CSCI 544: Group 46 Project Proposal
ImagiNarrate: Building a Narrative with Images and Generated Captions

Asmita Chotani , **Apurva Gupta** , **Chetan Chaku** , **Rutuja Oza** and **Priya Nayak**
University of Southern California
[achotani, apurvagu, chaku, rgoza, psnayak] @usc.edu

Abstract

In this paper, we introduce a new natural language processing (NLP) approach to solve the problem of visual storytelling that utilizes image features to generate captions and subsequently develop a coherent story line for the images. By incorporating image features in the caption generation process, our proposed approach aims to provide a more relevant and informative description of the images that can be used to build a cohesive and engaging narrative. We evaluate our model in comparison to the AREL model (Wang et al., 2018) used for story generation on the basis of traditional metrics like Meteor and Bleu as well as human evaluation of the generated stories.

1 Project Domain & Goal

Storytelling is one of the few human capabilities that can be considered complicated for a machine. Humans have the ability to add emotions as well as imagination when creating a story. In an attempt to work on storytelling capabilities using NLP, we aim to propose a model to produce a visual narrative using a collection of images and captions generated from those images.

To create a compelling story, it is necessary to use techniques from both image processing and natural language processing which goes beyond the coursework. We plan to use the generated captions to build up additional context for a story based on the series of images. Comparing our results with stories created only using images, we will be checking the importance and usefulness of the augmentation of captions and images for storytelling.

One of the largest impacts of this technique if successful would be creating novels from comic books automatically. Visually impaired people can not access comics in the usual way others do and translating comics to novels is a hard and time consuming manual task. Using our model these comics can directly be converted into textual stories which are easy to convert to braille, making them accessible to the visually impaired.

2 Related Work

(Ting-Hao et al., 2016) introduced the first dataset for sequential vision-to-language tasks and provided the first baselines for the task of visual storytelling on their proposed dataset - VIST in 2016. Their model is trained using GRUs for image sequence encoder and the story decoder. They also motivate the use of automatic metrics like Meteor for evaluation of the generated stories by providing a comparative analysis against human judgment.

(Gonzalez-Rico and Fuentes-Pineda, 2018) proposed the neural vision storyteller which extends the image description model by (Vinyals et al., 2014). The original model consisted of CNN encoder and LSTM decoder. The newly proposed model uses 5 different decoders one for each image in the sequence and then consolidates the output from the decoders by concatenation.

(Wang et al., 2018) proposed the AREL model in 2018 which was based on Reinforcement learning and produced SOTA results on Meteor scores for Visual Storytelling. The paper uses an encoder/decoder based policy model which is trained with the help of another CNN based reward model.

Further, (Kim et al., 2018) proposed GLAC Net, that generates visual stories by combining global-local (glocal) attention and context cascading mechanisms. Some knowledge graph (Hsu et al., 2019) and common-sense (Chen et al., 2021) based models were also introduced over the years, but weren't able to provide any significant improvement for the task at hand based on traditional metrics (Wang et al., 2022).

(Wang et al., 2022) performed a comparative study of some of the above discussed models using traditional automatic metrics like Meteor and Bleu but came to the conclusion that they don't correlate enough with human evaluation and also proposed their own metric to overcome this hurdle.

3 Datasets

The project uses the MSCOCO dataset(Lin et al., 2014) for training the image to caption model and the VIST dataset(Ting-Hao et al., 2016) for the storytelling task. Both the datasets are commonly used benchmarks in the field of CV & NLP. These datasets have been used in different use cases to generate state-of-art models. Another advantage of these datasets is the large-scale, high-quality annotations that they provide.

MSCOCO consists of data related to large-scale object detection, segmentation, key-point detection, and captioning. The dataset contains a total of 328K data points annotated with captions (msc).

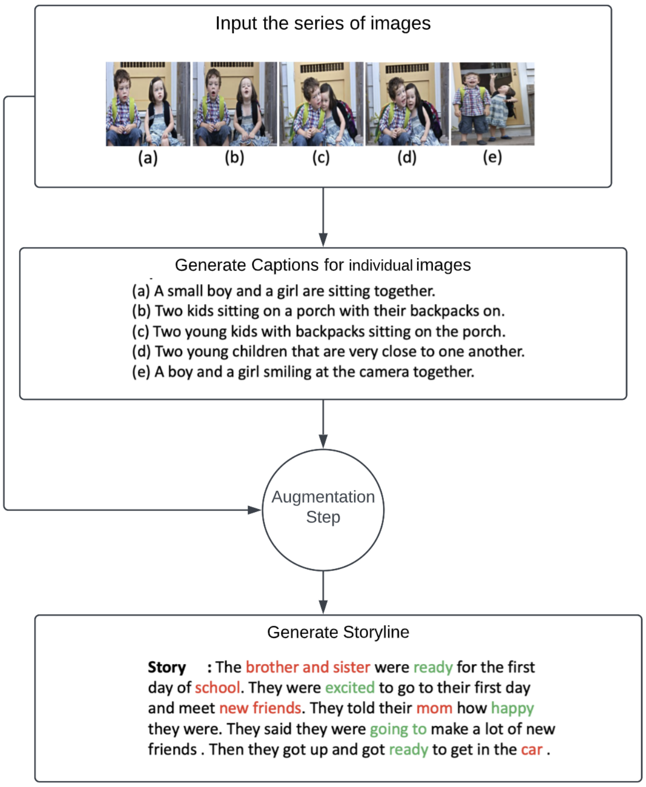
VIST is specifically for sequential vision-to-language tasks which makes it perfect for our use case. A single VIST sample consists of an image sequence paired with description for individual images and description of a human-constructed narrative over the sequence. The dataset consists of a total of 10,117 Flickr albums and 210,819 unique images (vis).

