KANNADA TEXT-TO-IMAGE SYNTHESIS

A report submitted to

RAMAIAH INSTITUTE OF TECHNOLOGY

Bengaluru

ISP SENIOR PROJECT

as partial fulfillment of the requirement for the award of degree of

Bachelor of Engineering (B.E) in Information Science and Engineering

By

Aston Glen Noronha	1MS18IS021
Madhura J Shet	1MS18IS050
Meghna Nair	1MS18IS051
Shruthi Iyer	1MS18IS101

Under the Guidance of
Dr. Pushpalatha M.N.
Assistant Professor

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING
RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, Affiliated to VTU)
BANGALORE - 54

JULY 2022

Department of Information Science and Engineering

Ramaiah Institute of Technology

Bengaluru - 54

CERTIFICATE

This is to certify that Aston Glen Noronha (USN- 1MS18IS021) , Madhura J Shet (USN- 1MS18IS050), Maghna Nair (USN- 1MS18IS051) AND Shruthi Iyer (USN- 1MS18IS101) who were working for their ISP, SENIOR PROJECT under my guidance, have completed the work as per my satisfaction with the topic Kannada Text-to-Image Synthesis. To the best of my understanding the work to be submitted in dissertation does not contain any work, which has been previously carried out by others and submitted by the candidates for themselves for the award of any degree anywhere.

Name and Signature of the Guide Signature of the HOD Signature of the Principal

External Viva

Name of the examiners Signature with date

1.
2.

DECLARATION

We hereby declare that the entire work embodied in this ISP SENIOR PROJECT report has been carried out by us at Ramaiah Institute of Technology under the supervision of Dr. Pushpalatha M.N. This project report has not been submitted in part or full for the award of any diploma or degree of this or any other University.

Aston Glen Noronha	1MS18IS021
Madhura J Shet	1MS18IS050
Meghna Nair	1MS18IS051
Shruthi Iyer	1MS18IS101

Place: Bangalore
Date:

ABSTRACT

The automatic synthesis of realistic images from text is quite intriguing and highly essential. However in the past few years powerful neural networks have been developed for the text feature representation. There has also been a massive increase in improvement in convolutional networks, which has helped to provide semantic information which is useful while transferring the content to another style. The model could give us outputs that represent images described in the text. Stack Generative Adversarial Network(Stack-GAN) has been generating images which belong to specific categories such as human faces, animals, albums etc.

GAN's are an accost to generative modeling using deep convolutional methods. Generative modeling is unsupervised learning that can discover and learn the patterns for the input in a manner that the model can generate an output which can be obtained from the dataset. The primary role of the discriminator is that it tries to classify the real or fake images. The models are trained by the zero sum game that is adversarial. GAN's are an interesting and upcoming field since they can generate realistic images across various topics.

Some of the challenges include processing time. To run a generative adversarial network model we need an efficient GPU which could handle the large data. To generate high-quality images, we require a GPU which has a high computational power. It is also extremely difficult to generate large and fine grained images.

ACKNOWLEDGMENT

We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project. We would like to express our profound gratitude to the Management and Dr. N.V.R Naidu, Principal, M.S.R.I.T, Bengaluru for providing us with the opportunity to explore our potential.

We extend our heartfelt gratitude to our beloved Dr. Sanjay H, HOD, Information Science and Engineering, for constant support and guidance.

We wholeheartedly thank our project guide Dr. Pushpalatha M.N., for providing us with the confidence and strength to overcome every obstacle at each step of the project and inspiring us to the best of our potential. We also thank her for her constant guidance, direction and insight during the project.

Finally, we would like to express sincere gratitude to all the teaching and non-teaching faculty of ISE Department, our beloved parents, seniors and my dear friends for their constant support during the course of work.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation and Scope

Images convey information universally and they can also be understood easily by any individual. They can also be understood by illiterate people. People can be addressed using images rather than spoken language. This helps make communication between people easier.Nowadays we can't find a picture for every situation. People who are not able to read can be assisted using images. Since there is also a vast language barrier. Individuals can also communicate using images. Some of the major applications include 3D object generation, photograph editing and creating image datasets.

Finding images for every scenario is quite hard and if a mental picture has been described it's quite difficult to get the exact picture hence Machine Learning and Deep learning has made this quite possible.Whenever we read a story we draw images pertaining to it. Hence there is an intricate relation between the visual world and languages. In terms of memory spatial navigation and reasoning mental imagery plays a very crucial role. Description of text is very diverse with respect to choice of words.

