

**MYANMAR TEXT TO IMAGE SYNTHESIS USING
GENERATIVE ADVERSARIAL NETWORKS**

Nang Kham Htwe

UNIVERSITY OF COMPUTER STUDIES, YANGON

March, 2024

Myanmar Text to Image Synthesis Using Generative Adversarial Networks

Nang Kham Htwe

University of Computer Studies, Yangon

A thesis submitted to the University of Computer Studies, Yangon in partial
fulfillment of the requirements for the degree of
Doctor of Philosophy

March, 2024

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

.....17.3.2024.....

Date

.....

Nang Kham Htwe

ACKNOWLEDGEMENTS

Firstly, I would like to thank the Minister, the Ministry of Science and Technology for full facilities support during the Ph.D. course at the University of Computer Studies, Yangon.

Secondly, I would like to express very special thanks to Dr. Mie Mie Khin, Rector, University of Computer Studies, Yangon, for allowing me to develop this research and giving me invaluable advice for my research.

I extend my sincere gratitude to Dr. Mie Mie Thet Thwin, former Rector of the University of Computer Studies, Yangon, for giving the opportunity to do this research.

I would like to show my sincere gratitude to Dr. Win Lelt Lelt Phyu, Professor and Course-coordinator of the Ph.D 13th Batch, University of Computer Studies, Yangon, for her guidance, mentorship and support during the research.

My greatest pleasure and the deepest appreciation to my supervisor, Dr. Win Pa Pa, Professor, Natural Language Processing Lab., University of Computer Studies, Yangon, for her kindly support, encouragement, excellent mentorship, invaluable guidance, constructive comments and practical advice throughout the research period.

I am extremely thankful to the external examiner, Dr. Ei Ei Hlaing, Rector, University of Computer Studies (Taungoo), for her patience in critical reading, valuable comments and suggestions in the preparation of the dissertation.

Moreover, I would also like to mention my special appreciation and thank to Dr. Khin Mar Soe, Professor, Head of Natural Language Processing Lab, University of Computer Studies, Yangon, for her excellent advice and suggestions, and providing me with the institutional resources.

I would like to grateful thank to all my teachers for their valuable comments to my thesis. To the reading committee teachers, especially Daw Aye Aye Khine, Associate Professor, Head of English Department, I would like to thank her for valuable supports and editing my thesis from the language point of view.

Finally, I would like to thank my family and friends for their unwavering support and encouragement, especially during the difficult times. Their love, understanding, and patience have been my source of strength and motivation.

ABSTRACT

The goal of text-to-image synthesis is to automatically create an image that corresponds to a given text description. It is the process of training a computer model to understand natural language and translate it into visual representations. One of the challenges of text-to-image synthesis is the semantic gap between natural language and visual representations. Natural language processing and computer vision techniques can be used to bridge this gap by mapping textual input to visual representations, which helps to generate more accurate and meaningful images. In text-to-image synthesis, computer vision is used to generate images that correspond to the textual input. On the other hand, natural language processing is used to process the textual input and extract meaningful information from it.

Text to image synthesis has gained popularity in recent years according to the advancements results in deep learning. It has become an active research area in artificial intelligence and has attracted searchers, practitioners, and the general public to focus on this research. However, Text to image synthesis for Myanmar is a challenging research problem because there are several factors that make generating images from textual descriptions difficult. One of challenge is the scarcity of large-scale annotated datasets of textual descriptions and corresponding images in Myanmar. Therefore, Myanmar caption corpus is manually built based on Oxford-102 flowers dataset to build Myanmar text-to-image synthesis (T2I) model.

In this dissertation, Myanmar T2I model is proposed using Generative Adversarial Networks (GANs). Firstly, Myanmar T2I based DCGAN is proposed to create images from Myanmar text descriptions. However, this model can generate low resolution images (64x 64). For this reason, AttnGAN and DF-GAN are used to investigate which model enable to generate high-resolution images (256 x 256) with semantic accuracy from Myanmar text descriptions. In this comparison, DFGAN gives better result for Myanmar T2I.

