VisionTales: Generating Stories from Visual Sequences

Nayna Baghel nb67@rice.edu
Ishaan Iyer ii10@rice.edu
Department of Computer Science, Rice University

Abstract

The Visual Storytelling (VIST) dataset presents a unique challenge of aligning sequences of images with coherent natural language narratives. In this project, we build a visual storytelling system that generates short stories conditioned on three-image sequences drawn from VIST. Our approach combines CLIP for visual feature extraction with a pretrained GPT-2 language model for story generation. The pipeline is designed to process curated image-caption pairs, leveraging both vision and language modalities to produce contextually relevant narratives. During training, we condition the language model on the visual context encoded by CLIP and optimize it using standard language modeling loss. We evaluate the system through training loss curves and qualitative examples, which demonstrate the model's potential to produce semantically aligned and syntactically fluent stories.

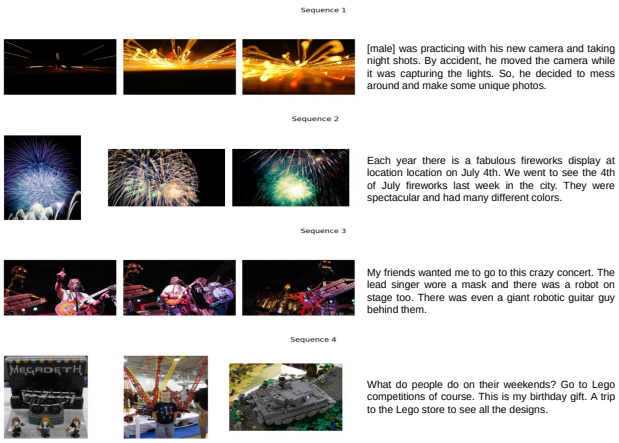

Figure 1. Sample images from the Visual Storytelling (VIST) dataset. VIST contains photo albums collected from Flickr, where each album consists of five images and corresponding human-written captions forming a narrative. The dataset was created to support research in aligning visual content with coherent multi-sentence storytelling.

1. Introduction

Storytelling is a fundamental aspect of human communication, often combining visuals and language to express ideas, emotions, and narratives. In the field of multimodal machine learning, generating coherent stories from sequences of images presents a unique challenge that extends beyond static captioning by requiring temporal understanding and narrative structure across multiple visual scenes. To address this task, we use the Visual Storytelling (VIST) dataset [2], which contains photo albums composed of five images, each annotated with a corresponding sentence. For this project, we select a subset of the data consisting of three-image sequences, filtered to ensure valid annotations and the availability of associated image files. These sequences form the basis for training and evaluating our visual storytelling model. Figure 1 illustrates the structure of the dataset and the preprocessing steps used to construct the final training input pipeline.

Our approach combines two pretrained models: CLIP [3] is used to encode semantic visual information from each image, and GPT-2 [4] serves as the language generation component conditioned on the sequence of captions. The training pipeline is designed to be lightweight and modular, enabling story generation from curated visual inputs without full-scale end-to-end training. This setup allows us to explore how well pretrained vision and language models can be composed for grounded narrative generation.

2. Related Work

Generating coherent stories from images has evolved significantly with the progress in both vision and language modeling. Early captioning models such as Show and Tell [5] and Show, Attend and Tell [6] focused on generating descriptions for individual images using encoder-decoder frameworks and attention mechanisms. The introduction of the Visual Storytelling (VIST) dataset [2] enabled work on generating connected stories from image sequences, highlighting the need for narrative flow and temporal reasoning.

Several approaches have since been proposed to model

this complexity, including hierarchical LSTMs [7] and transformer-based models like Meshed-Memory Networks [1]. More recently, pretrained models such as CLIP [3] and GPT-2 [4] have demonstrated strong performance in cross-modal understanding and language generation, respectively.

