Visual Story Generator

Divyanshi Thapa, Tatvavid Tripathi
School of Computer Science and Engineering
Vellore Institute Of Technology, Chennai.

Abstract

The recent developments in deep learning models like CNN, LSTM, GRU, Transformers etc. have significantly improved the tasks related to visual and language learning. One of such task is visual story generation in which an image is given as input and according to that a story is generated by using the deep learning models. Our aim is to develop a similar application which can generate a story and recite it to the user. The input image features are extracted using VGG-16 model and from that, the captions are generated using LSTM which are given to GPT2-124M model for story generation.

1 Introduction

There are various perks of teaching machines to understand the context and come up with short stories like knowing how to compose unique responses and even quick story lines by which chat and customer support bots would become capable of holding more practical discussions and intelligently responding to customer queries, no matter how complex they are. Additionally, these algorithms could help marketers deal with such tasks as creating product image captions and product descriptions for images, along with various other possibilities to use the concept of story generation.

Some years ago, Ryan Kiros from the University of Toronto published an open-source project called neural storyteller on GitHub. The neural network was trained on many romance novels to deliver descriptions for images. However, modeling paragraphs of readable text with a high-level structure was a problem as most of the NLP algorithms would only generate word-by-word summaries accurately. They would not think ahead and map out a good story plot. After that, further work like (Swanson and Gordon, 2012) and (Mitchell et al., 2018) has been done in the field of ML storytelling, which were able to generate stories from a written text but in the process parting ways with the main idea and deviating off-topic altogether by shifting its focus on some unimportant pieces of text.

In our project, we will be using images as input to generate relevant stories, maintaining the coherency between sentences and paragraphs, adding some characters and subplots, and following the same theme and idea throughout the story.

VGG-16 is a CNN based model and a significant milestone in the computer vision domain which uses 3x3 convolutions making it very simple and easy to work with. In this model a image with 244x244 dimension is passed with three channels that is red green and blue (RGB) with normalisation of these RGB values as pre-processing. The image is passed through first stack of 2 layers which has 64 filters each, followed by a second stack having 128 filters. Third stack of 3 convolutional layers each containing 512 filters, after which 3 more connected layers with a flattening layer in present. The output layer is followed by softmax activation layer. It can be used in various applications like image recognition or classification and image detection and localisation.In most scenarios this is a simple and easy to use model but there are some scenarios where it has some issues like there is no specific measure to control the exploding or vanishing gradient problem which was later addressed in the ResNet.

Long Short Term Memory (LSTM) network is a type of RNN capable of remembering all the past knowledge and forgetting the nonrelevant data, using function layers called gates which make it far more effective than a traditional RNN model.

GPT2 is a large transformer based language model trained on a dataset of 40gb having 8 million web pages and has about 1.5 billion parameters and hence easily outperforms the other language models trained on specific domains The objective of the model is pretty simple, that is predicting the next

word given the previous words of a text. When used to generate a lengthy text, it can come up with very high quality of text which can be used for many applications like question answering, translation and summarization. The model can also be fine tuned on the datasets to suit itself for a particular domain and generate the text similar to the data being given, which provides it even more potential and control over the generated text and maintaining the coherency till the end. It creates a persistent TensorFlow session which stores the training configuration, and runs the training for the specified number of steps.

These models have been combined in the proposed architecture by using some of the VGG-16 layers to generate features from the images and then using these features as input for the LSTM for sequence prediction. By doing this, the proper context of the input image is recognized, and relevant captions are generated as output.

This generated caption is then used as input for a GPT2-124M model finetuned on three genre of story corpus which include horror, sci-fi and humor, each in different sessions of the GPT2 model. This allows us to generate stories of three different genres using the same caption and image. The benefit of using the GPT2 model is that the overall idea of the story and coherency is maintained between sentences and paragraphs. The plot and thematic ideas of the story are carried along to the very end, which can generate stories having good narration, making them much more enjoyable.

