Successive Image Generation from a Single Sentence

Amogh Parab[1], Ananya Malik[1], Arish Damania[1], Arnav Parekhji[1], and Pranit Bari[1]

[1] Dwarkadas J. Sanghvi College of Engineering, Mumbai, India

Abstract - Through various examples in history such as the early man's carving on caves, dependence on diagrammatic representations, the immense popularity of comic books we have seen that vision has a higher reach in communication than written words. In this paper, we analyse and propose a new task of transfer of information from text to image synthesis. Through this paper we aim to generate a story from a single sentence and convert our generated story into a sequence of images. We plan to use state of the art technology to implement this task. With the advent of Generative Adversarial Networks text to image synthesis have found a new awakening. We plan to take this task a step further, in order to automate the entire process. Our system generates a multi-lined story given a single sentence using a deep neural network. This story is then fed into our networks of multiple stage GANs inorder to produce a photorealistic image sequence.

Keyword - Deep Learning, Generative Adversarial Networks, Natural Language Processing, Computer Vision, LSTM, CNN, RNN, GRU, GPT-2.

1. INTRODUCTION

Tell me and I will forget, Teach me and I will remember, Show me and I will learn.

Imagining a visual story from a text description is the ultimate test of semantic, visual, and spatial world knowledge. Artificial Intelligence's arrival can be effectively used to create expert systems that will exhibit intelligent behaviour, provide solutions to complicated problems, and further help to develop stimulations equivalent to human intelligence within machines.With the advent of technologies within the NLP (Natural Language Processing) and CV (Computer Vision) communities, there exists a strong initiative to promote vision and language research. Until now this field has been heavily populated with research focused on the vision to language stream that has promoted visual perception to understanding diverse linguistic representations. With our work we intend to focus on the language to vision path which will allow vision understanding and offer new avenues of visual content creation. We intend to be able to formulate this tool to communicate and express stories, and correspondingly data better.

We propose generating a story and it's visualisation in the form of a sequence of images with a single textual input. Our system uses a transformer to generate a multi-lined story from the input. This story is then fed into a network of multiple two staged StackGANs to develop a sequence of images.

2. LITERATURE SURVEY

We have used Yitong Li et al [1] paper as our base paper, as it aims to solve a similar problem, that of story visualisation. The paper suggests that given a multi sentence paragraph, the need to generate an image storyline that must accurately represent the story, scene and context of the given input and focuses less on the continuity of the motion between images. In order to maintain the intertextuality between the images, the paper proposes the use of a Deep Context Encoder that identifies the context. It makes the use of a sequence of GANs which are modified two accommodate two discriminators at the image level and the story level. The main aim of the paper is to be able to generate images on a multi sentence paragraph input that comprises the entire story. The uniqueness of this paper lies in the initial fact that it proposes an entirely new task of Story Visualisation. The challenges faced by this system is to maintain consistency in the images as each sentence might not depict the entirety of the scene and to keep the layout of the scene similar, if not same throughout the entire sequence.

Zhang et al. [2] proposes the StackGAN to convert a single line of text to a singular image. This model uses a stacked architecture of GANs. They propose two techniques to implement the same - using a two stage architecture and multi-stage architecture. The two stage architecture is a conditional architecture using two GANs, the multi stage architecture advocates use for conditional and unconditional generative tasks by suggesting a tree like structure in the Stage II. The proposal of StackGANs [2] was a pathbreaking move as it allowed visuals of higher resolution to be generated with conditional GANs. The conditioning augment module allows conditional smoothening, that is it focuses on the details of the image as well. The two layer architecture is useful in our application of Story Visualisation, as it provides a two level discriminator check to ensure that the image doesn't lose the contextual consistency. Unlike the implementation in StackGAN v1 where both level 1 and level 2 discriminators are fed with the entire text embedding in order to ensure that the background text is not lost, our application demands a contextual support. Hence we are able to modify the stack based architecture proposed to ensure that the sentence is fed in discriminator 1 maintains the

textual intricacy and the stage 2 discriminator can yield to increasing a higher resolution output.