What sets our work apart is its modularity: instead of training a task-specific model from scratch, we combine off-the-shelf pretrained components—CLIP for extracting visual semantics and GPT-2 for story generation. This design reduces training overhead while still producing coherent and contextually relevant narratives.

3. Methodology

Our method follows a modular, vision-to-language pipeline designed for generating coherent narratives from sequences of images. The Visual Storytelling (VIST) dataset [2] serves as the basis for training and evaluation. From each photo album, we extract sequences of three images with valid captions, filtering out incomplete or missing data. The three captions are concatenated to form a unified narrative that serves as the training target.

To represent the visual input, we use the CLIP model [3], which provides strong visual-semantic embeddings pretrained on a large image-text corpus. Each image in the sequence is passed through CLIP's vision encoder to obtain a feature vector, and the three vectors are averaged to summarize the visual context of the sequence.

On the language side, we use GPT-2 [4], a pretrained generative transformer model. Instead of modifying its architecture or input layers, we use a prompt-based approach to condition the model. For each input, we prepend a textual prompt such as "Tell a story based on the following images," followed by the ground-truth caption sequence. This allows GPT-2 to remain unchanged while still adapting to the storytelling task.

The model is trained using a cross-entropy loss between the generated tokens and the target captions. CLIP remains frozen throughout training, and only the GPT-2 parameters are updated. This modular design reduces computational cost and simplifies training while still leveraging the strengths of large-scale pretrained models.

4. Experimental Settings

To build our visual storytelling pipeline, we began by designing a training framework that could integrate pretrained models while minimizing computational complexity. Given our focus on modularity and efficiency, we chose to freeze the vision encoder and only train the language generation component.

We use the Visual Storytelling (VIST) dataset [2], which contains around 50,000 images grouped into photo albums, each accompanied by descriptive sentence-level annotations. While the original dataset provides five-image stories, we extract sequences of three images to simplify the input and reduce noise. To ensure quality, we filter out incomplete samples and retain only those with available images and non-empty captions. After preprocessing, our final dataset contains approximately 3,000 valid image-caption sequences.

Each image is resized to 224×224 and normalized using torchvision transforms. These preprocessed images are passed through the frozen CLIP ViT-B/32 vision encoder [3], producing 512-dimensional feature vectors. The embeddings from the three images are averaged to obtain a single semantic representation of the visual sequence.

For language generation, we use GPT-2 [4], which has shown strong performance in open-ended text generation tasks. Instead of modifying the model architecture, we adopt a prompt-based setup: the visual context is used to generate a text prompt such as "Tell a story based on the following images:" followed by the ground truth story text. This design choice avoids end-to-end fusion while still providing visual grounding.

Training is carried out using PyTorch with the AdamW optimizer, a learning rate of 5e-5, and batch size 1. The model is trained for 7 epochs. To manage memory and ensure smooth performance on limited hardware, we keep CLIP frozen and only update GPT-2's parameters. We track training progress using average token-level loss and sample qualitative outputs after each epoch for inspection.

5. Results Analysis and Model Evaluation

In this section, we present a detailed evaluation of our visual storytelling system from both quantitative and qualitative perspectives. The model's behavior during training, its ability to generalize from visual inputs, and the nature of its generated outputs are examined. We also highlight key observations, including the system's strengths, current limitations, and areas for potential improvement. Results are supported by training metrics and representative examples that illustrate the model's effectiveness as well as its occasional shortcomings in handling visually ambiguous input.

5.1. Quantitative Trends

Our training process shows a consistent downward trend in average cross-entropy loss across epochs, indicating that the language model is learning to generate more accurate and fluent outputs over time. Even with CLIP frozen and training performed on limited hardware, the model adapts well to the prompt-based conditioning approach.

As shown in Figure 2, loss decreases steadily across seven epochs, suggesting stable convergence and supporting the efficiency of our modular setup.

