

Multi-modal Creative Story Generation from Images using captioning

Kashish Goel

kg3044@dcolumbia.edu

Abstract

Being able to understand the context of an image using properly formed English sentences is a very challenging task, and to generate coherent stories describing the image is even more difficult. But it could have great impact, for instance in generating creative narratives for marketing campaigns, generating social media posts, etc. In all the previous works of story generation from images, normally only a single sentence or small descriptions are generated. This work generates more creative and coherent stories using those images. The stories are generated using a two part framework as descriptive image captioning using smaller NLP models and then using a large language model for generating creative stories using those. The aim of the work is to generate stories which are coherent with the images and can incorporate the themes and context presented¹. The work successfully generates creative stories for sequence of three random images, where the stories are relevant as well as coherent with the themes present in the images.

1 Introduction

For automated story generation, the research in creating narratives that are not only coherent but also captivating has led to significant advancements in Natural Language Generation (NLG) models. The challenge lies in transcending the constraints of traditional text generation tasks, to embrace the open-ended and creative demands of storytelling (Fan et al., 2018) especially for multi-modal text generation.

The integration of multimodal content is increasingly prevalent, with images frequently accompanying textual narratives in various media. And another extremely useful application is for images to story generation, which caters to generating coherent descriptions of the images. However, exist-

ing models that combine vision and language are primarily trained for tasks other than storytelling, such as image captioning and visual relation prediction (Lu et al., 2019; Wu and Goodman, 2019). These models do not capture the essence of story generation, which requires not just the description of images but the weaving of those images into a compelling narrative.

One of the core challenges in story generation from images is maintaining long-term coherence and ensuring that the narrative adheres to an overarching theme while evolving in composition and direction (Jain et al., 2017). Current storytelling models predominantly focus on text, neglecting other modalities such as images (Yao et al., 2019; Fan et al., 2018; Ippolito et al., 2019). This singular focus can lead to a lack of generalization and robustness, as models trained on a single modality may not effectively transfer learning across diverse inputs and domains (Radford et al., 2019). Conversely, multi-modal generative models, capable of synthesizing information across different modalities, can foster representations that emphasize object relations and interactions (Baltrušaitis et al., 2018). Further, these models by themselves can be used to generate creative and descriptive stories from the sequence of random input images. Nonetheless, predominantly caption a sequence of image snapshots in same setting to generate the story rather than from random images (Wang et al., 2019).

In response to these challenges, this research proposes a novel approach to multi-modal story generation. It introduces a sequential language-and-vision framework that synergizes text and imagery, aiming to produce narratives that are not only coherent and contextually rich but also engaging, encompassing the themes represented in the images. The approach is grounded in the belief that the interplay between text and images can lead to

¹<https://github.com/kashishgoel9/Multi-modal-Creative-Story-Generation-from-Images-using-captioning>

more immersive and relatable stories which can capture the context in the images and generate stories using those. In addition to these contributions, further scope of the research is discussed in Appendix under 'Contributions and Novelty' section.

2 Related Work

The field of creative story generation from images is an interdisciplinary domain that intersects computer vision, natural language processing (NLP), and creative writing. The objective is to transform visual inputs into coherent and imaginative stories, leveraging advanced deep learning techniques and neural networks. This section reviews the related work in this area, focusing on the previous approaches, and evaluation that have been previously considered, and how they relate to this work.

The domain of story generation has seen substantial progress with various methods aiming to produce coherent and engaging narratives. Notably, controllable story generation techniques have been at the forefront of this advancement (Ippolito et al., 2019; Wang et al., 2020; Yao et al., 2019). These methods have been pivotal in guiding the generation process to yield text that is not only coherent but also follows a desired plot trajectory.

2.1 Visual Storytelling and Interactive Models

For Image captioning, there has been a significant amount of work that has been done to generate context-aware captions which can capture the themes presented. These captions are one sentence descriptions which describe the image. Most prior research in visual story generation has relied on the Visual Storytelling dataset (VIST) (Huang et al., 2016). These works primarily utilized global image features extracted from a general vision backbone model trained for image classification (Huang et al., 2016). More recent studies have explored the use of local features, such as object-level information, to enhance the generation of visually grounded stories (Wang et al., 2020).