Text to image synthesis aims to build photo realistic image generation. This is done using a generative adversarial network (GAN). GAN is a model in which there are two neural networks. These two networks tend to compete with each other. GAN is about creating a portrait or composing music. GAN's are generative models that create new data which tend to resemble the data we have trained. A typical example is when we create an image similar to faces of people even if they don't exist. GAN models have several applications of unsupervised learning representation, denoising and semi supervised learning. Even though there are several GAN models, they have a quantitative evaluation.

Inspection that is done visually is extremely time consuming and cannot capture several characteristics. This is an important factor for unsupervised learning. If we need better GAN models we need to design it in such a manner that it should overcome the quantitative limitations. Recently, several GAN's evaluation benchmarks have both been introduced with the emergence of new models.

Some of the GAN model applications are:

- Using image datasets for generating examples.
- Generating photographs that are realistic images
- Generating photographs that are realistic
- Generating cartoon characters
- Translation of image - to - image
- Translation of text - to - image
- Face frontal view generation
- Generating new human pose

- Editing photographs
- Aging of face
- Blending of photos
- In painting of photos
- Translation of clothing
- Prediction of video
- Generation Of 3D objects

1.2 Issues and Challenges

- Processing time: In order to run a Generative Adversarial Network Model, a CPU along with an efficient GPU is required which can handle data since image generation can be tedious.
- Resolution: A GPU memory is required while generating images of high quality. However an unstable GAN model will be made if we take smaller batch size. Therefore it is difficult to get fine detailed and large images.
- Accuracy of conversion: Few GAN models do not accurately generate the image which matches the input text. This usually happens when the GAN model does not have a detailed and accurate dataset which consists of the required data to generate the right image which matches the text description.

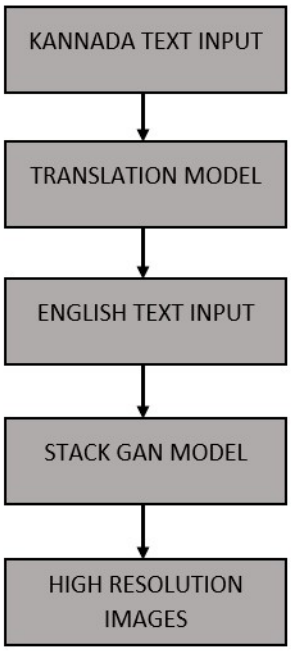

Figure 1.1: Proposed Model

1.3 Problem Statement

In this project a framework is proposed to formulate image generation conditioned on the text input. Converting regional language text descriptions into images using Stack Generative Adversarial Network(Stack-GAN) and Gated Recurrent Unit (GRU).

Text-to-Image generation is a landmark task in multi-modal machine learning and it have major applications in creating image datasets, 3D Object generation, generate photographs, photograph editing and so on.

The objective is to develop a model which helps in generating photo realistic images of high resolution which are consistent semantically with the input text description.

1.4 Proposed Model

This section presents the methodology used for the synthesis of high-quality images from Kannada text descriptions. The first step in the model is the data collection phase. This data is sent to the translation model to translate the data from Kannada text to English. The translation model consists of encoder and decoder Gated Recurrent Unit (GRU) to perform the translation. The English text data generated is sent as input to the Stack-GAN model consisting of two stages. In Stage-1, the images generated are of low resolution 64x64 images trying to capture the basic features described in the text description. In Stage-2, the defects in the low-resolution image from Stage-1 are corrected and it reads the text description again to complete the the details of the object, producing a high-resolution image. Figure 1.1, shows the proposed model.

1.5 Organization of the Report

In order to explain the developed system, the following sections are covered:

Chapter 2 provides Literature Review which describes the study of the existing systems and techniques taken into account prior to development of the proposed system.

Chapter 3 provides System Design and Analysis which describes a detailed walk through of the software engineering methodology adopted to implement the model, an overview of the system and the various modules incorporated into the system.

Chapter 4 provides the Implementation which describes a deeper insight into the working of the model. The various modules and their interactions are depicted using relevant descriptive diagrams.

Chapter 5 provides Experiment and Results which describes the results obtained and comparison between existing models.

Chapter 6 provides Conclusion about the results obtained after successfully running the model and Future Scope of the model is highlighted.