1.1 Real world applications

1. **Education and imagination skill enhancement for kids**

   This application can be very useful for the kids and elementary school students as it will provide the description of the image in a form of a story which will help them to imagine a given scene into an entirely different way.

2. **Source of entertainment for old people**

   With an added component to convert text to speech, it can be used as the source of entertainment for the elderly people whenever they feel bored.

3. **Provide a base model for VQA**

   Visual question answering (VQA) is a one of the recent research area in NLP where the picture and a question related to it is fed in the application and the application tries to answer it. This application has a wide use for visually impaired people. But the main problem is generation of answer for the question asked in context of the image given as an input. Generation of the captions and deciding the theme for the story component ca be useful for the VQA problem.

2 Related Work

With the advancement in technology, the NLP has attracted a lot of researcher's attention and many tasks and challenges were addressed. One such task was Natural Language Generation (NLG). The task of coherent story generation from independent descriptions, describing a scene or an event has been one of the special topics of interest in NLG. The authors in (Parag Jain, 2017) worked on two popular text-generation paradigms. The first one was Statistical Machine Translation, posing story generation as a translation problem and second is deep Learning, posing story generation as a sequence-to-sequence learning problem. For SMT, they implemented a RNN architecture which encoded sequence of variable length input descriptions to corresponding latent representations and decoded them to produce well-formed comprehensive story like summaries. Their future work included the designing of trainable metrics for evaluating stories holistically to include aspects on creativity, coherency, novelty and other parameters compared to current score computation which is based on exact match. RNNs have also been used widely for sequence-to-sequence learning,for example in sequence to sequence models for data to text NGL(Natural Generation Language) (G. Jagfeld and Vu, 2018). Standard sequence-to-sequence models have found application in open-ended content generation, but such straight forward encoder-decoder setups fail to generate content conditioned on a given starting seed. Researchers proposed a state-of-the-art hierarchical neural fusion architecture using two seq2seq models along with multi-scale gated attention mechanism ensuring relatedness of generated content to a given prompt. Since simple encoder-decoder architectures fail to model important meaningful representations

of words and phrases,used Gated Recurrent Unit (GRU)-based neural checklist models for recipe generation. GPT2 was developed by OpenAI in 2019. With that, the idea of text translation and generation took a huge leap because it generated text output on a level that sometimes it was indistinguishable from that of humans. However, Bowen Tan , Zichao Yang, Maruan Al-Shedivat, Eric P. Xing, Zhiting Hu (Bowen Tan, 2021) revealed that it is still challenging for such model to generate coherent long passages of text (e.g., 1000 tokens), especially when the models are fine-tuned to the target domain on a small corpus. For this, they proposed a simple but effective method of generating text in a progressive manner, inspired by generating images from low to high resolution. Their drawback and future work was they were not able to do the generation at multi stages. Combining the art of image captioning and generation of story, researchers have tried to develop a system to automatically generate stories from the images. B Venkat Raman, Nagaratna P Hegde, NenavathVenkateshNaik, Allu Siva Kishore Reddy (B Venkat Raman, 2019) used CNN model trained on MS COCO dataset for image feature extraction and captioning. For the story generation, they used LSTM network. One of their drawback was the tested results for a small stories dataset may not be accurate. It took time in training the of the models. An LSTM just like RNN which is also used for sequential data and it will be used to predict the next word of sentence when we pass one word as input which is a previous word of the predicted word. LSTM also used in Language Models also for the purpose of text generation.LSTM and RNN were used for characters level predictions also. A study of overall model stability and performance showd that fine-tuned GPT2 language models have the least deviation in metric scores from human performance. Using a diverse set of automated metrics, the authors (Avisha Das, 2020) compared the performance of transformer-based generative models OpenAI's GPT2 (pre-trained and fine-tuned) and Google's pre-trained TransformerXL and XLNet to human-written textual references. The authors of (Kyungbok Min, 2021) developed an unsupervised deep learning-based framework that combines a recurrent neural network (RNN) structure and encoder-decoder model for composing a short story for an image, and a huge story corpus, which included two different genres (horror and romantic), manually collected and validated. They used two separate datasets. The first was the books downloaded from the Smashwords website, which is a website where the authors share their unpublished books with the community and second was the Conceptual captions dataset, which is a huge captions dataset introduced by Google in 2018. One of the things that they didn't consider was grammar and detail context modules. Another weakness of the system is the deep learning model. Only a basic CNN model with the GRU method was implemented, but it was originally developed for image processing.

