Visual Story Telling: Generating Stories from images using Transformers

Sanjana S. Khan
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
sanjana.sabah.khan@g.bracu.ac.bd

Mehreen Khan
Computer Science
Brac University
Dhaka, Bangladesh
mehreen.khan@g.bracu.ac.bd

Annajiat Alim Rasel
Senior Lecturer
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
annajiat@bracu.ac.bd

Humaion Kabir Mehedi
Research Assistant
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
humaion.kabir.mehedi@g.bracu.ac.bd

Abstract

This research paper presents a novel approach for generating stories from images using transformer-based language models. The proposed method leverages the ability of transformers to capture complex dependencies in sequential data and effectively integrate visual and textual information. We evaluate the performance of the proposed approach on a popular benchmark dataset and demonstrate its effectiveness in generating engaging and coherent stories that are grounded in the provided images. Our research shows how certain models and approaches outperform certain state-of-the-art methods in terms of various evaluation metrics, including fluency, coherence, and engagement. Additionally, we provide an in-depth analysis of the generated stories to gain insights into the underlying mechanisms of the model. The findings of this study have potential implications for a wide range of applications, including content generation, image captioning, and visual storytelling.

Keywords: Visual storytelling, VIST, Transformer, Generative Image-to-text Transformer (GIT), Vision Transformer (ViT)

1 Introduction

Storytelling has been a vital part of human communication for centuries, serving as a means of conveying experiences, ideas, and emotions. In recent years, with the proliferation of digital media, there has been a growing interest in leveraging artificial intelligence (AI) techniques to automate the process of storytelling. One particular area of interest is visual storytelling, which involves generating a narrative based on a given set of images. This task is challenging because it requires the integration of visual and textual information and the generation of a coherent and engaging story. Recently, transformer-based language models have achieved state-of-the-art results on a wide range of natural language processing tasks, including language translation, language modeling, and text summarization. These models have the ability to capture long-range dependencies and effectively integrate visual and textual information. In this research paper, we study a novel approach for generating stories from images using transformer-based language models. The method leverages the ability of transformers to capture complex dependencies in sequential data and effectively integrate visual and textual information. We evaluate the performance of said approach on a popular benchmark dataset and demonstrate its effectiveness in generating engaging and coherent stories that are grounded in the provided images.

2 Literature Review

With the help of concrete description modeling and figurative and social language provided in the sampled dataset and storytelling task, it is possible to utilize artificial intelligence from the basic understandings of natural visual scenes and move towards a much more human-like understanding of a grounded event structure as well as subjective expression (Huang et al., 2016). This paper introduced a new dataset for sequential vision-to-language called SIND. The data collection was done through generating a list of storyable event types. The paper managed to keep possessive phrases heads if they could be classified as an EVENT in the WordNet 3.o while relying on manual winnowing for targeting collection efforts. The terms were used for collecting albums using the Flickr API. The paper also crowdsourced photos as part of its data collection efforts. Using automatic evaluation metrics, baseline experiments were performed using RNN and GRU. Several decode-time heuristics were explored which saw the improvement of the METEOR score. This approach shows that temporal context and narrative language facilitates the direct modeling of literal and abstract visual concept relationships.

Incorporation of context encoder along with a variety of independent decoders to existing image description architectures and story generation from sequences of images is also not unheard of (Gonzalez-Rico and Fuentes-Pineda, 2018). The utilization of an independent decoder for each separate position of the image sequence allowed the modified storyteller for more specific language model building, where the first state used the context vector while the first input was set as the image embedding. Competitive results were achieved using the METEOR metric. Utilizing the VIST dataset, it was observed that the model achieved the highest scores using the METEOR and BLEU-3 metrics.

Generating story from images is also a concept that has been tinkered with in the past (Min et al., 2021). The suggested framework recommends an AI-based story writer where the model uses Natural Language Processing (NLP) and an encoder-decoder structure for generating a variety of stories under different genres. The research that suggests this framework worked with two distinct datasets – books downloaded from Smashwords website, where the books contained over 20,000 words and were crawled for noise reduction and too short stories, and the conceptual captions dataset from Google which contains over 3.3 million pairs of images and captions generated automatically through filtering and caption annotation extraction from the internet. An NLP and encoder-decoder based model was then utilized to generate stories under a wide range of genres. The stories retrieved in this manner were semantically understandable and were also found to be based on the general content as per the images.

In recent times, vision transformer-based models for interpreting image sets as stories have also been explored. (Malakan et al., 2022). Such a framework is robust in the sense that it uses a sequence encoder receiving multi-view patches of images retrieved from a vision transformer (ViT) as the input to a bidirectional-LSTM, and then using a decoder fitted with a standard LSTM further enhanced using a mogrifier-LSTM. Using the Visual Story-Telling dataset (VIST) dataset, it was observed that this proposed model was able to achieve better scores and outputs compared to existing HGS and CAMT models.