Chapter 2

Literature Review

In this paper [5] A text to image generation (T2I) approach tries to produce photorealistic images that are semantically coherent with the text descriptions. Existing T2I models have made considerable strides thanks to recent developments in generative adversarial networks (GANs). However, a detailed examination of their produced photos reveals two significant shortcomings.:

- The local semantics are ignored as the condition batch normalisation procedures are applied uniformly to all of the picture feature maps;
- Semantic spatial aware GAN is put forth to learn better representations for image creation and so that the text encoder can be fixed throughout training and taught along with the image generator.This GAN is trained from beginning to end so that the text encoder can take advantage of superior text data.

This paper [3] presents a novel GAN model, that is, ControlGAN. This GAN model effectively generates high resolution images while also controlling parts of image generation based on the text descriptions. In order to obtain this, word level spatial and channel wise attention driven generator, word level discriminator was introduced.The Generator detangles various visual features, allowing the model to concentrate on developing and altering sub-areas of the images corresponding to the key terms of description.It is proposed that the discriminator provides minute supervisory feedback by equating phrases with image features, thereby enabling effective training of a generator.

[8], proposed a novel Generative Adversarial Network (AttnGAN) that synthesizes minute details at all of the regions of the image by focusing to the required words in the text description and generates high quality image through a multi-stage process. A Deep Attentional Multimodal Similarity Model (DAMSM) is also proposed.Using attention mechanism, DAMSM additionally provides a fine-grained image-text matching loss for training the generator of the AttnGAN and computes the resemblance between the sentence and the created image using both the fine-grained word level information and the broader sentence level information..

[10], proposed Stack Generative Adversarial Networks (StackGAN) with conditioning augmentation to generate 256×256 photo-realistic images constrained on text descriptions. Stage-1 GAN gives us the object based on the text descriptions' basic color and shape constraints. Stage -2 GAN adds more results to stage-1 and corrects the results of stage-1 as well. This gives us high resolution images which has better quality. The text-to-image generative models that exists this method gives us high resolution images with more details.

To improve the performance of previous text-to-image models, this [9] proposed the contrastive learning method. Synthetic pictures' quality and semantic consistency are improved via contrastive learning. Contrastive loss was employed in the pretraining stage to acquire the consistent textual representations for the captions that accompanied the same image. To improve the consistency of the images generated from the captions, the contrastive learning method was used in the second stage. The framework is evaluated on two baselines, Attentional Generative Adversarial Networks (AttnGAN) and Dynamic Memory Generative Adversarial Networks (DM-GAN), on dataset CUB and COCO respectively. This method improves the FID score by 30% and by 22% over AttnGAN and DM-GAN respectively.

[6] proposed a workflow for developing images based on "machine-generated" descriptions. This is critical when captions for a specific domain are unavailable. As a result, the proposed solution entails using a generic dataset of captions, like COCO dataset, to enable the image captioning module to generate captions for a specific domain. To accomplish this, a pipeline comprised of an Image Captioning Module and a GAN Module was built. To generate multiple captions for the input image, the Captioning Module is trained on a generic captioned dataset. After that, true images are fed into the Trained Image Captioning Module, which generates multiple captions for each image. The generated captions are then fed to the GAN Module, which learns to generate images based on the "machine-generated captions"..

[2] proposed a novel method that successfully tackles controllability, human perception, quality and resolution pivotal gaps, while attaining state-of the-art results in the task of text-to-image generation. Their method provides a new type of control complementary to text, enabling new-generation capabilities while improving structural consistency and quality. Furthermore, they propose explicit losses correlated with human preferences, significantly improving image quality, breaking the common resolution barrier, and thus producing results in a resolution of 512×512 pixels. Their method is comprised of an autoregressive transformer, where in addition to the conventional use of text and image tokens, we introduce implicit conditioning over optionally controlled scene tokens, derived from segmentation maps. During inference, the segmentation tokens are either generated independently by the transformer or extracted from an input image, providing freedom to impel additional constraints over the generated image. Contrary to the common use of segmentation for explicit conditioning as employed in many GAN-based methods, their segmentation tokens provide implicit conditioning in the sense that the image generated and image tokens are not constrained to use the segmentation information, as there is no loss tying them together. In practice, this contributes to the variety of samples generated by the model, producing diverse results constrained to the input segmentations.

In this paper [7], to bridge these developments in text and picture modelling and successfully translate visual notions from letters to pixels, they created a revolutionary deep architecture and GAN formulation. The main distinction of their work is that instead of using class labels, build requirements on text information. This is the first character-to-pixel architecture that can be differentiated from end to end. A manifold interpolated regularizer for the GAN generator is also included, which considerably enhances the quality of the samples that are generated, even those for CUB's held out zero shot categories. The approach used here is to develop a hybrid character-level convolutional recurrent neural network that can encode text features in order to train a convolutional neural generative adversarial network (DC-GAN).