3 Proposed Work

The main aim of this work is to generate a story from a given image. For this, we divide this work in two sub parts. The first one is the image caption generation in which the image will be fed and the captions will be generated and the second part the story generation where a story will be generated from the given keywords. The work flow will be as follows:

The input image will be passed on to the Image caption generator model to obtain features present in the input image. The image features will be generated from the VGG-16 model. Based on the obtained features, it will generate captions for the input image. The output will look like a random sentence about the image's content. Convolution Neural Network (CNN) and Long Short Term Memory Network (LSTM) are the two models used in this step.

From the generated caption, the story will be created.

3.1 Dataset

The whole work is divided in two parts. For the image caption generation, the dataset used is the **Flickr 8k** dataset which is a new benchmark collection for sentence-based image description and search, consisting of 8,000 images that are each paired with five different

captions which provide clear descriptions of the image. The images were chosen from six different Flickr groups, and tend not to contain any well-known people or locations, but were manually selected to depict a variety of scenes and situations.

For the story generation part, the dataset contains various files which have various sci-fi, scary and humorous stories. Web scrapping is also used to scrape out some of the famous authors like Sir Arthur Conan Doyle's work. All the text files containing stories together make the text corpus.

3.2 Architecture

Any image is passed to the image captioning model where VGG-16 layers are used to extract features from the image and LSTM model for generating captions from those features. The generated caption is then passed to a GPT-2 124M model trained on various genres of story corpus and a story from the image is obtained as final output.

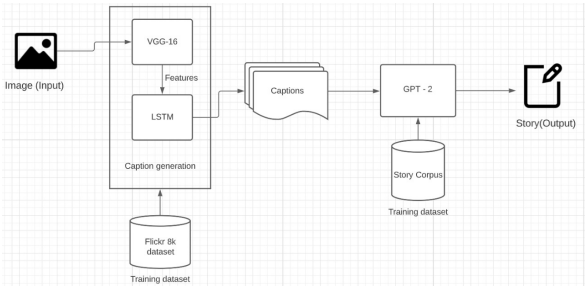

Figure 1: System Architecture

There are three genres in which story can be generated. The story corpus contains scare, sci-fi and humorous stories on which GPT-2 is trained.

3.3 Image captioning

3.3.1 Extracting image features

For this part, VGG-16 model has been restructured. Since the fully connected layer of the VGG-16 model for predictions are not needed, only the previous layers to extract the features from the image have been kept and the last two layers have been excluded from the model. A dictionary stores the image id and features extracted by the model. First the image is converted to a numpy array and then reshaped for the model in order to extract the features. After pre-processing the data is passed to the model and final features data of all the images in our data is obtained in the dictionary.

3.3.2 Working with the caption data

In the dataset we have a text file containing the captions for every image. This data will also be used in the caption generation process. Each image id has been mapped to the captions in a dictionary. The pre-processing and cleaning like converting to lowercase, removing digits, special characters and whitespaces are removed. After this, all the captions are stored in a single list is used for tokenisation process to create a vocabulary of unique words.

3.3.3 Data generator function

Given the large amount of data in the flickr8k dataset, to prevent system crashes in training process, a data generator function is used which gets the data in small batches and load into the model. By doing this, one caption is taken, complete processing is done for that which includes encoding and padding and the data is stored. Only after this a new caption is taken.

3.3.4 Caption Model Creation

This model will use the output from the restructured VGG-16, so the shape of input for this model is kept similar to the shape of output produced by the feature extraction model. For image processing, three layers are present which are input layer, dropout layer and dense layer. For the text part, 4 layers are there which are input layer, embedding layer, dropout layer and LSTM layer. Both the text and image part is concatenated into a single dense layer on which categorical one hot encoding is done. The word with a higher probability will be taken. After compilation, this model has been trained on the image features from flickr8k dataset and the vocabulary created in previous steps for 20 epochs.