Ouyang et al. [3] have posed a generation of realistic images from text description by cumulating two neural network structures of LSTM and GAN. Their approach surmounted the problem of creating high resolution images from text descriptions, which required multiple training of models. This new network thus helps to generate a sequence of images depicting evolution of image based on the sentence. This paper is distinctive from the other image generating ones because it blends two separate neural networks LSTM and Conditional GAN, thereby giving more detailed images. The problem solved by them was to create a sequence of images on a text description. Our model does the same thing, but we are to produce output of a sequence of images on the generated story. We are proposing to replace the LSTM in this paper with the GPT-2 transformer to help maintain more context along the generated sentences. The main challenge will be to not only check the generated image is acceptable but withal ascertain that the background story stays intact and does not divert from the actual scene. To do so we are using stacked GANs.

Clark et al. [4] represents an alternate approach to language modelling for neural text generation. It uses the concept of context to maintain consistency throughout the generated story. This context is maintained by the use of entities, whose representations are stored as vectors and updated regularly.This paper works on combining entity recognition with neural text generation. This is what sets this paper apart from other basic text generators using LSTMs or GRUs. It uses an additional vector for the entities which are then combined with the context of the previous sentence. The model provides better results than other models which are verified by Mention Generation and Pairwise Sentence Selection in addition to Human Evaluation.

Pawade et al. [5] covers automatic text generation which would be required to create more text from a given sentence. Their paper aims to generate a new story based on a series of inputted stories. It makes use of Recurrent Neural Networks - Long Short-Term Memory (RNN-LSTM) to generate a new sequence of text, that is, it forms sentences based on a provided input text. The generated stories received an accuracy rating of 63% by humans rating the stories on the basis of grammar, event linkage, interest level and uniqueness. However, the limitations of the system were that the stories generated were monotonous and repetitive.

3. PROPOSED ARCHITECTURE WITH MODULAR DESCRIPTION

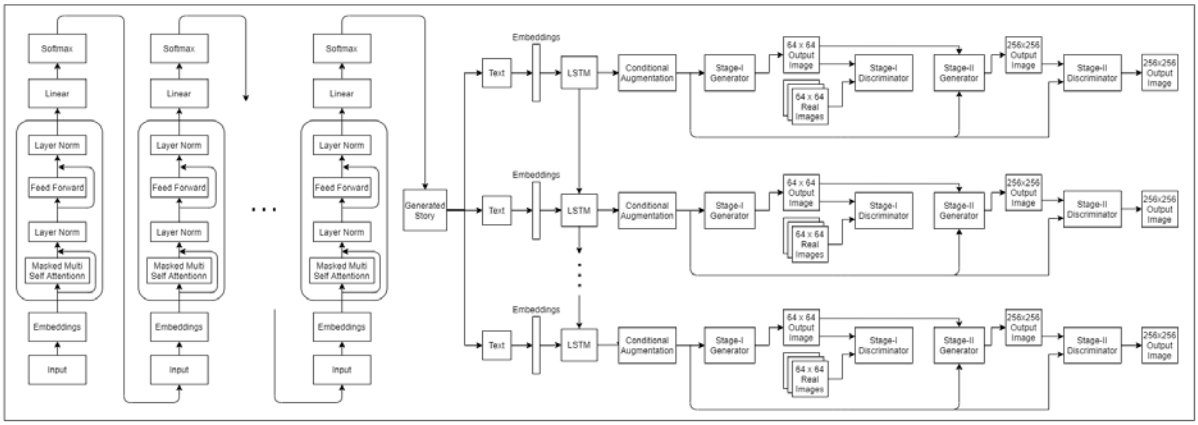

Fig. 3.1. Illustration of proposed model.

Fig 3.1 is the complete system architecture of the proposed model. It comprises GPT-2 decoder and StackGAN model which will be explained in detail in the following sections.

The user inputs a sentence, which will be used to generate a story. The sentence is mapped to vectors which are passed on to the GPT-2 decoder. The output of the decoder is passed through the convolution layer and then a softmax function is applied on it to predict the next word. This procedure continues till a story is generated. After the story is generated, it is passed on to the StackGANs model which will be used to generate images.