There are existing architectures like Coherent Neural Story Illustration which propose an encoder-decoder framework for retrieving a coherent sequence of images from sources. The encoder-decoder framework has been widely used, with advancements like attention mechanisms enhancing the relevance of generated captions to the content of images (Xu et al., 2015). Although this model performed well in generating captions, but The

challenge still remained to create captions that are not only accurate in describing the present elements but also in capturing the subtleties and context of the scene in the images. More recent and advanced work has been using the CLIP model which provides a better groundwork to capture the context as well as the themes in the images (Mokady et al., 2021). There are much more recent models which have shown better results than any other so far, like BLIP. But all these previous work have only been able to generate short descriptive captions from images, rather than full-fledged creative stories.

2.2 Controllable Story Generation

Prior research has delved into controllable story generation, aiming to produce coherent text with engaging plots (Fan et al., 2018; Bensaid et al., 2021). Topic-conditioned models, generate stories based on compact topic inputs (Yao et al., 2019). These models offer the advantage of creating concise, progressively evolving storylines. However, these models, particularly those based on the Seq2Seq architecture, tend to lose track of the overarching plot as they progress, often resulting in repetition or deviation from the intended storyline.

2.3 Multi-modal Approaches

The intersection of visual and textual modalities for story generation is less explored. Although there are no known multi-modal architectures specifically designed for storytelling, the integration of vision and language for joint representation has been explored in various successful models. To bridge the gap between visual and textual storytelling, multi-modal approaches have been proposed. Models like MVAE (Wu and Goodman, 2019) have demonstrated the power of combining modalities for a unified representation, albeit not specifically for storytelling.

Story visualization techniques have been developed to either retrieve or generate images that align with a given narrative. These methods, however, often require descriptive text and are computationally intensive, making them less feasible for real-time applications. Our proposed multi-modal story-generation framework leverages robust representations of stories, decoder-based transformer architecture, and controllable text generation to generate creative and coherent short tales. Transformer models like GPT-2, GPT-3, and TransformerXL have demonstrated success in generating diverse, stable text. In summary, this proposed multi-modal story-

generation combines ideas from various domains to generate creative and coherent stories, leveraging the strengths of transformer-based models and controllable text generation. Evaluating this framework on the complex story-generation task holds promise for advancing the field.

3 Data

Different datasets have been used in this research for the two subparts of the work.

Dataset statistics			
	MS COCO	Flickr8	Writing Prompt
Corpus Size	328,000	8000	300,000
Train size	56021	-	15,620
Validation size	13979	-	-
Test size	-	8000	-

Table 1: Dataset statistics

3.1 Image Captioning Datasets

The main dataset used for training the image captioning model is MS COCO (Large dataset with 328,000 images and 5 captions per image) (Lin et al., 2014). It is known for its extensive collection of images with multiple captions, making it valuable for training deep learning models. It is one of the most widely used datasets for the image captioning task. For training the Baseline model as well as the CLIP Prefix model, only MS COCO dataset was used. Another smaller dataset, Flickr8k (Rashtchian et al., 2010) consisting of 8,000 images, each associated with five different captions was used as the test dataset for the both the baseline as well as the final CLIP Prefix model. These images capture a diverse range of scenes and descriptions and would be good test set to check how well the models generalize as well.

3.2 Story Generation Dataset

The WritingPrompts dataset (Fan et al., 2018) comprises of 300,000 human-written stories paired with writing prompts from an online forum. This dataset is rich in creative narratives and serves as an excellent resource for story generation tasks. Writing Prompts dataset is primarily used for the story generation subpart of the research. The dataset is divided into prompts and stories, both of which are utilized to train the model to generate creative and coherent text. During training, the data is concatenated in the format: <prompt + ' <sep> ' + story>

For image captioning, the input consists of images from the datasets along with predefined 5 captions for each image, and the output are the captions generated by the model for all the test images, stored as pair of three captions together. Whereas for story generation, the inputs are the writing prompts (consisting of a prompt along with the pair of three captions), and the output are the generated creative stories. Dataset summary is presented above in Table 1.

4 Methods

The aim of the project is to build an image to story generation system. A pipeline using two different models is used to create a multi-modal system to generate creative stories using initial prompt and relevant random images. To generate stories which capture the themes of the images and are coherent, the system is divided into sub-systems. The first is the Image captioning model which takes the images as the input and generates coherent captions and subsequently a story generation model is trained to generate creative and coherent stories taking these captions from random images as the input along with a story prompt to generate a creative story which is coherent with the themes present in all those images.

