. He, "Semstyle: Learning to generate stylised image captions using unaligned text," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8591–8600, 2018.