The use of a generative image-to-text transformer for mapping input images and then associating them with subsequent text descriptions on large-scale image and text pairs is another avenue that has been explored (Wang et al., 2022). Fine-tuning the proposed GIT model, it was found that the model achieved a score of 92.9, which is more than the existing score of 91.9 tested on similar standard MJ+ST datasets.

Extensive ablation studies were further conducted on various datasets to understand the impact of a variety of parameters (Strudel et al., 2021). It was observed that the performance was better suited when large models and small patch sizes were included in the parameters. The suggested segmenter succeeded in achieving excellent results for semantic segmentation while outperforming previous studies on similar datasets. The application of a simple point-wise linear decoder to the patch encodings helped yield excellent results. For further improvement on this, using a mask transformer for decoding seemed to have worked wonders. The overall research was a stepping stone towards a more unified approach for semantic, instance, and panoptic segmentations. The segmenter-based model was trained end-to-end using a per-pixel cross-entropy loss. During the inference stage, argmax was applied to obtain a single class per pixel after upsampling. The ADE20K, Pascal Context, and Cityscapes datasets were utilized in this research. Comparison with state of the art shows that for both ADE20K and Pascal Context datasets, improvements were noticed on all avenues. For the Cityscapes dataset, while a strong improvement was not reported, it was observed that the proposed model is able to competitively match the performance levels of existing state-of-the-art methods.

Research has also been conducted on the aspects of using visual and narrative components for image-based storytelling (Smilevski et al., 2018). It was observed that through the use of an image sequence encoder, it is possible to capture temporal dependencies between the image sequence and the sentence-story. Further improvement on this was made possible by the use of a previous sentence-story encoder. This improvement meant that it was possible to achieve a much better story flow. The final result of all this was stories that were generated with better story flow and length compared to previous experimentations. This meant that it was possible to generate long human-like storytelling which was capable of describing the images provided as well as add narrative and evaluative content or language to it.

There has also been research done to look into the use of an adversarial reward learning (AREL) framework for learning an implicit reward function which is then used for optimizing policy search (Wang et al., 2018). Human evaluation conducted on said approach reveals that it achieved significant improvement in generating better human-like stories when compared to existing state-of-the-art methods. When looking at the output sentences individually, it is evident that the results produced by this method make more grammatical and semantic sense. Connecting the sentences together, it was further observed that the AREL story feels more coherent and is capable of describing the photo stream with more accuracy. The AREL model proposed here significantly outdoes the existing XEss model on all qualitative aspects. The output also won the Turing test where the majority of the participants believed that the output story was human-made.

3 Dataset and Preprocessing

For this paper, 65-75% of data in the dataset used will be from the Visual Storytelling Dataset (VIST) (Huang et al., 2016) while the remaining 30-35% of data will be taken from Google Images. This ratio may change in the future depending on acquired results for further experimentation. The initial plan is to utilize an estimate of 5000 images, a number that may increase or decrease depending on the results obtained. The train, validation, and test split will be in the ratio of 80:30:20. VIST is a dataset that contains 210,819 unique photos and 50,000 stories. Images in this dataset have been collected from albums on Flickr. Albums include 10 to 50 images and all the images in an album were taken in a 48-hour span. Each story included in VIST has five sentences and each sentence is paired with an appropriate image. The words and inter-punctuation signs in the stories are separated by a space character while all the location names are replaced using the word 'location'. Names of the people have also been replaced using the words 'male' and 'female' depending on the genders of the individuals. This dataset can be downloaded from their official website.

For data pre-processing purposes, the paper will conduct normalization, skew correction, image scaling, and noise reduction on the image dataset. Normalization will change the pixel intensity range, bringing the image to a range that is normal to sense. Skew correction will allow the removal of any skewness of the image making the data more efficient for processing. The images will then be scaled to a certain PPI (pixel per inch) allowing better processing. Finally, noise reduction will smoothen the image by removing any small patches or dots that have higher intensity compared to the rest of the image.

For analyzing the data, the paper used Exploratory Data Analysis (EDA). Target feature correlation is shown with the correlation matrix through a heatmap.

References

Diana Gonzalez-Rico and Gibran Fuentes-Pineda. 2018. Contextualize, show and tell: a neural visual storyteller. *arXiv preprint arXiv:1806.00738*.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 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.

Zainy M Malakan, Ghulam Mubashar Hassan, and Ajmal Mian. 2022. Vision transformer based model for describing a set of images as a story. In *Australasian Joint Conference on Artificial Intelligence*, pages 15–28. Springer.

Kyungbok Min, Minh Dang, and Hyeonjoon Moon. 2021. Deep learning-based short story generation for an image using the encoder-decoder structure. *IEEE Access*, 9:113550–113557.

Marko Smilevski, Ilija Lalkovski, and Gjorgji Madjarov. 2018. Stories for images-in-sequence by using visual and narrative components. In *International Conference on Telecommunications*, pages 148–159. Springer.

Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. 2021. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.

Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. 2018. No metrics are perfect: Adversarial reward learning for visual storytelling. *arXiv preprint arXiv:1804.09160*.