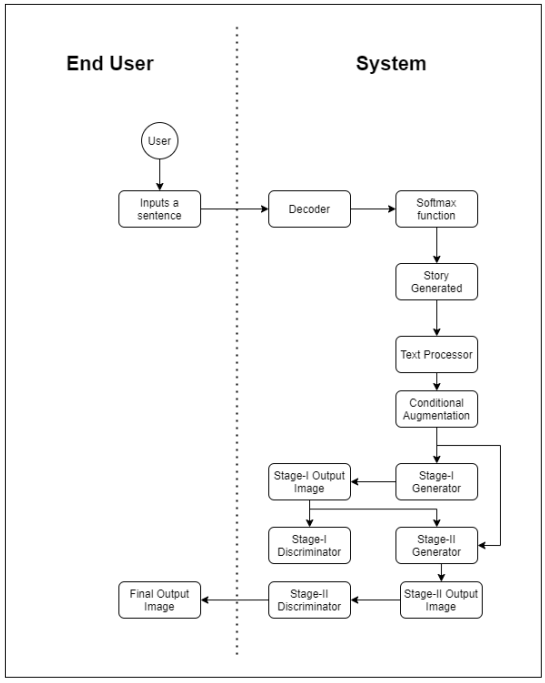

Fig. 3.2. Illustration of system flow.

The generated story is passed on through a text processor which keeps the story's background intact. A condition is augmented to the output of the processor and then is passed as an input to the StackGAN Stage-I generator. The generator generates an output which is verified with the Stage-I discriminator. These outputs are then passed on as input to the Stage-II generator along with the processed text. The output of the stage-II generator is validated with the stage-II discriminator and then the final image is returned. This process continues till the story ends. After the end of the story, sequences of images describing the story are presented to the user.

4. EXPERIMENTATION AND RESULT

Image Generation Module: Since we employ StackGAN to generate images from text, we experimented on the Bird Dataset by Caltech-UCSD 2010-11 dataset. Our method aims to use the two GAN architecture, like the StackGANS++. The first level, or the first GAN structure aims to model the scene, i.e. sketch the basic shape and colour, leading to the production of a low resolution image. In order to obtain a more detailed image in higher resolution, we now use the previously obtained image and sentence as input and generate the higher resolution image.

Fig. 4.1 Generated images using StackGAN.

Text Generation Module: We are using the pre trained GPT-2 transformer to perform the task of language generation. Transfer learning is used to fine tune the transformer architecture on our dataset to generate the new sentences. GPT2 is a decoder only transformer. It makes use of the concept "auto-regression". The GPT2 transformer uses masked self attention. We made use of a Flintstones Dataset comprising 25,184 video clips of the animated TV show The Flintstones.

```
Input text:
Fred is standing in a room. He is looking at something intently.

Generated story:

Fred is standing in a room. He is looking at something intently. In
the middle of the room a man sits down, and looks at the light that
shines on it. It is a single, blue-yellow. It is not the light of a
car, it is a light of his. He takes out his camera.
```

Fig. 4.2. Generated Story using GPT-2 transformer.

After working separately on the two modules, we have started integrating them. Experimenting on Flintstones dataset. we expect the following outputs.

```
Input text:
Dino is in the living room.

Generated Story:

Dino is in the living room. He lays on a rock chair in front of the
television. He looks up and into the other room. He jumps off of the
chair and lays on the floor.
```

Fig. 4.3. Expected Text Output on Flintstones Dataset.

We will be using the PORORO-SV dataset which contains a video clip with a written description. A compilation of 40 video clips form a story, and in total there are around 16,000 video clips of 408 movie stories. Thus the entire dataset comprises 15,336 description-image pairs. We have preprocessed the data and is ready to use for training.