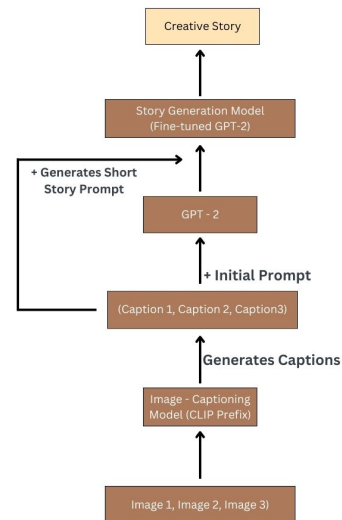


Figure 1: End-to-End System Architecture

4.1 Captioning Models

This initial baseline model takes the images as the input and generates short descriptive captions as the output. A model encodes variable-length inputs (images) into fixed dimensional vectors, and then

uses these representations to decode them into desired output sentences. It also utilizes stochastic gradient descent during training. This model is a Hierarchical Neural Model which utilized vision based model.

4.1.1 Clip Prefix Model

This model uses the expressive embedding of CLIP (Contrastive Language-Image Pre-Training) (Radford et al., 2021) for visual representation. The method leverages the semantic embedding of CLIP, which encapsulates essential visual data, as a condition to the caption. This condition is treated as a prefix, allowing the use of an autoregressive language model (GPT-2) to predict the next token without considering future tokens. The objective function is formulated as:

$$\max_{\theta} \sum_{i=1}^N \sum_{j=1}^{\ell} \log p_{\theta}(c_{ij} | x_i, c_{i1}, \dots, c_{ij-1}) \quad (1)$$

where c_{ij} represents the j -th token of the i -th caption, x_i is the image, and θ are the parameters of the model.

GPT-2 is employed as the language model, with its tokenizer projecting the caption to a sequence of embeddings. The visual information from an image x_i is extracted using the visual encoder of a pre-trained CLIP model. A mapping network F then maps the CLIP embedding to k embedding vectors:

$$p_{i1}, \dots, p_{ik} = F(\text{CLIP}(x_i)) \quad (2)$$

Each vector p_{ij} has the same dimension as a word embedding. These vectors are concatenated with the caption embeddings to form the prefix-caption concatenation used during training.

The training objective is to predict caption tokens conditioned on the prefix in an autoregressive manner. The mapping component F is trained using cross-entropy loss:

$$\mathcal{L}_X = - \sum_{i=1}^N \sum_{j=1}^{\ell} \log p_{\theta}(c_{ij} | p_{i1}, \dots, p_{ik}, c_{i1}, \dots, c_{ij-1}) \quad (3)$$

The Mapping network in particular converts CLIP embeddings into the format used by GPT-2. Here the language model is kept frozen and a transformer architecture, known for its global attention mechanism and efficiency in handling long sequences is used, which also gives the flexibility in experimenting with the prefix length.

The transformer network takes two inputs: the visual encoding from CLIP and a learned constant. This constant serves two purposes: extracting relevant information from the CLIP embedding and adapting the fixed language model to new data. The language model is static (frozen) and the transformer network learns a specific set of embeddings.

In contrast to the Neural Image Captioning model, This CLIP Prefix model should perform better as it is a lightweight captioning approach that utilizes pre-trained frozen models for both visual and textual processing. For this reason, for the image captioning task, both the Neural as well as CLIP models are experimented with and finally the better performing model is used as the final model.

4.2 Story Generation Pipeline

The benchmark model architecture generates text from a given prompt using a decoder-based transformer (GPT-2), fine-tuned on Writing Prompt dataset. It retrieves image captions from the captioning model as the input.

Further, for the final integrated pipeline, the images guide the text generation process where in addition to the taking the captions generated by the CLIP Prefix model as in the previous section, an additional decoder only GPT-2 model is trained in zero-shot setting to generate a short story prompt for each input of the pair of three captions. For this the model is prompted in the format: "Use all the three captions and generate a short story:<caption 1, caption2, caption3>"

A pre-trained GPT-2 model is then fine-tuned to generate creative stories, which takes input prompts as the captions/descriptions generated from the previous smaller model along with the short story prompt, concatenated together and generates a creative story plot around it for the respective images, as the output. GPT-2 model especially performs very well for the generation task for creative stories and hence it is used as the primary model for the task here. In addition to GPT-2 model, encoder-decoder models like T5 and BART are also fine-tuned with Writing Prompt dataset. Further details are present in Appendix. The final model pipeline using GPT-2 model for story prompt and then a fine-tuned GPT-2 model, improves upon the benchmark by revising both the text generation and image retrieval methods.