3.3.5 Caption generation

The words that we will get from the model will be indexes from the tokenizer so using

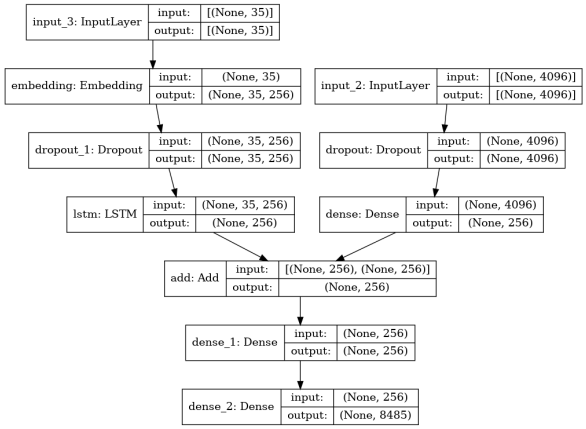

Figure 2: VGG16 image captioning model

the index, the word will be fetched for the caption. A start tag is used to start the caption, after which model is given image features to predict the caption. From the model, indexes of words having higher probability to be there in the caption are returned which is converted back to actual words.

3.4 Story Generation

In our project we have fine-tuned the gpt2-124M model for 1000 steps with three genre of story corpus which include horror, sci-fi and humor, each in different sessions of the gpt2 model. Through this 3 differently trained models of gpt2 were obtained for the three genres respectively. For the story generation process we passed the image caption generated from the caption model as prefix into the gpt2 model trained on the corpus on which we want the story to be focused. Length of the story can also be controlled by passing the number of words while calling the generate function. By using the gpt2 model in our project we are able to generate grammatically correct and coherent text with good narration, while also maintaining the initial idea of the story and making them enjoyable to read.

4 Results

After all the steps of proposed work and training the caption model for 20 epochs,it has been tested for accuracy of prediction of caption with BLUE-1 and BLUE-2 scores as metrics. The obtained values were 0.53006 for BLUE-1 the score and 0.305475 for BLUE-2 which are both, respectable scores. Here the GPT2 takes the caption generated from the VGG16 encoder decoder model as the first line of the story and then builds a stroy around it. Any of the three genres can be selected for creating the story.

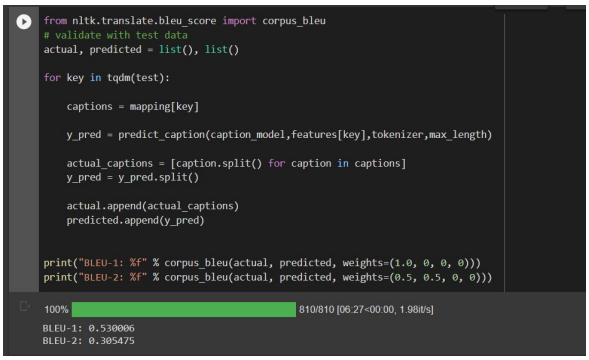

Figure 3: Obtained BLUE-1 and BLUE-2 scores.

When the image path were passed to the function made for caption generation, the system was able to understand the content of image provide and generate the captions for it accordingly.

Figure 4: two girls are playing in sand

However,in some cases the system falsely detects a man as woman because of more number of images of woman in the flickr8k dataset which was used to train the model.