While the images would be preprocessed using techniques like image pixel standardization, and tokenization to reduce the noise in the data, all the textual data would be preprocessed using cleaning, tokenization, removal of stop words, lemmatization & vectorization.

4 Technical Challenge

Generating correct captions for images is one of the project's technical challenges. The NLP model must interpret the context and content of the image in order to produce correct and expressive captions for the images. This requires the model to extract the image's most important features and efficiently learn the links between the two modalities and demands a strong understanding of both image and text processing. Another challenging task is constructing a cohesive narrative from the order of image captions and the images. The model must understand these relationships and provide a logical and interesting story.

While there are several research papers that focus on image captioning and story generation, few of them have specifically focused on the combination of the two tasks to create visual storytelling. We propose to create an architecture consisting of 2 models, one model will take images as input and generate captions, while the other model will take both the images and caption embedding and generate a coherent story from them.


Input the series of images
(a)
(b)
(c)
(d)
(e)
Generate Captions for individual images
Augmentation Step
Generate Storyline

(a) A small boy and a girl are sitting together.
(b) Two kids sitting on a porch with their backpacks on.
(c) Two young kids with backpacks sitting on the porch.
(d) Two young children that are very close to one another.
(e) A boy and a girl smiling at the camera together.

Story : The brother and sister were ready for the first day of school. They were excited to go to their first day and meet new friends. They told their mom how happy they were. They said they were going to make a lot of new friends . Then they got up and got ready to get in the car .

Figure 1: Proposed Model Architecture
(images and data from (Wang et al., 2018))

Traditional metrics like BLEU and METEOR have limitations in evaluating the semantic characteristics of images and the overall coherence of a story. They rely on string matching and may not be strongly correlated with human evaluations. Some models may perform well on these metrics but lose coherence and semantic correctness in the story. (Wang et al., 2022)

Since, automatic metrics do not help in evaluating the performance of the story model model, to evaluate the coherence of the generated story, we can use human evaluators to assess the story's quality. Human evaluators can provide qualitative feedback on the overall coherence and engagement of the story, as well as identify any grammatical or logical errors in the text.

To evaluate our model, we will be comparing the results of our model with the AREL model which only uses the image to generate the stories on the basis of the above metrics.

5 Division of Labor

- Rutuja Oza - Data Pre-Processing
- Priya Nayak - Caption Generation model from images
- Chetan Chaku & Apurva Gupta - Story Generation model from images and captions
- Asmita Chotani - Model Evaluation

References

Mscoco dataset.

Vist dataset.

Hong Chen, Yifei Huang, Hiroya Takamura, and Hideki Nakayama. 2021. Commonsense knowledge aware concept selection for diverse and informative visual storytelling. arXiv.

Diana Gonzalez-Rico and Gibran Fuentes-Pineda. 2018. Contextualize, show and tell: A neural visual storyteller. arXiv.

Ansar Hani, Najiba Tagougui, and Monji Kherallah. 2019. Image caption generation using a deep architecture. pages 246–251.

Chao-Chun Hsu, Zi-Yuan Chen, Chi-Yang Hsu, Chih-Chia Li, Tzu-Yuan Lin, Ting-Hao 'Kenneth' Huang, and Lun-Wei Ku. 2019. Knowledge-enriched visual storytelling. arXiv.

Chi-Yang Hsu, Yun-Wei Chu, Ting-Hao 'Kenneth' Huang, and Lun-Wei Ku. 2021. Plot and rework: Modeling storylines for visual storytelling. arXiv.

Taehyeong Kim, Min-Oh Heo, Seonil Son, Kyoung-Wha Park, and Byoung-Tak Zhang. 2018. Glac net: Glocal attention cascading networks for multi-image cued story generation. arXiv.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft coco: Common objects in context. arXiv.

Ting-Hao, 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. arXiv.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator. arXiv.

Eileen Wang, Caren Han, and Josiah Poon. 2022. Rovist:learning robust metrics for visual storytelling. arXiv.

Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. 2018. No metrics are perfect: Adversarial reward learning for visual storytelling. arXiv.

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.