In the work done by Cristian Bodnar [1],generating the image is trained on a condition vector which is applied to the GAN model. The additional inputs will teach the networks to adapt and change their parameters. A CNN processes the images for mapping, while a hybrid convolutional recurrent network converts the description. A common skip through vector assigns similar vectors to sentences with similar syntax and semantics. When embeddings are used inside convolutional networks, this property results in improved performance.

[4] This study set out to preserve other contents that are unrelated to the text while semantically editing certain portions of an image that corresponds to its respective text that defines what was intended. A unique generative adversarial network (ManiGAN) with two crucial components—the text-image affine combination module (ACM) and the detail correction module—was presented to accomplish the same (DCM). For efficient manipulation, the ACM chooses image regions pertinent to the text that is provided and then correlates those regions with semantic words that belong to those regions. In the meantime, it encrypts the original image features to aid in the reconstruction of text-unrelated elements. The DCM fills in absent data from the images generated and rectifies mismatched image properties. Numerous tests on the CUB and COCO datasets show that the suggested technique performs better than others.

Chapter 3

System Design and Analysis

3.1 Use Case Design

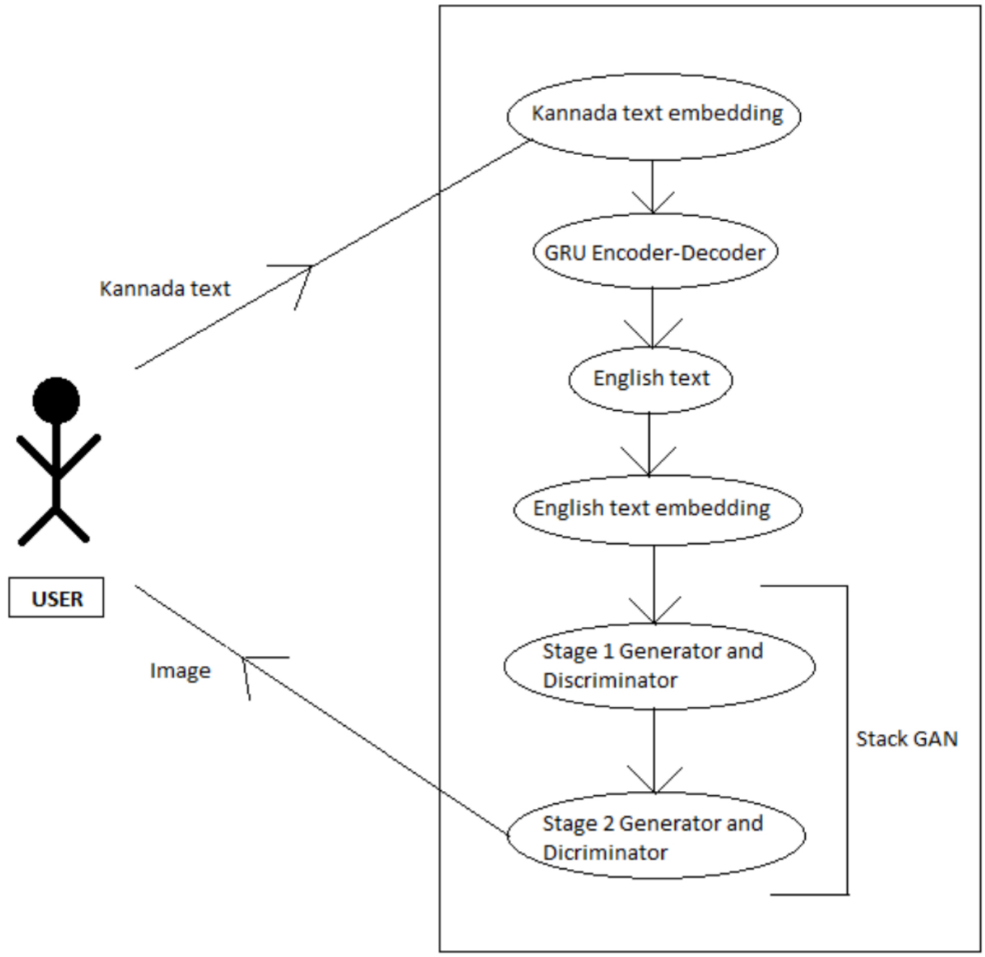

Figure 3.1: Use Case Diagram

Figure 3.1 shows the use case design of the proposed Kannada text-to-image synthesis model with one actor-User, someone who is fluent only in Kannada. The user entered Kannada text is vectorized and is converted to

English text using GRU Encoder-Decoder model. Stack GAN having two stages uses English text embeddings to generate a semantically consistent synthetic image back to the user.