Moreover, DF-GAN+MSM (multimodal similarity model) is proposed in order to generate semantically consistency images with precise in shape for Myanmar language because there are artifacts that need to enhance on the generated images of DF-GAN. In DF-GAN+MSM, DFGAN is applied to generate images from Myanmar text descriptions. Multimodal similarity model is used to evaluate the matching score

between Myanmar text and the generated images during training of the model. This model contains two networks: text encoder and image encoder.

The evaluation on the performance of the models-based Myanmar T2I is done on two areas: quantitative analysis and qualitative analysis to assess the quality of the generated images. In quantitative analysis, DFGAN+MSM got the highest inception scores and the lowest FID scores of the generated images. In addition, DFGAN+MSM obtained the highest preferences scores based on qualitative evaluation by human perception. Moreover, the proposed model is also implemented with UCSD-CUB birds dataset annotated English to prove that this model gives the progressive results for the quality of synthesized images from different languages and different dataset.

Table of Contents

Acknowledgements	i
Abstract	ii
Table of Contents	iv
List of Figures	viii
List of Tables	x
List of Equations	xi
1. INTRODUCTION	
1.1 Problem Statements	3
1.2 Motivations of the Research	3
1.3 Objectives of the Research.....	4
1.4 Contributions of the Research	5
1.5 Organization of the Research	6
2. LITERATURE REVIEW	
2.1 Text to Image Synthesis.....	7
2.1.1 First Text to Image Synthesis.....	7
2.1.2 GAN with Stacked Architectures	8
2.1.3 Attention Architectures.....	9
2.1.4 Memory Architectures	10
2.1.5 Hierarchically Nested Architectures.....	11
2.1.6 Hierarchically Fused Architectures.....	11
2.1.7 Mirroring Architecture.....	11
2.1.8 Bridge Architecture.....	12
2.1.9 Cycle Consistency Architectures.....	12
2.1.10 Single Stage Architectures.....	14
2.2 Evaluation Metrics.....	14
2.2.1 Inception Score.....	14
2.2.2 Fréchet Inception Distance.....	16
2.3 Summary.....	17
3. THEORETICAL BACKGROUND	
3.1 Generative Adversarial Networks.....	18
3.2 Types of Generative Adversarial Networks.....	19
3.2.1 Conditional GAN.....	20

3.2.2 Deep Convolutional GANs (DCGANs).....	20
3.2.3 CycleGAN	20
3.2.4 StyleGAN	21
3.2.5 BigGAN	21
3.2.6 Super Resolution GANs (SGANs)	22
3.2.7 Generative Adversarial Text to Image Synthesis.....	22
3.2.8 Adversarial Autoencoder (AAE).....	23
3.3 Application Areas of Generative Adversarial Networks.....	23
3.4 Representations.....	24
3.4.1 Unimodal Embedding.....	25
3.4.1.1 Language Representations.....	25
3.4.1.2 Visual Representations.....	26
3.4.2 Multimodal Embedding.....	26
3.4.2.1 Joint Representations.....	26
3.4.2.2 Coordinated Representations.....	27
3.5 Other Visual Language Generative Models.....	28
3.5.1 DALL-E.....	28
3.5.2 GLIDE.....	29
3.6 Text Encoder and Image Encoder.....	30
3.6.1 Convolutional Neural Networks.....	30
3.6.2 Bi-directional Long Short-Term Memory.....	33
3.7 Text Encoding Techniques.....	34
3.7.1 Indexed-Based Encoding.....	35
3.7.2 Word2Vec.....	35
3.8 Summary.....	36

4. DATASET PREPROCESSING

4.1 Dataset for Myanmar T2I.....	37
4.1.1 Building Annotated Myanmar Caption Corpus.....	37
4.1.2 Caption Preprocessing	37
4.1.3 Text Encoding using Indexed-Based Encoding	39
4.2 Dataset for English T2I.....	41
4.3 Image Preprocessing.....	42
4.4 Summary.....	43