Fig. 4.4. Expected Image Output on Flintstones Dataset.

```
0----------------------------------------------------------
 pororo is trapped on an ice pillar. the ice pillar is very high and in the middle of the air.
 pororo is trapped on an ice pillar with pororo sled. pororo is scared and doesn't know what to do.
 eddy is telling pororo not to move. eddy is leaving to call other friends.
 eddy and friends hurriedly came back. pororo and eddy  are standing on a steep ice cliff.
 eddy is asking rody to help. rody is stretching rody arms and legs to reach the ice pillar.

1----------------------------------------------------------
 Pororo says be quiet to Crong. Crong holds a book out to Pororo.
 Pororo asks what the book is while Pororo says that Pororo is sleepy.
 Even if Pororo says Pororo is sleepy Pororo takes and opens the book to read it.
 Pororo starts to read a book Crong sits beside Pororo.
 Pororo skips the middle of the book so as to finish it quickly.

2----------------------------------------------------------
 poby grabs a cookie and give a try.
 harry eats a cookie and finds out that something is wrong.
 pororo's face turns red. and crong drops cookies.
 petty and loopy look at each other and wonder why.
 pororo suggests another cookie to crong. and crong denies.

3----------------------------------------------------------
 pororo found loopy easily. pororo is surprised to find loopy so easily. loopy drooped.
 loopy is telling pororo and petty is okay. loopy dusting the earth off loopy clothes.
 pororo is happy that poby found loopy. loopy is feeling sorry for loopy.
 pororo is happy that poby found loopy. but loopy is feeling sorry for loopy.
 pororo is entering into a large tree. harry and petty are watching pororo entering. harry and petty are on the top of the
 staircase.
```

Fig. 4.5. Generated Story.

Fig. 4.6. Stage I Low Resolution Image

Fig. 4.7. Stage II High Resolution Image

Fig. 4.8. Ground Truth

5. CONCLUSION

We thus have been able to study story generation and visualisation as one concentrated task. The proposed model deals with the task jointly by generating the story from a single sentence and simultaneously generating the images for visualisation. Unlike other systems, this proposed solution allows the generation of the story and images under one system. Generation of text using GPT2 increases the controllability of the text and using StackGAN allows accurate depiction and photo-realistic images. The sequential generation model followed here allows context to be maintained. Using statistical and human evaluation factors we are able to determine that the proposed system is an improvement over current baseline models.

Today, story generation and visualisation are viewed as separate tasks, primarily because the maintenance of context across the streamlined procedure is tough. Thus, by developing more state-of-the-art context-maintenance and embedding systems we can greatly improve the efficacy of this task. By expanding the dataset beyond the realms of the Pororo cartoon we will be able to include stories that are applicable to current events. This will allow the applications of such a system to extend to media companies for the generation of stock images and in the publishing of children's books and comics.

REFERENCES

[1] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, Jianfeng Gao. StoryGAN: A Sequential Conditional GAN for Story Visualization. 2019

[2] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, Dimitris Metaxas. StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks. 2018

[3] Xu Ouyang, Xi Zhang, Di Ma, Gady Agam. Generating Image Sequence from Description with LSTM Conditional GAN. 2018

[4] Elizabeth Clark, Yangfeng Ji, Noah A. Smith. Neural Text Generation in Stories Using Entity Representations as Context. 2018

[5] Dipti Pawade, Avani Sakhapara, Mansi Jain, Neha Jain, Krushi Gada. Story Scrambler - Automatic Text Generation Using Word Level RNN-LSTM. 2018

[6] K.-M. Kim, M.-O. Heo, S.-H. Choi, and B.-T. Zhang. Deepstory: Video story qa by deep embedded memory networks. 2017.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. Attention Is All You Need. 2017

[8] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, Aniruddha Kembhavi. Imagine This! Scripts to Compositions to Videos. 2018

[9] Van-Khanh Tran, Le-Minh Nguyen and Satoshi Tojo. Neural-based Natural Language Generation in Dialogue using RNN Encoder-Decoder with Semantic Aggregation. 2017

[10] Xu Tao, Pengchuan Zhang, Qiuyuan Huang, Zhang Han, and Xiaodong He. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. 2017

[11] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, and Honglak Lee. Generative Adversarial Text to Image Synthesis. In International Conference on International Conference on Machine Learning. 2016

[12] Zizhao Zhang, Yuanpu Xie, and Yang Lin. Photographic Text-to-Image Synthesis with a Hierarchically-nested Adversarial Network. 2018

[13] Fuwen Tan, Song Feng, and Vicente Ordonez. 2018. Text2Scene: Generating Abstract Scenes from Textual Descriptions. 2018

[14] Bowen Tan, Zichao Yang, Maruan Al-Shedivat, Eric P. Xing, Zhiting Hu. Progressive Generation of Long Text with Pretrained Language Models. 2021

[15] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. 2019

[16] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. 2018

[17] Tingting Qiao, Jing Zhang, Duanqing Xu, Dacheng Tao. MirrorGAN: Learning Text-to-image Generation by Redescription. 2019