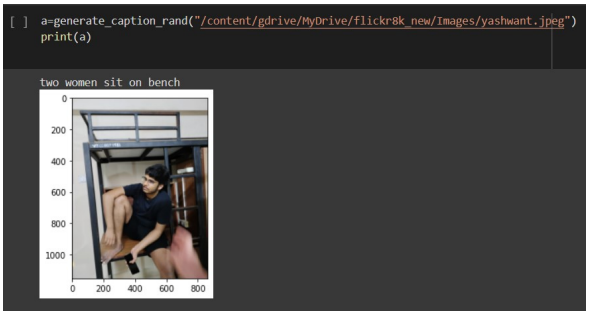

Figure 5: two women sit on a bench

These captions were passed to the fine tuned GPT2-124M model for the next part of the project that is story generation, through three different functions made for each genre that the model was fine tuned on. The caption served as prefix for the story and text was generated to continue the caption in form of a story.

Figure 6: A scary themed story generated from the caption "two girls are playing in sand".

Story: *two girls are playing in the sand when they find them. They eat popcorn, melt down ice, and go for it. The old woman's name is Daisy. She lives on a farm about eight miles away and is famous for her dancing. She and her husband live in a house on a farm about eight miles away and the road leads back their way. The boy runs up on the farm and beats his dog to make it quiet. "That's a nice dog," he says. "I'll put it in the soup, and we'll have it for supper." That night the boy plays the piano. "What do you come for?" he asks. "To play chess and go to the movies." To make it more cheerful, the boy adds a little dingo: "A big mean dog," says the wife. "I'll put it in the soup, and we'll have it for supper." That night the boy plays the piano. "What do you come for?" he asks. "To play chess and go to the movies."*

In a similar way, we can generate humorous story with the instance of GPT2 that was fine-tuned on the humour genre of story corpus by passing the image caption generated into the function specified for story generation of humour genre.

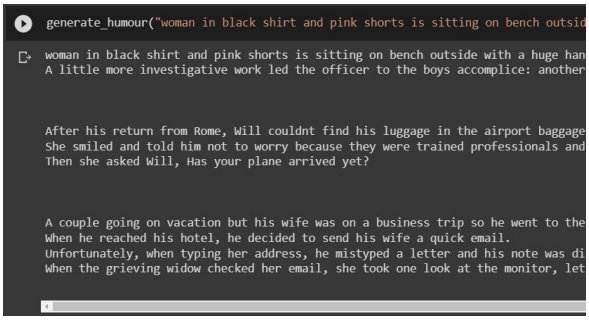

Figure 7: A humour themed story generated from the caption "women in black shirt and pink shorts is sitting on bench outside".

Story: *woman in black shirt and pink shorts is sitting on bench outside with a huge hand painted sign which said Radar Trap Ahead.A little more investigative work led the officer to the boys accomplice: another boy about 100 yards beyond the radar trap with a sign reading TIPS and a bucket at his feet full of change.After his return from Rome, Will couldn't find his luggage in the airport baggage area. He went to the lost luggage office and told the woman there that his bags hadn't shown up on the carousel.She smiled and told him not to worry because they were trained professionals and he was in good hands.Then she asked Will, Has your plane arrived yet?A couple going on vacation but his wife was on a business trip so he went to the destination first and his wife would meet him the next day.When he reached his hotel, he decided to send his wife a quick email.Unfortunately, when typing her address, he mistyped a letter and his note was directed instead to an elderly preachers wife whose husband had passed away only the day before.When the grieving widow checked her email, she took one look at the monitor, let out a piercing scream, and fell to the floor in a dead faint.*

Similarly the GPT2 instance that was trained on the science fiction story corpus has been used in the function specified for generation of science fiction themed story.The length of the story can be specified in the function, which allows the flexibility to generate the story of desired genre and length. Image caption has to be passed to the function and story will be generated by the model.When passing the same caption into the function for different genre, the idea and context derived from the caption changed completely to match with the

new genre and a new story was generated on the it. Hence, the system is capable enough to derive multiple context from a single image and caption as required which can be seen in Figure 6 and Figure 9.

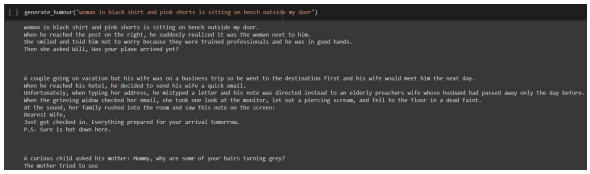

Figure 8: A story of science function genre generated with caption "two girls are playing in the sand".