5. DEEP CONVOLUTIONAL NEURAL NETWORKS FOR MYANMAR T2I	43
5.1 Deep Convolutional GANs based Myanmar T2I	43
5.2 Detailed Implementations.....	45
5.2.1 Dataset Creation.....	45
5.2.2 Preprocessing the Dataset.....	45
5.2.3 Training Stages.....	46
5.2.4 Experimental Results.....	46
5.2.5 Evaluation.....	48
5.3 Summary.....	48
6. GENERATIVE ADVERSARIAL NETWORKS FOR MYANMAR T2I	49
6.1 Architecture of Attentional Generative Adversarial Networks	49
6.2 Deep Attentional Multimodal Similarity Model (DAMSM)	50
6.3 Architecture of Deep Fusion GANs	51
6.4 Training of Myanmar Text to Image Synthesis	52
6.5 Experiment Results and Comparison	54
6.6 Summary	57
7. MULTIMODAL GENERATIVE MODEL BASED T2I	58
7.1 Architecture of Multimodal Similarity Model	58
7.2 Architecture of DF-GAN+MSM	60
7.3 Implementation Details	61
7.3.1 Dataset	61
7.3.2 Preprocessing Dataset	62
7.3.3 Pretraining Multimodal Similarity Model	63
7.3.4 Training DFGAN+MSM	63
7.4 Experimental Results and Analysis on Myanmar T2I	64
7.4.1 Quantitative Analysis.....	64
7.4.2 Qualitative Analysis	66
7.5 Experiment Results and Analysis on English T2I	69
7.5.1 Quantitative Analysis.....	69
7.5.2 Qualitative Analysis	69
7.6 Summary	70

8. CONCLUSION AND FUTURE WORK	72
8.1 Advantages and Limitations of the Proposed System.....	74
8.2 Future Works.....	75
Author’s Publications	75
Bibliography	76
Appendices	83
List of Acronyms	95

LIST OF FIGURES

1.1	Example of Five Annotated Captions and Corresponding Image	2
2.1	Steps of Evaluation of Inception Score	15
2.2	Steps of Evaluation of FID Score	17
3.1	Framework of Generative Adversarial Networks	18
3.2	Joint Representations	27
3.3	Coordinated Representation.....	28
3.4	DALL-E	29
3.5	Simple Framework of Convolutional Neural Network	30
3.6	Convolution.....	31
3.7	Max-Pooling.....	32
3.8	Average-Pooling.....	32
3.9	Fully-Connected Layers	33
3.10	Framework of Bi-LSTM.....	34
4.1	Sample of Segmented Word	38
4.2	Sample of Flower Images with Annotated Captions.....	39
4.3	Sample of Bird Images with Annotated Captions	40
4.4	Sample of Original Images and Preprocessed Images	41
5.1	Framework of DCGAN based Myanmar T2I	44
5.2	Images Generated from Myanmar Text Descriptions based on DCGAN	47
6.1	Frameworks of AttnGAN	50
6.2	Frameworks of DF-GAN	52
6.3	The Image Generation Process of (a) AttnGAN (b) DF-GAN	53
6.4	The Images Generated from Myanmar descriptions while changing some Words	55
6.5	The Images Generated from Myanmar Text Descriptions.....	56
7.1	Architecture of MSM	60
7.2	Framework of DF-GAN+MSM	61
7.3	Figure 7.3. Flowchart of Myanmar T2I	62

7.4	Comparison of FID Scores on AttnGAN, DF-GAN and Our Proposed Model	65
7.5	Comparison of IS Scores on AttnGAN, DF-GAN and Our Proposed Model	66
7.6	Comparison of Preference Scores.....	67
7.7	The generated images from Myanmar Textual Inputs	68
7.8	The Bird Images Generated from English Text Description	70
A.1	Home Page of Myanmar T2I	83
A.2	The Generated Images from Query Text 1	84
A.3	The Generated Images from Query Text 2	86
A.4	The Generated Images from Query Text 3	86
A.5	The Generated Images from Query Text 4	87
A.6	The Generated Images from Query Text 5.....	87
A.7	The Generated Images from Query Text 6	87
A.8	The Generated Images from Query Text 7	88
A.9	The Generated Images from Query Text 8.....	89
A.10	The Generated Images from Query Text 9.....	90
A.11	The Generated Images from Query Text 10.....	90
A.12	The Generated Images from Query Text 11.....	91
A.13	The Generated Images from Query Text 12.....	92
A.14	The Generated Images from Query Text 13.....	92
A.15	The Generated Images from Query Text 14.....	93
A.16	The Generated Images from Query Text 15.....	93
A.17	The Generated Images from Query Text 16.....	94
A.18	The Generated Images from Query Text 17.....	94