3.2 Architecture

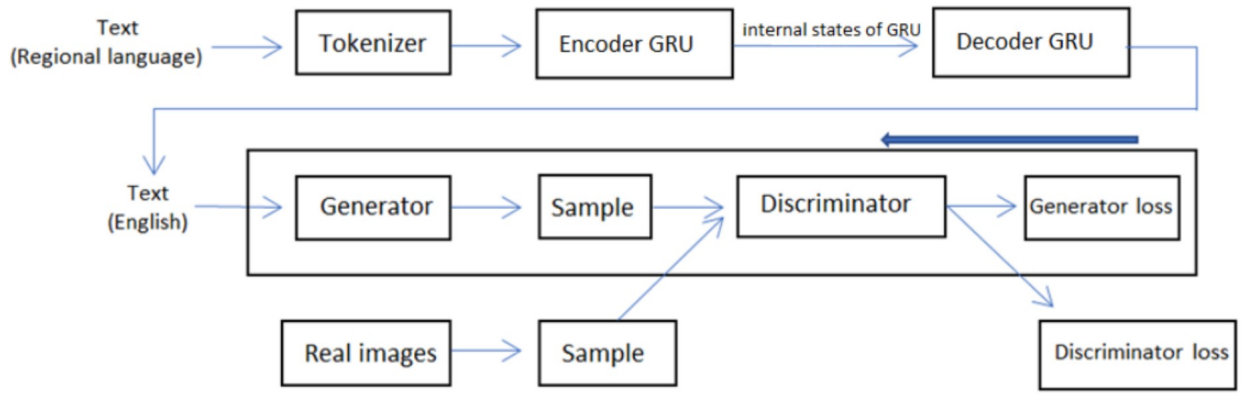

Figure 3.2: Architecture of the model

Kannada language description of the desired image is tokenized for better understanding of the text. The tokens then are mapped to its corresponding vector of real numbers which will be useful in finding next word predictions or for word similarities/semantics.This word embeddings are sent to the seq2seq model. As a recurrent unit Gated Recurrent Unit is being used. A GRU is used to compensate for a standard recurrent unit's short-term memory loss. An encoder in the GRU compresses an input sequence into a vector through its hidden states, which is then unfolded by a decoder network into a new sequence, which is the output of the translation model. The embeddings of the English text, that is the result of the translation model is sent to the GAN model.

The GAN model that is used is the Stack-GAN which comprises two stages;Stage 1 and 2.In both of the stages the network consists of text encoder, a Conditioning Augmentation network, a network of Generator and Discriminator.

In Stage 1, the model focuses on the color and shape of the object depending on the text embedding. The Generator of Stage 1 generates a synthetic image of low resolution 64x64, along with random noise. The image generated is forwarded to the Discriminator, where it tries to distinguish between the real data(low resolution 64x64 real image) and the data generated by Generator. A discriminator who incorrectly labels a real data instance as fake or a fake data instance as real is penalized by the discriminator loss. The Generator loss is to maximize the output of the discriminator for its fictitious instances. The discriminator updates its weights through the back propagation of the discriminator loss in the discriminator network. In the second stage, the input will be the synthetic image generated and its text embedding. The main concern in this stage is to fix any flaws of stage 1 and to generate a high resolution 256x256 photo-realistic image. Stage 2 Discriminator distinguished between the real image and the image generated by Stage 2 Generator. Again the Generator and Discriminator loss is calculated.

3.3 Tools and Technologies

- **Python 3.8-** A high level ,interpreted computer programming language.Python is used in this project .It consists of many in built functions which supports tensorflow.
- **Jupyter notebook-** Jupyter notebook is the application used to create and run the kannada text to image synthesis model.
- **Tensorflow-** Tensorflow is a open source library used for artificial intelligence , machine learning,neural networks.The tensorflow library is used to develop and train the Stack-GAN model.

- **Pytorch-** PyTorch is an open source framework on the torch library used to develop the translation model.
- **Pandas-** Pandas is a machine learning package which is used in this project for data analysis.
- **Keras-** Keras is used as an interface in the tensorflow library.
- **Numpy-** Numpy is used to calculate the mathematical and logical operations like the generative loss.
- **Pillow-** The pillow library has generative loss which is used for creating and processing images.

