# Children Stories Generator From Hand Drawings

Suhong Kim, Mohammad Mazraeh, Sara Jalili
Department of Computing Science, Simon Fraser University

## Introduction

Children sometimes express their feelings and thoughts through their drawings. Adults can make children stories for them but it would be awesome if we can capture their drawings and interpret their own stories. For this purpose, we suggest the children stories generator from hand drawings using the convolutional neural network (CNN) and recurrent neural network (RNN).

## Main Objectives

Based on the neural-story teller model, we did experiments as follow:

- Converting real images to sketch form using edge detection
- Training the model to generate children stories.
- Research extension on the recent work.

## Methods

This model requires two training stages: Image Captioning (Encoder) and Story Generation (Decoder). We trained CNN + RNN model on Microsoft COCO datasets to obtain standard image captions. Then, We fed them into the GRU Network Decoder which was trained on Facebook Children Stories. The Style Shifting is used to fill the gap between the caption style and children story style by matching each paragraph with a Skip-Thought Vector.

**Image-Sentence Embedding:**
Image features are extracted from the CNN network and projected into the GRU hidden states inside RNN network. With stochastic gradient method, we optimized our model to generate image-sentence embedding from the vocabulary of Microsoft COCO image captions.

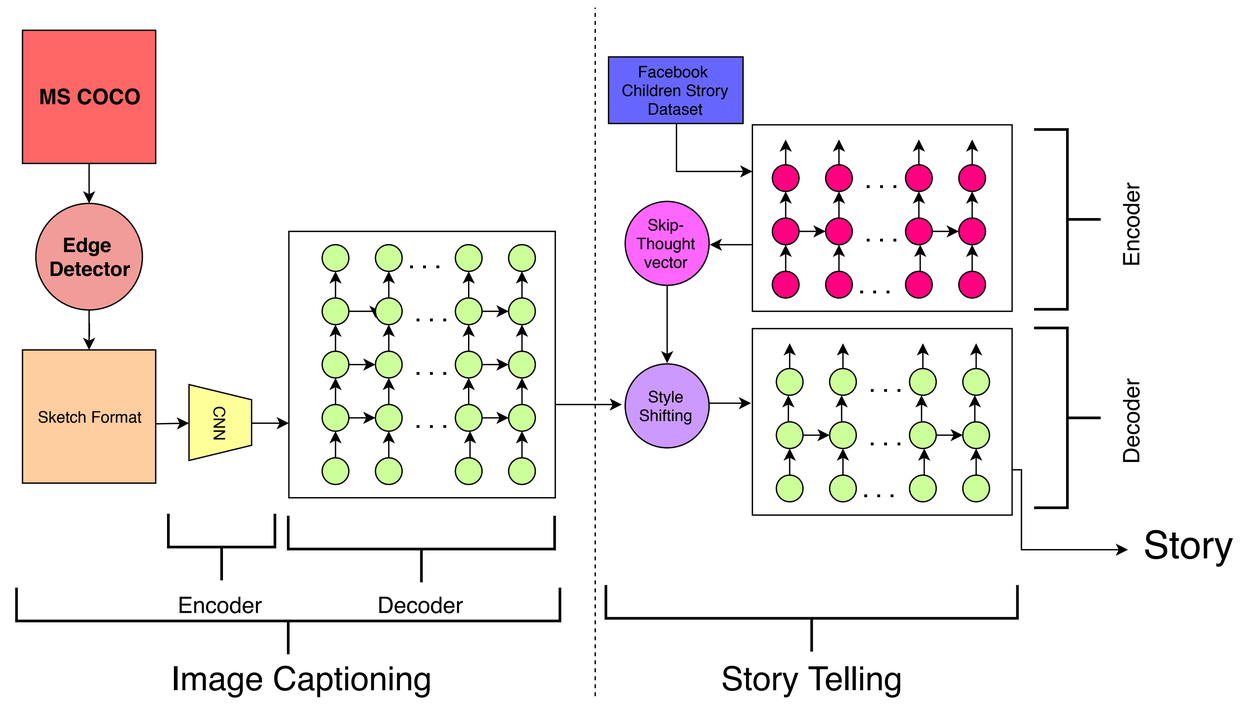
**Skip-Thought Vectors:**
Skip-thought vectors is a work inspired by word2vec. Word2vec learns a vector space in which words with the same meaning are close to each other. Skip-thoughts is a model for learning fixed length representations of sentences in any Natural Language without any labeled data or supervised learning. The only supervision/training signal Skip-Thoughts uses is the ordering of sentences in a natural language corpus.