5 Experiments

5.1 Experimental Setup

A neural network architecture consisting of both CNNs (Encoder) and Decoder with attention mechanism is used to automatically generate captions from images as the baseline model. Additionally, a Clip Prefix model is implemented using the CLIP embeddings and pre-trained CLIP model along with GPT-2. Based on the comparison of results and quality of captions generated, CLIP Prefix model is used for generating captions in the final Image-to-Story pipeline. For story generation, the baseline model used is a fine-tuned GPT-2 model which generates creative captions based on the prompts given (image captions along with initial prompt). This model is modified to give more context to it and eventually an additional GPT-2 model is also used to first generate short story prompts based on the pair of three captions for each set and then the fine-tuned GPT-2 model is used to generate stories for each pair of these three random images. BART and T5 models are also fine-tuned using the Writing Prompt dataset to generate creative stories, but due to computational limitation and GPU available, the experiments were tested specifically with the GPT-2 models for the final pipeline for the research.

5.1.1 Image Captioning Model

The baseline image captioning model employs an encoder-decoder architecture, consisting of a pre-trained InceptionV3 convolutional neural network (CNN) with Multi-head Attention Transformer layer as the encoder to process images. It extracts features from the images, which are then reshaped to form a suitable input for the transformer layers. The transformer encoder layer is designed to process these image features, utilizing attention mechanisms and layer normalization for efficient learning. The encoder processes image data and generates a feature vector, while the decoder generates captions for the images based on the encoded information.

For the text generation part, a transformer decoder layer is used. It incorporates embeddings that combine token and position information, and employs multi-head attention mechanisms to focus on different parts of the input sequence. The decoder also includes feed-forward networks and dropout layers for regularization. The model is trained to minimize loss and maximize accuracy,

with the ability to augment images during training for better generalization.

Data preprocessing involves converting text to lowercase, removing special characters and numbers, and tokenizing the sentences. Word embeddings are generated using an embeddings layer. Data is generated in batches to manage resource consumption effectively. During training, image embeddings and corresponding caption text embeddings are used as inputs, with the transformer layer generating captions word by word. Attention masks are implemented to enhance model accuracy and reduce overfitting.

Clip Prefix Model Training For this model, an input image x is used to extract its visual prefix using the CLIP encoder and the mapping network F . The caption generation begins with this visual prefix, and the language model (GPT-2) guides the prediction of subsequent tokens. For each token, the model calculates probabilities for all vocabulary tokens. These probabilities are then used to select the next token, either through a straightforward selection method or a more complex beam search strategy.

For this purpose, there are different components that are implemented as follows: A MLP (Multi Layer Perceptron) implements a simple neural network with two linear layers and dropout, used for transforming inputs. Further, a multi-head attention mechanism is implemented, crucial for capturing different aspects of the input data in parallel. Transformer Layer then combines multi-head attention with a feed-forward network. The Transformer Mapper, maps the inputs and processes the CLIP embeddings. Then the CLIPCaption Model finally integrates the GPT-2 model with either MLP or Transformer for generating captions from CLIP embeddings. Then for generating text, either a beam function can be used, which considers multiple possible next steps to find a more optimal sequence, or another function which uses a top-p sampling strategy, selects the next word based on a probability threshold, allowing for more diverse and creative outputs.

For the MLP mapping networks, the prefix length is set to $K = 40$, incorporating a single hidden layer within the MLP. In the case of the transformer mapping network, it is configured with $K = 40$ constant tokens and equipped with eight layers of multi-head self-attention, each having eight heads. For the training process, the map-

ping type used was Transformer. To train the model quickly, the additional training stage generally used to optimize metric such as applying self-critical sequence training as well as effective cross-entropy, are avoided.