Chapter 4

Implementation

4.1 Dataset

The dataset most frequently used for fine-grained visual categorization tasks is Caltech-UCSD Birds-200-2011 (CUB-200-2011). It consists of 200 subcategories, 11788 images of birds. Each image in the dataset has 10 single sentence description in English. These descriptions are translated manually to Kannada which becomes the input to our translation model.

Figure 4.1: Example of Dataset

4.2 Translation Model

The unidirectional translation model uses Neural Machine Translation (NMT); Sequence-to-Sequence Model (seq2seq2), which is the state-of-art translation technology. Encoder-Decoder architecture and the Gated Recurrent Unit (GRU) as a Recurrent Neural Network (RNN) unit with optimization technique are used in the construction of the seq2seq model.

The user's entered Kannada text is converted to English using the translation model before being fed into the GAN model. The data for this translation model is in a text file containing captions of flowers in English and its Kannada translation separated by a space delimiter.To produce better results, data is pre-processed and then added to the encoder-decoder model. The encoder is a set of GRU cells where each receives one element from the input sequence, gathers information for that element, and propagates that data forward. In the final

hidden state of the GRU encoder, the encoder vector seeks to include all input element information in the final hidden state in order to aid the decoder in making precise predictions. This vector only is given as the first hidden state of the decoder. The decoder takes a word and context vector, which predicts the next word in the sequence and outputs a hidden state for use in the subsequent iteration. Up until an end of statement (EOS) token is produced, it keeps creating words. The decoder uses an attention mechanism to optimize the result: all the encoded input vectors are weighted, with the highest weight assigned to the most relevant vectors, to efficiently handle and enable the decoder to employ the pertinent portions of the input sequence.

Figure 4.2: Translation Model

4.3 Stack-GAN Model

The output from the translation model becomes the input to the GAN model. Like any other text data, to make use of it in machine learning algorithms, it needs to be converted into numbers or vectors. This text data is converted into text embeddings and a random noise is added to this embedding vector giving rise to a d-dimensional array. This process is called Conditioning Augmentation which yields more training pairs given a small number of image-text pairs, to encourage robustness and increase performance of the model. The reason why Conditioning Augmentation is used is to introduce randomness for modelling text-to-image translation and to avoid the same text embedding and image pair to give similar outputs with different combinations.