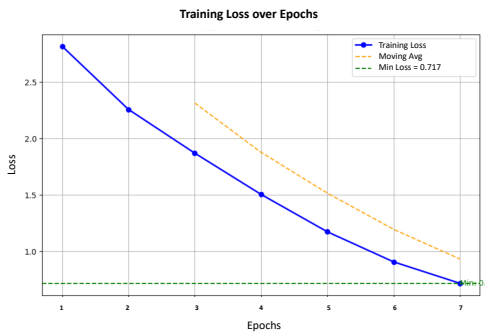

Figure 2. Training loss across 7 epochs. The consistent decline indicates stable convergence using our lightweight setup.

5.2. Qualitative Output Behavior

Generated stories tend to follow the themes present in the input image sequences, especially when the visual content is distinct and easily interpretable. In many samples, the model maintains coherence and fluency, successfully narrating short, descriptive stories.

However, occasional misinterpretations arise due to the fixed CLIP embeddings. For example in Figure 3, in one sequence, an image of a bridge with an arch-like structure was mistaken for a ferris wheel. As a result, the generated story included references to a fairground ride, illustrating how visual ambiguity can influence narrative generation. These moments reflect both the richness and limitations of semantic embeddings from frozen vision models.

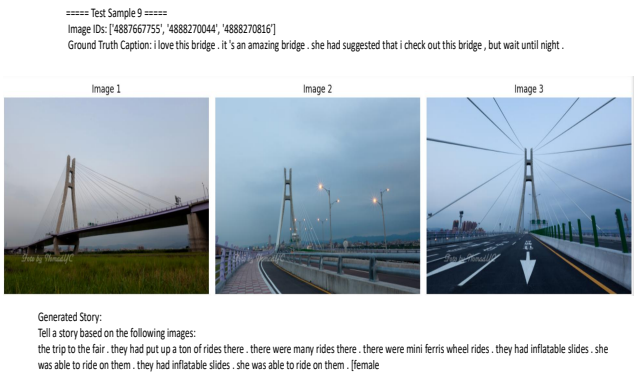

Figure 3. Example of a visual misinterpretation. The model mistook the circular structure of a bridge for a ferris wheel, and generated a story involving a fairground scene. This illustrates how semantic similarity in visual features can influence the narrative, even when the interpretation is incorrect.

5.3. Strengths and Limitations

Our system benefits from a modular and efficient design. By leveraging powerful pretrained components—CLIP for extracting semantic image embeddings and GPT-2 for language generation—we were able to construct a fully functional visual storytelling pipeline with minimal task-specific tuning. The use of prompt-based conditioning enabled integration without modifying model architectures, making the system both easy to implement and adaptable to new datasets or prompts.

One of the key strengths is that the model generalizes well from relatively small amounts of training data. Despite working with a filtered subset of the VIST dataset and training on limited compute, the model exhibited steady convergence and produced coherent, context-aware stories for most image sequences. This shows that combining pretrained vision and language models, even without joint training, can still lead to meaningful multimodal outputs.

However, the approach has its limitations. Since CLIP and GPT-2 were not fine-tuned together, the system lacks explicit alignment between visual regions and textual tokens. As a result, generation sometimes reflects generic language patterns rather than fine-grained visual understanding. In ambiguous cases, the model may invent or misinterpret visual content, as seen in the ferris wheel example. Additionally, without incorporating temporal modeling or deeper reasoning, the system may struggle with more complex narrative structures beyond short, descriptive storytelling.

6. Conclusion

Our project set out to generate compelling visual stories using only sequences of images and the strength of pretrained models. By combining CLIP's ability to capture visual semantics with GPT-2's fluency in language generation, we built a system that performs surprisingly well without the need for complex architectural redesigns or large-scale finetuning.

The training process was stable, and the stories produced were often coherent, expressive, and reflective of the visual input. Even in edge cases—where visual ambiguity led to unexpected interpretations—the model generated contextually believable narratives, showcasing its creative adaptability. This outcome highlights the expressive power of prompt-based conditioning, even when the visual and language models are trained independently.