5.1.2 Story Generation Model

The implemented model is GPT-2, which involves first zero-shot transfer learning, where the model is pre-trained on language modeling without task-specific fine tuning. The model is then also fine-tuned using writing prompt dataset, and then used to generate stories based on these prompts. Data preparation involves using the validation dataset as the training dataset due to time constraints. Prompts and corresponding stories are combined into a single file for input. GPT-2 tokenizes text using Byte Pair Encoding (BPE) and enforces a maximum sequence length of 512 by truncating longer sequences and padding shorter ones. Labels (targets) for training are generated by excluding padding tokens and aligning them with input tokens. The pretrained GPT-2 model is readily available through the transformers package. Before fine-tuning, the model's performance is evaluated on a validation dataset, measuring the average perplexity of the evaluation results. After fine-tuning, the same prompts are used to generate stories with the improved model.

For the final Story Generation model, changes were made to the baseline GPT-2 model to improve the generation by introducing another zero-shot GPT-2 model at the beginning to provide more context to the story model related to the images. A pre-trained GPT-2 model and tokenizer are used initially to generate a short story prompt in continuation to the given input of pair of three captions from the captioning model. The tokenizer is configured to pad on the left side, and the end-of-sequence token is used as the padding token. The input text is then tokenized, and an attention mask is created. The tokenization is truncated or padded to a maximum length of 512 tokens. The output along with the caption and the prompt are concatenated together, which is then passed to the fine-tuned GPT-2 model as trained above. This fine-tuned model then generates the story for each pair of three captions. This final model was then integrated with the Image Captioning model to generate coherent stories from a sequence of random images.

5.2 Evaluation Metrics

Evaluation is done separately for the image captioning as well as the story generation model. And in addition to these the whole Image-to-Story model is also evaluated using different metrics.

5.2.1 Image Captioning

The most commonly used automatic evaluation metric so far in the image description literature has been the BLEU score (Papineni et al., 2002), which is a form of precision of word n-grams between generated and reference sentences. For the image captioning models as well, BLUE score is used to evaluate their performance. Further human feedback/evaluation process was also used for evaluation (context-aware) in addition to the BLUE scores.

5.2.2 Image-to-Story System

The GPT-2 models are evaluated using the 'Perplexity' scores. The main goal is to study the stories generated with respect to the image modality. For this, 'Reference-free metrics' are used to evaluate the coherence between the text generated and the themes presented in the images - Fluency and Readability (Using the Flesch-Kincaid readability test to measure the readability of the text) (Sai et al., 2021), Lexical Diversity(Using Type-Token Ratio (TTR) to measure the diversity of vocabulary), Narrative Structure(analyze simpler aspects like sentence length variability as a proxy for complexity). A separate new metric (WorLap) is also created to evaluate the whole model by calculating the percentage of the number/ratio of words from the pair of captions, which are also present in the story generated as well. For this purpose specifically, the story prompts generated by GPT-2 are modified to only contain the short prompt and not the captions themselves while comparison. This ensures that the word overlap from the captions themselves are not considered. In addition to this, human annotators also annotate the stories generated for coherence and relevance.

5.3 Experiments

First, I evaluated the Image Captioning models by evaluating the quality of captions generated by the model. For this, the captions generated are evaluated using the BLEU score. The tokenized generated captions are compared against the original captions for both the Neural as well as the CLIP Prefix model.

In the second experiment, the stories generated are evaluated. 'Perplexity' is used to evaluate how well the model is able to predict/generate the next word. The GPT-2 model is fine-tuned to generate coherent stories on Writing Prompts dataset. Further, multiple experiments were performed to analyse how coherent the stories are, when generated using a fine-tuned GPT-2 model as compared to without fine-tuning.

The main experiment was to evaluate the entire pipeline for coherence and relevancy of the stories generated from the pair of three captions. An initial experiment was done for the baseline pipeline, where the three captions (generated by the image captioning model) by the first model were used as input prompts to generate stories from the GPT-2 model. These results were evaluated manually for this experiment along with the separate evaluations. Then further the best model was chosen for the Image Captioning task and was integrated with the final, more-context aware Story Generation model to generate results and evaluated based on Automatic and newly created metrics. Further human evaluations were also performed for the stories generated by the entire final system.

6 Results

6.1 Image Captioning

To gain a comprehensive understanding of the optimal hyper-parameters, I followed the guidelines outlined in the paper (Liu and Brailsford, 2023). For the training process, I initially set a minimum word count threshold at 5, employed an embedding size of 512, and utilized a hidden size of 512 as well. The initial training duration lasted for 3 epochs. Upon initial inspection, the loss exhibited the expected decreasing trend. However, after training for an extended period of 20 hours and subsequently evaluating the network on test images, it became evident that the model had overfit to the training data.