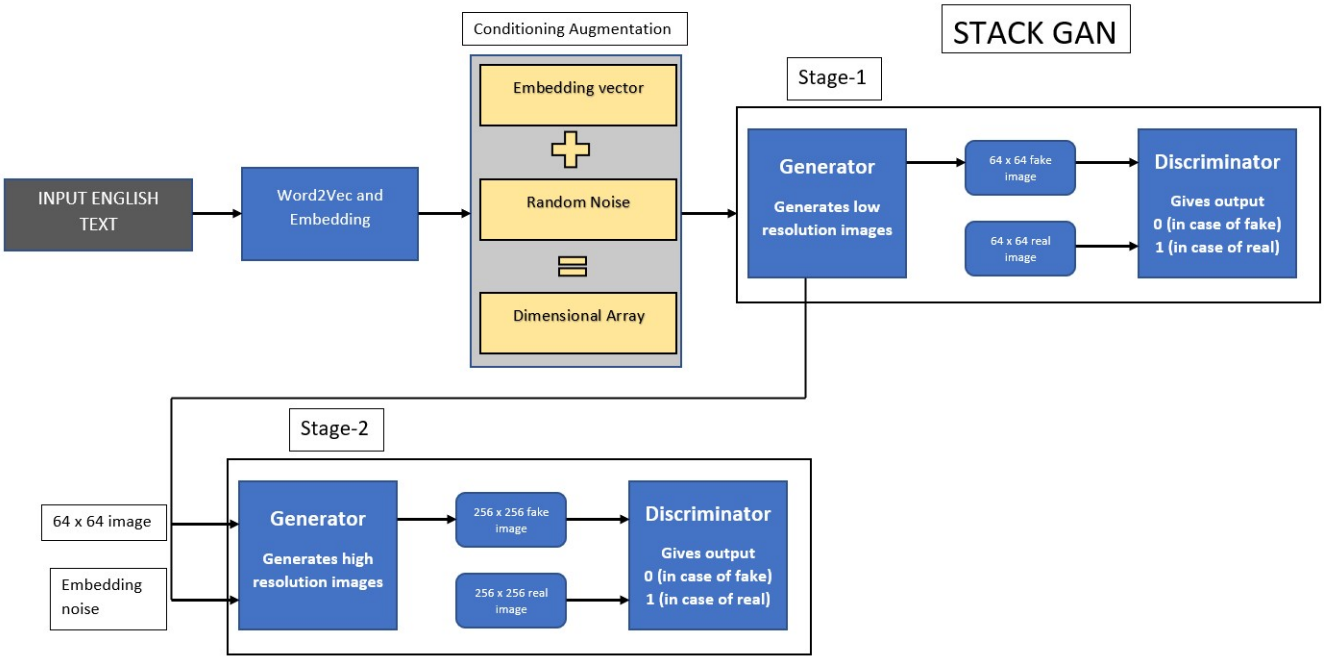

Figure 4.3: Stack-GAN Model

This d-dimensional array consisting of image-text pairs becomes the input to Stage-1 Generator.

Stage-1 : From the text description that has been provided, the generator has been provided, the generator sketches the shape as well as the colors of the object and creates the background design using noise vector which

is random and produces a 64x64 image of high resolution. This low resolution 64x64 fake image generated by the generator is sent as input to the discriminator along with the corresponding low-resolution 64x64 real image with its text description. The discriminator classifies these images and gives an output 0 in case of fake image and 1 in case of real image.

Stage-2 : This low-resolution 64x64 fake image generated by the Stage-1 generator in addition with a random noise becomes input to the Stage-2 generator. Here, the generator attempts to fix any errors in the low-resolution image from Stage-I and completes the object's details by reading the text description once more, generating a high-resolution photo-realistic 256x256 image. This high-resolution 256x256 fake image generated by the generator is sent as input to the discriminator along with the corresponding high-resolution 256x256 real image with its text description. The discriminator classifies these images and gives an output 0 in case of fake image and 1 in case of real image.

Chapter 5

Experiments and Results

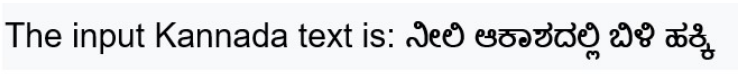

Figure 5.1: Kannada Input Text

The English translation of this text is "A white bird in the blue sky".This translation is done using the GLU translation model .The English text is then passed through the GAN model .In the Stage-1 of the Stack-GAN model the below image (64x64) was generated after the 1st epoch.The 1st epoch takes a time duration of ten minutes.

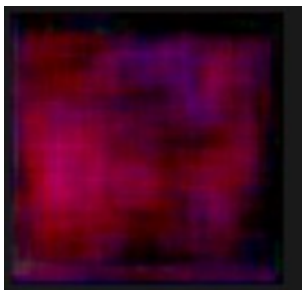

Figure 5.2: Stage-1 Generator First Epoch

After Stage 2 GAN a higher resolution image (256x256) was implemented the below image was generated. The Stage-2 Stack-GAN indicates that the processing text descriptions at Stage-2 helps improvise Stage-1 results.

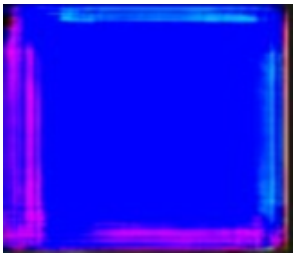

Figure 5.3: Stage-2 Generator First Epoch

After running for a period of 3 hours 30 minutes the below image was generated in the 17th epoch after the

implementation of Stage-2 Stack-GAN.

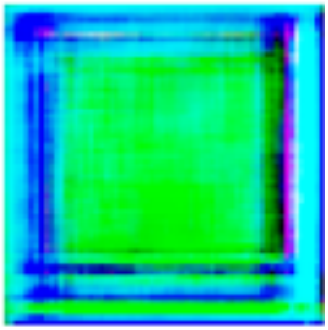

Figure 5.4: Stage-2 Generator 17th Epoch

After 50th epoch, Stage-2 Stack-GAN implementation the below image is generated.

Figure 5.5: Stage-2 Generator 50th Epoch

The final generated image is an image of "A white bird in a blue sky". The below image is generated after 80th epochs. The time duration to run for 80 epochs is 15 hours. The generation loss which is calculated based on discriminator's classification ability. The generative loss value shows the effectiveness of the model. The generator loss in the model is 1.56.

Figure 5.6: Stage-2 Generator 80th Epoch

In order to get a better resolution image, the model must be trained for more than 100 epochs and should be run in GPU in order to reduce processing time. Since the dataset is large it is necessary for the model to be trained for more epochs in order to get a high-resolution image.