Despite using limited resources and a filtered subset of the dataset, our system maintained both performance and interpretability. These results show that well-aligned, pretrained models can meaningfully solve complex multimodal tasks without the burden of full-scale retraining.

Looking ahead, this foundation offers exciting directions

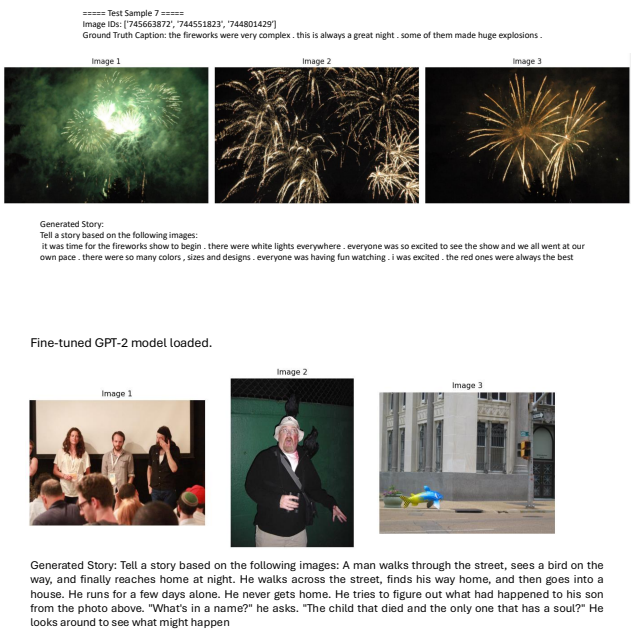

Figure 4. Narrative generated from a three-image input sequence. The output demonstrates effective semantic alignment between CLIP-extracted visual embeddings and GPT-2 language generation, facilitated through prompt-based conditioning without end-to-end multimodal training.

for improvement—such as integrating visual attention, fine-tuning on narrative structure, or extending to longer image sequences. Overall, the success of this work demonstrates that simplicity, when paired with the right tools, can lead to powerful and scalable storytelling systems.

7. Future Implementations

Looking ahead, we aim to explore end-to-end multimodal fine-tuning to better align visual semantics with narrative generation, moving beyond frozen CLIP embeddings. Incorporating scene graphs or structured representations could enhance contextual grounding and logical flow, while integrating a story planning module may help the model maintain thematic coherence across sentences. Additional avenues include training the model to adapt to different storytelling styles or user preferences for creative applications, and expanding to multilingual datasets to improve accessibility and cultural relevance. These enhancements would advance VisionTales toward more coherent, diverse, and user-aligned storytelling capabilities.

8. Ethical and Societal Considerations

Story generation from images can raise issues related to bias, cultural assumptions, and factual inaccuracy. Since the system draws from patterns in large-scale data, it may reflect stereotypes or generate content that is plausible but misleading. These risks are especially important in contexts where fairness, inclusivity, or truthfulness matter.

In our case, the model is used strictly for academic purposes on a publicly available dataset. No personal or sensitive data is involved, and outputs are monitored throughout. Moving forward, careful evaluation and bias mitigation will be crucial as similar models are applied to broader, real-world use cases.

Acknowledgments. We would deeply like to thank Professor Vicente Ordóñez Román for his guidance and for designing a course that encouraged hands-on exploration of vision and language models. We also acknowledge the use of OpenAI's ChatGPT to support the writing and formatting of this report. All implementation, analysis, and final decisions were made by the project team.

References

[1] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020.

[2] T.-H. K. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, C. L. Zitnick, D. Parikh, L. Vanderwende, M. Galley, and M. Mitchell. 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. Association for Computational Linguistics, 2016.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[5] 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*, pages 3156–3164, 2015.

[6] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057. PMLR, 2015.

[7] Y. Yu, J. Wang, Z. Huang, and A. L. Yuille. Bert and hierarchical lstms for visual storytelling. *Computer Vision and Image Understanding*, 202:103103, 2021.