This overfitting was evident in the generated captions, which were unrelated to the test images. I repeated the inference with models trained after each epoch, but the results remained unsatisfactory. Consequently, I decided to reduce the embedding size to 256 and retrained the model with attention masks, this time for only 1 epoch, which led to a significant improvement in performance. And it was seen that on increasing the number of epochs to 2, the performance increased. For the CLIP Pre-

Evaluation scores		
Model	BLEU	Coherency (Human)
Neural	0.209	0.69
CLIP Prefix	0.275	0.87

Table 2: BLEU scores for the Image Captioning Models

fix model, I first used mapping network as MLP and tried it with the prefix length set to 10. The results were comparable to the Neural model, but on changing the mapping type to transformer and increasing the prefix length from 10 to 40, gave better and more relevant captions. Furthermore, the CLIP Prefix model achieved a better BLEU score of .275 as compared to the BLEU score for the Neural model and therefore, it was also used as the final model integrated with the story generation model in the pipeline. The results are shown in Table 2 and the final loss in Fig 3. The model was trained on the MS COCO dataset but tested using the Flickr8k dataset. Hence the evaluations and BLUE scores reported are of the model trained on MS COCO and how well they generalized to Flickr8k dataset. As an example, the captions for the same image from both the models can be seen in Fig 2

6.2 Story Generation

I initially performed a zero-shot story generation evaluation for the GPT-2 model and the evaluation yielded an average perplexity score of 39. For this, I selected a prompt from the validation set and used it as an input prompt for the model, instructing it to generate a 300-word story. I utilized the model's built-in generate method, which supports various decoding options. Then I fine-tuning the model, utilizing a training dataset (Writing Prompt) comprising 15,620 samples. Training the model on a single GPU required approximately 21 minutes to complete a single epoch. Following one epoch of training, the perplexity for the validation dataset improved, decreasing to around 24, which represented an enhancement compared to the pre-fine-tuning perplexity score. Hence, for the final model, only the fine-tuned model was used.

6.3 Image-to-Story System

The entire system for the image-to-story pipeline was evaluated using various automatic evaluation metrics as well as through human evaluation. The CLIP Prefix model was tested on the Flickr8k

dataset, out of which the initial 150 images were taken and divided into 50 pairs of 3 captions each. These 50 pairs of captions were given as input to the GPT-2 model to generate short story prompt for each of the pairs of captions, which were then passed to the fine-tuned GPT-2 model to generate creative stories for each of the 50 pairs. The results for the stories generated using different models are shown in Fig 4. The aggregated evaluation scores for all the metrics are presented in the Table 3, along with the human evaluations done. Human

Evaluation Scores	
Metric	Clip Prefix + Fine-tuned GPT2
WorLap	0.40
Readability	0.032
TTR (lexical diversity)	0.55
Complexity	0.27
Human Evaluation	0.69

Table 3: Evaluation scores for the Creative Stories from the entire Image-to-Story pipeline

evaluations regard the stories highly for coherence and much better than the new automatic evaluation metric devised and also in general we can see that the scores also represent that although not great, but the stories generated are coherent and relevant to the images. The stories were able to identify the underlying context in each of the images and use those to weave a narrative. Further, the results show that using the zero-shot GPT-2 model first to generate short story prompts and then using the fine-tuned GPT-2 model performs much better than the baseline GPT-2 model for the different metrics. And we also see that using a better captioning model (CLIP Prefix), generates better and more context-aware captions which contributed in generating much more relevant and coherent creative stories for the sequence of images.

7 Error Analysis

7.1 Captions

For the Baseline Neural Image Captioning model, the results seemed decent, but there were a few instances where the model generated captions which depicted incorrect information. In Fig 2, the model generated details that are not present in the image, such as objects or actions (4 people, whereas in the image there are only two people) (Hallucination). The caption was also too generic and not specific enough and hence failed to capture dis-

tinctive elements of the image. It did not identify the drawings/paintings as well as the scene behind (snow). It also inaccurately described elements of the image. As in this case, it thought the people were wearing costumes and was not very good at capturing the context or the relationship between the elements in the image.