**Style Shifting:**
With Skip-Thought Vectorś 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. 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.

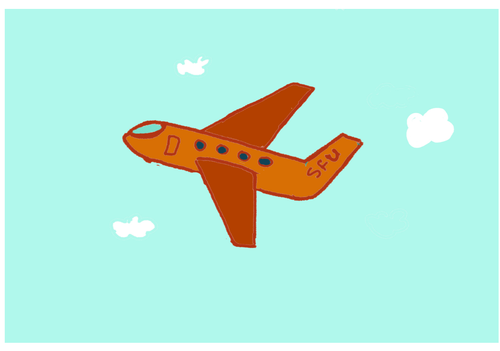
## Model Structure




## Results

**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 did n'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:** I could barely contain the flash of blue in my eyes , and for the first time in Rome , I felt as if she were dead . The truth was , I had no idea what to do about us . It was also the most beautiful thing I 'd ever seen in my life . In the background of the plane , he pulled out a pair of tattered jeans and a pair of black cargo pants that hung from the ceiling and sent them flying around us . It seemed like an endless stretch of time , and I refused to let go . I had no clue what happened to me , the most powerful and painful memory .

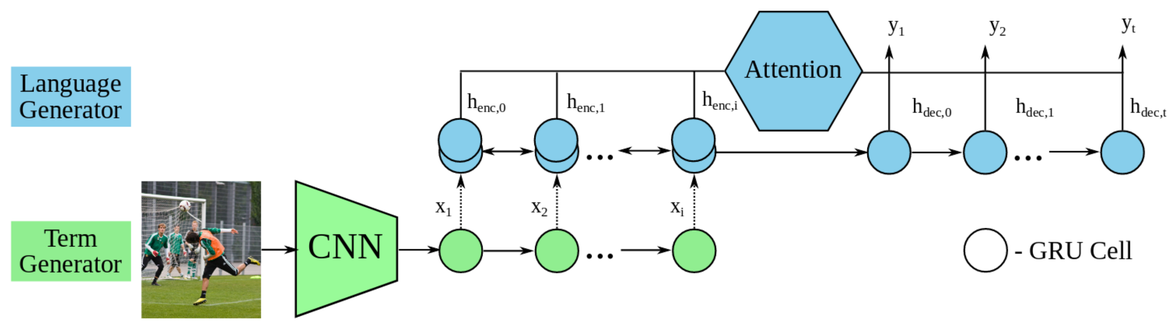


**Drawing:** We were flying to the bottom of the plane , and I gasped in surprise . As much as I hated the feeling , it had nothing to do with the power flowing from me , but I had no idea how long it would be for us to remain silent until the sun came up . The air in the air made me feel more and more powerful . By the time he pulled her into his arms , I could n'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 .

## Discussion

Our base model has several limitations such as the loose connection between image and story, and unnatural English sentences. Thus, our future work includes below

- Extract more relevant features from images with applying Inception V3 model on CNN.
- Generate more human style sentences with using attention based GRU, which is based on the recent work "Semstyle"[2]




## References

[1] Kiros R, Zhu Y, Salakhutdinov RR, Zemel R, Urtasun R, Torralba A, Fidler S. *Skip-thought vectors. InAdvances in neural information processing systems* 2015 (pp. 3294-3302).

[2] Mathews A, Xie L, He X. *SemStyle: Learning to Generate Stylised Image Captions using Unaligned Text.* InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018 May 18 (pp. 8591-8600).