ಹಕ್ಕಿಗೆ ರೆಕ್ಕೆಗಳು ಕೆಂಪು ಮತ್ತು ಕಿತ್ತಳೆ ಬಣ್ಣದ ಬಿಲ್ಲುಗಳನ್ನು ಹೊಂದಿದ್ದವು

Translation: the bird had wings that are red and has a orange bill

Figure 5.7: Result 1

ಈ ಹಕ್ಕಿಯು ಅಗಲವಾದ ರೆಕ್ಕೆಯನ್ನು ಹೊಂದಿದ್ದು, ಉದ್ದನೆಯ ಕೊಕ್ಕಿನೊಂದಿಗೆ ಬೂದುಬಣ್ಣದ ದೇಹದ ಗರಿಯನ್ನು ಹೊಂದಿರುತ್ತದೆ.

Translation: the bird has a wide wing span, with a grayish body feather with a long pointed beak.

Figure 5.8: Result 2

ಈ ಹಕ್ಕಿಯು ಕಂದು ಬಣ್ಣದ ದೇಹವನ್ನು ಹೊಂದಿದ್ದು ಕಪ್ಪು ಕೊಕ್ಕನ್ನು ಹೊಂದಿದೆ

Translation: this bird has brown body with black beak

Figure 5.9: Result 3

ಹಕ್ಕಿ ನೀಲಿ ರೆಕ್ಕೆಗಳನ್ನು ಮತ್ತು ಕಂದು ಕೊಕ್ಕಿನೊಂದಿಗೆ ಕಂದು ದೇಹವನ್ನು ಹೊಂದಿದೆ

Translation: the bird has blue wings and brown body with brown beak

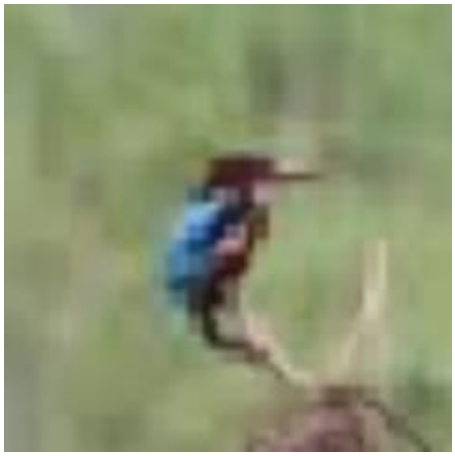

Figure 5.10: Result 4

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

In this project, the model proposed generates an image which is semantically consistent with the the input Kannada text .The proposed model translates the input Kannada text to English text, using a Gated Recurrent Unit (GRU) which is a machine translation model. The English text is then accessed by the Stack-GAN model which generates the image.

The Stage-1 implementation of the Stack-GAN creates the object with the primary colors and shape from the input text description. This is followed by the Stage-2 Stack-GAN implementation which rectifies flaws in the Stage-1 GAN implementation results and adds more features, resulting to better resolution images. Comprehensive qualitative values show the efficacy of the model. Compared to existing methods which only generate images from input text which are in English ,the Kannada text-to-image synthesis method generates images where the input text can be Kannada or English.

6.2 Future Scope

Generative Adversarial Networks is a hot topic in Deep Learning and Machine Learning. There has been immense growth in this field. GAN models with better outputs have been invented and such GAN's require high processing capacity which would generate better resolution images.
To keep up with the high processing capacity requirement, a good GPU can be used to reduce the time for training and testing. A GPU can run many epochs in a short duration of time. A photo-realistic image with more diverse photo-realistic details can be generated .

Bibliography

[1] Cristian Bodnar. Text to image synthesis using generative adversarial networks. 2018.

[2] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors, 2022.

[3] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[4] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H.S. Torr. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[5] Wentong Liao, Kai Hu, Michael Ying Yang, and Bodo Rosenhahn. Text to image generation with semantic-spatial aware gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18187–18196, June 2022.

[6] Marco Menardi, Alex Falcon, Saida S. Mohamed, Lorenzo Seidenari, Giuseppe Serra, Alberto Del Bimbo, and Carlo Tasso. Text-to-image synthesis based on machine generated captions. In Michelangelo Ceci, Stefano Ferilli, and Antonella Poggi, editors, *Digital Libraries: The Era of Big Data and Data Science*, pages 62–74, Cham, 2020. Springer International Publishing.

[7] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1060–1069, New York, New York, USA, 20–22 Jun 2016. PMLR.

[8] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] Hui Ye, Xiulong Yang, Martin Takác, Rajshekhar Sunderraman, and Shihao Ji. Improving text-to-image synthesis using contrastive learning. *CoRR*, abs/2107.02423, 2021.

[10] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.