The caption generated by the CLIP Prefix model is more accurate and correctly identifies that people are standing in the snow and does not hallucinate, however, fails to capture more detailed contextual information such as presence of artwork on display, which indicates that while the model has a better grasp of the general context, it may not be sufficiently detailed in its analysis of the scene. Although CLIP Prefix model performs better and generates more contextually aware descriptions, it seems to still lack specificity.

7.2 Generated Stories

The generated story by the Neural model with fine-tuned GPT-2 is incoherent and disjointed as we can see in the Fig 4. It includes elements that do not relate to the captions or the images, such as robbery, a green lawn, and anime. And hence, the model hallucinates a little as it generates unrelated content. There is also an inconsistency in the narrative and lack of continuity between sentences, implying the model struggles with creating a cohesive story. The mention of specific themes like "NSFW" and "anime" show that the model does create creative stories but the themes are highly unrelated.

Whereas, for the story generated by the final system, we see that the story is more coherent and follows a clearer narrative. It describes emotions and a scenario involving family revolving around the themes presented in the image. However, the narrative contains elements of judgment and speculation about the characters' feelings and morality, such as calling a character "a terrible parent," which may not be directly concluded from the images.

The final system provides more detailed and specific captions, but both models struggle with creating a detailed story that remains true to the images. The final model seems to have a better understanding of the context, but both models fail to consistently apply this understanding to create a coherent story that logically follows from the images and captions.


Image	Original Caption	Predicted Captions	
		Neural Model	Clip Prefix Model
	man on skis looking at artwork for sale in the snow	Four people in costumes on skis	A couple of people that are standing in the snow

Figure 2: Difference in Captions generated by the Two models. We see that the CLIP Prefix model understands the context a little better than Neural model. It also identifies 'snow'

8 Conclusions, Limitations, and Future Work

In this paper, the primary focus was on generating creative and coherent stories from images. This was achieved using a two-part framework, consisting of descriptive image captioning, followed by the generation of creative stories using a large language model. The study aimed to create narratives that are coherent with the images and incorporate the themes and context presented. The research encountered several limitations. One of the key challenges was maintaining long-term coherence in the generated stories and ensuring they adhere to an overarching theme while evolving in composition and direction. For future work, there could be further work done to improve the coherence between the story and the images by experimenting with encoder-decoder models to see if the stories generated from them are more coherent than decoder-only models. Further, curating a new dataset consisting of captions and corresponding stories on which the story generation model can be trained and fine-tuned can help in generating more coherent stories which can better identify the underlying themes and context of the images. These suggestions aim to improve the capabilities of multi-modal story generation systems building on this work.

References

- Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.
- Eden Bensaïd, Mauro Martino, Benjamin Hoover, and Hendrik Strobelt. 2021. Fairytaylor: A multimodal generative framework for storytelling. *arXiv preprint arXiv:2108.04324*.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.
- 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.
- Daphne Ippolito, David Grangier, Chris Callison-Burch, and Douglas Eck. 2019. Unsupervised hierarchical story infilling. In *Proceedings of the First Workshop on Narrative Understanding*, pages 37–43.
- Parag Jain, Priyanka Agrawal, Abhijit Mishra, Mohak Sukhwani, Anirban Laha, and Karthik Sankaranarayanan. 2017. Story generation from sequence of independent short descriptions. *arXiv preprint arXiv:1707.05501*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.
- Haixia Liu and Tim Brailsford. 2023. Reproducing “show, attend and tell: Neural image caption generation with visual attention”. In *Journal of Physics: Conference Series*, volume 2589, page 012012. IOP Publishing.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.
- Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using amazon’s mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon’s Mechanical Turk*, pages 139–147.

Ananya B Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M Khapra. 2021. Perturbation checklists for evaluating nlg evaluation metrics. *arXiv preprint arXiv:2109.05771*.

Ruize Wang, Zhongyu Wei, Ying Cheng, Piji Li, Haijun Shan, Ji Zhang, Qi Zhang, and Xuanjing Huang. 2019. Keep it consistent: Topic-aware storytelling from an image stream via iterative multi-agent communication. *arXiv preprint arXiv:1911.04192*.

Su Wang, Greg Durrett, and Katrin Erk. 2020. Narrative interpolation for generating and understanding stories. *arXiv preprint arXiv:2008.07466*.

Mike Wu and Noah Goodman. 2019. Multimodal generative models for compositional representation learning. *arXiv preprint arXiv:1912.05075*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.