LIST OF TABLES

Table 4.1	Sample of Converting Words to Index Numbers	38
Table 5.1	FID Scores and Inception Scores of the Generated Image from DCGAN	46
Table 6.1	Training Parameters of AttnGAN and DF-GAN.....	54
Table 6.2	Inception score and FID score of two models evaluate based on test data.....	54
Table 7.1	Hyperparameters setting of MSM.....	63
Table 7.2	Hyperparameters setting of DF-GAN+MSM.....	64
Table 7.3	Comparison of FID score and Inception Score.....	66
Table 7.4	Comparison of FID and Inception Scores of the Proposed model with others on CUB dataset.....	69
Table A.1	Query Input 1.....	84
Table A.2	Query Input 2.....	85
Table A.3	Query Input 3.....	87
Table A.4	Query Input 4.....	88
Table A.5	Query Input 5.....	89
Table A.6	Query Input Text on English T2I.....	91

LIST OF EQUATIONS

Equation 2.1.....	15
Equation 2.2.....	16
Equation 3.1	19
Equation 3.2	33
Equation 3.3	33
Equation 3.4	33
Equation 6.1	50
Equation 6.2	51
Equation 6.3	52
Equation 7.1	59
Equation 7.2	59
Equation 7.3	59
Equation 7.4	59
Equation 7.5	61

CHAPTER 1

INTRODUCTION

Text-to-image synthesis (T2I) is one of the exciting research fields of artificial intelligence (AI) where the purpose is to create algorithms that can output realistic and semantically meaningful images from text descriptions. This is the combination of natural language processing and computer vision, and has the potential to be applied in many industries including art, design, and virtual reality.

The process of T2I synthesis involves training deep learning models on large datasets of paired text and image data. In this work, T2I models learn and recognize the relationships between text and images, and then generate new images based on text input. It has been applied in many applications, generating personalized content, creating realistic images for video games, and designing new products. However, there are still challenges to overcome, such as generating best-quality images with fine-grained details and ensuring that the generated images are visually realistic and semantically refer to text data.

However, the generation of images from Myanmar text is a relatively new and challenging field of research. Myanmar is a complex and morphologically rich language, with many unique features and characteristics that leads to many challenges to develop deep learning models for T2I tasks. Despite these challenges, there have been some progressive technologies to apply in Myanmar text-to-image (T2I) synthesis. Generative Adversarial Networks (GANs) have widely used by many researcher to output images conditioned on English descriptions and has remarkable progress in image and video generation. Therefore, this research is implemented using Generative Adversarial Networks to synthesize images from Myanmar text descriptions. GANs is a kind of deep learning models that involves two networks: a generator network and a discriminator network.

In T2I based GANs, the generator takes a text input and generates corresponding image. The discriminator is used to discriminate between the real images in the dataset and the generated images. The generator task is to produce more realistic images and makes the discriminator indistinguishable real image from fake. Typically, this work involves training a neural network on a large dataset of paired text descriptions and visual inputs, and then using the trained model to generate images from

new text descriptions. Therefore, an annotated images dataset extended from Oxford-102 flowers is manually constructed to train Myanmar T2I. This dataset contains 5 Myanmar text descriptions for each image in this dataset.

Moreover, multimodal similarity model (MSM) is applied to the generator side to evaluate the visual-semantic consistency. This model contains two sub-encoders, text encoder and image encoder. Bidirectional Long Short-Term Memory is used to build text encoder. Convolutional Neural Networks is used as image encoder to extract image features. MSM is pretrained using real images with text pairs in the dataset. The pretrained text encoder is used to extract the features of Myanmar given as input to generator. Furthermore, text-to-image synthesis also implemented on Caltech-Birds datasets annotated on English to investigate the progressive results of the image generation process due to applying multimodal on generator side.