Story: *two girls are playing in sand this sand." "Why sand?" Larkin demanded. "A sand-girl beach trip always ends in disaster. Nobody likes being swamped." "But not here." "Why? Why? Why – " Tee jay said nothing more about it. Then she turned to Steve. "If this is as serious as bragging, then we can hardly blame him. After all, he was a boy again. And now he's a boy again!" The lights blinked out, and the darkness was gone. The only sound was the restless cursing of the ledgers, and the grinding of door bodies. Steve ran for his life, to his knees. He ran again, and found himself walking aimlessly. Overhead he could hear the grinding of an air-foil, and the crashing of the suddenly come alive landing. Long slavering feet he failed to control. A scream came, then a fat thump, and a metallic thump, and another. He could see the retreating frontiers of the world from a thousand miles away, no longer pursu'd by the great Monsters, but by some random imposter who had some magic formation which he sent pointed the way.*

5 Conclusion

In this paper, we suggest a way of creating story from an image with the help of latest deep learning models. This is an attempt for creation of simple but better story generators. With the help of VGG16, LSTM and GPT2, we are able to generate stories from the given image. The features are extracted by the VGG16 model and the LSTM encoder decoder model is trained for generating captions. As the dataset used was Flickr8K, it didn't take much time. But to improve its accuracy, the larger datasets like MS COCO can be used. For the image captioning purpose we also tested the VGG19 and Inception V3 but VGG16 gave the best results on our dataset. The GPT 2 is fine tuned on 3 genres: humor, sci-fi and scary. The stroy corpus for this is created by the web scraping of some famous author's works. We believe this is a worthy direction for the advance of visual storytelling.

6 Future works

One of the future work can be grammar and spelling check. This module can be added to make the story grammatically correct. Furthermore, the image captioning model can be trained on the large datasets like MS COCO for getting the better results. For the story corpus, large dataset can be taken but the time taken by GPT and other deep learning model can be very large.

Also the text-to-speech functionality that can be integrated into the project which will make it more easy to use, even to the kids and add into the real world application part of it.

Since the passing of caption and calling different functions for different genre of story is a complex task especially for people who are out of computer science domain, a flask application can be used to design a front end for the project which will again make it very simple and easy for any person to use and access.

References

Rakesh M. Verma Avisha Das. 2020. Can machines tell stories? a comparative study of deep neural language models and metrics. *IEEE Access, Open Access Journal*.

NenavathVenkateshNaik Allu Siva Kishore Redd B Venkat Raman, Nagaratna P Hegde. 2019. A deep learning for the generation of textual story corresponding to a sequence of images. *International Journal of Recent Technology and Engineering (IJRTE)*, 8.

Maruan Al-Shedivat Eric P. Xing Zhiting Hu Bowen Tan, Zichao Yang. 2021. Progressive generation of long text with pretrained language models. *arXiv:2006.15720 [cs.CL]*.

S. Jenne G. Jagfeld and N. Thang Vu. 2018. Sequence-to-sequence models for data-to-text natural language generation: Word- vs. character-based processing and output diversity. *arXiv:1810.04864*.

Hyeonjoon Moon Kyungbok Min, Minh Dang. 2021. Deep learning-based short story generation for an image using the encoder-decoder structure. *IEEE Access, Open Access Journal*.

Margaret Mitchell, Ting-Hao Huang, Francis Ferraro, and Ishan Misra. 2018. Proceedings of the first workshop on storytelling. In *Proceedings of the First Workshop on Storytelling*.

Abhijit Mishra-Mohak Sukhwani Anirban Laha Karthik Sankaranarayanan Parag Jain, Priyanka Agrawal. 2017. Story generation from sequence of independent short descriptions. *SIGKDD'17*.

Reid Swanson and Andrew S Gordon. 2012. Say anything: Using textual case-based reasoning to enable open-domain interactive storytelling. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):1–35.