9 Appendix

9.1 Contributions and Novelty

This proposed method contributes to the field by introducing a novel two-model system that utilizes image captioning as a stepping stone towards generating creative stories. This approach is distinct from prior work that either focuses on generating short captions or uses text prompts for story generation. By fine-tuning an LLM like GPT-2 with captions generated from images, this project achieves

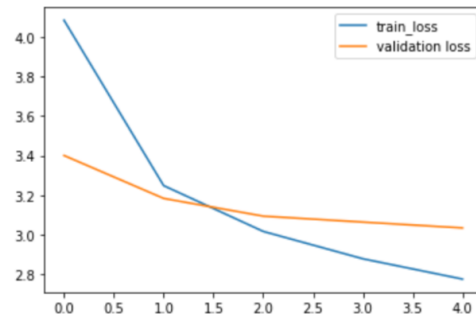


Figure 3: Loss Plot for Image Captioning Model

a more nuanced and contextually aware narrative output.

In summary, while there have been significant advancements in the field of creative story generation from images, this work introduces a unique approach by leveraging the strengths of both image captioning and LLMs. This method holds the potential to address some of the challenges faced by existing models, such as the generation of longer, coherent, and contextually enriched stories.

9.2 BART and FLaan-T5 training for Story Generation

In addition to GPT-2 model, encoder-decoder models like T5 and BART are also fine-tuned with Writing Prompt dataset to compare the quality of stories generated for coherence as opposed to decoder only model (GPT-2). The data pre-processing is done in the exact same way for these models. For the BART model, initially the facebook’s bart large model was tried to be fine-tuned, but due to computational limitation and more number of GPUs required for the compute, the model size had to be reduced. For training, I tried reducing the batch size, introducing gradient accumulation steps and also mixed precision training (fp16=True), but it still required more compute and GPU access for the model to be fine-tuned.




<u>Model</u>	  		
	<u>Captions</u>		
Neural	Two people near a gate	A man with hair	A girl with rope
CLIP Prefix	A man laying on a bench next to a dog	A man with a hat on his head	A little girl holding onto a ladder holding onto a rope
	<u>Stories Generated</u>		
	<u>Creative Story</u>		
Neural + Fine-tuned GPT2	<p>Two people near a gate, A man with hair, A girl with rope, A dog and a cat. They were robbed one evening while walking to work. Just being twenty years old now...She was getting ready to walk home from work, sweet mothergirl. So, she's been helping from the back of the van, kinda slow, just like the day she was last...That's a lovely green, green lawn...*One Piece Made By one to conquer. *On this monster, anime, NSFW * I've been a cop for ten years, and I still have it handy... The day I met her was a bummer. I didn't have time to think about the big decisions that would be to bring her home. Who would have thought she would have suddenly decided to visit me? Not me. I was thinking... `` Hey, Johnny, is something going on? " `` Um, it looks like something you just picked up, I just went off and had coffee with him.</p>		
CLIP Prefix + Zero-shot GPT2 + Fine-tuned GPT2	Story Prompt (Zero-shot GPT2)	Creative Story (Fine-tuned GPT2)	
	<p>The story is a bit of a mess, but it's a good one. It's about a man who is trying to get his life together, and he's trying not to be a burden to his family. He's not trying for</p>	<p>The story is a bit of a mess, but it's a good one. It's about a man who is trying to get his life together, and he's trying not to be a burden to his family. He's not trying for sympathy, he's trying to create. It's not one paragraph long. The story has to end with someone comforting the family, whatever that is. It starts with a boy sobbing on his mother's couch. He sits there crying, as though someone is going to come over and take care of the kids in his life. Then he just sits there and listens. He's not going to talk about how much he hurts, how hurt he gets. He's a terrible parent. He doesn't seem to care how his children die, or how he makes sure he meets them. Or what color life is, or if it has a black person, or a white person. He just seems to stay quiet. The main character, Nathan, is also crying while standing, and an older boy is leaning on a rope, and the younger boy looks up from the rope, scratching the cloth away, `` Oh god what the hell, how did you hurt her? " He's trying to remember the others.</p>	

Figure 4: Results for the Image-to-Story System: Creative Stories generated using the Neural Model and the CLIP Prefix Model with fine-tuned GPT-2 with and w/o additional zero-shot context GPT-2 model