- ဤအနီရောင်ပန်းပွင့်တွင်ထပ်နေသောပွင့်ချပ်နှင့်အဝါရောင်ဝတ်ဆံရှိတယ်
- ဤအနီရောင်ပန်းပွင့်ပေါ်တွင်အဝါရောင်ဝတ်ဆံနှင့်အနီရောင်ပွင့်ချပ်ရှိတယ်
- ကြီးမားသောအနီရောင်ပန်းပွင့်တွင်ထပ်နေသောပွင့်ချပ်နှင့်အဝါရောင်ဝတ်ဆံရှိတယ်
- ပန်းပွင့်တွင်အနီရောင်ပွင့်ချပ်အလွှာများနှင့်အဝါရောင်ဝတ်ဆံရှိတယ်
- ပန်းပွင့်သည်အနီရောင်ရှိပြီးထပ်နေသောပွင့်ချပ်များရှိတယ်

Figure 1.1 Example of Five Annotated Captions and Corresponding Image

1.1 Problem Statements

Myanmar is a complex and morphologically rich language, with many unique features and characteristics that lead to challenging in developing deep learning models for Myanmar T2I.

One of the main challenges in Myanmar T2I task is lack of annotated datasets. There is an annotated images dataset (extended from Flickr dataset) that include paired text and image samples. However, T2I synthesis research generally does not use Flickr dataset because it is not specifically designed for this task and contains images with varying degrees of quality and objects, including low-resolution images, images with poor lighting or composition, and images with watermarks or other artifacts. This variability in image quality can make it challenging to synthesize high-resolution images that accurately describe the textual descriptions.

Another challenge is the complexity of Myanmar language, which has many unique features such as stacked diacritical marks and context-sensitive spelling rules. These features make challenging to develop deep learning models that can accurately represent the meaning of text and generate high-quality images.

Myanmar T2I synthesis is a challenging research area that requires potential efforts to overcome the above challenges. Addressing these challenges need to develop specialized datasets, and the use of sophisticated T2I synthesis models that can generate best-quality and visually coherent images from Myanmar textual descriptions.

1.2 Motivations of the Research

The improvements in deep learning have a significant impact on computer vision applications and NLP techniques in recent years. Among these applications, the generation of images from text descriptions becomes one of active research areas because it provides many benefits for various fields, including art creation, image editing, virtual reality, etc. These applications demonstrate the importance of T2I synthesis areas, and deep learning models will continue to evolve with more exciting advancements in the future.

Images are often more understandable than text in conveying complex information for human beings. Therefore, converting images from textual descriptions can improve the accessibility of information and leading to better understanding of

complex information. For this reason, the creation of innovative applications and services will lead to provide benefits to individuals and society as a whole.

To accurately capture the nuances of language and understanding the context of the sentence is an essential step for generating appropriate images. The ability to generate the image with better lighting, accuracy in color, and precise in shape is also the important factor.

Moreover, it is required to create the images which are semantically described with textual description. This requires to match the relationships between objects and text features in depicting the generated images.

For these above reasons, state-of-the-art models are needed to be able to understand the complexities of Myanmar language and to generate the realistic image semantic consistency based on Myanmar text descriptions.

1.3 Objectives of the Research

The main intention of this research is to visualize images that are semantically reflects with the text descriptions written in Myanmar language. The multiple stages of image generation with attention mechanism and single stages of images generation with fusing of text and images at every blocks of generator are compared and analyzed to depict which method can generate best-quality and diverse images that reflect the semantic of the sentence. The following are the other objectives:

This research is a relatively new and lack of annotated image dataset that contains Myanmar text descriptions with corresponding images. The building of an annotated image dataset is an essential process for Myanmar text-to-image synthesis. Therefore, construction of Myanmar captions corpus is one of the objectives of this research.

The preprocessing steps are essential for text-to-image synthesis because they help to ensure that the data is consistent, error-free, and in a form that can be easily processed by the deep learning models. Proper preprocessing can enhance the quality of generated images and increase accuracy of the model. Therefore, segmentation of Myanmar sentence and encoding this segmented sentence to numerical data is also one of the objectives.

The generated images should be semantically consistency with descriptions. The similarity alignment between text and the artificially created image is an important

factor in evaluating the quality of image. Therefore, the other objective is to propose multimodal similarity model that enables to evaluate visual-semantic consistency during training stages.

1.4 Contributions of the Research

To develop this research, the first important factor is the availability of large-scale annotated image dataset. But there is the lack of an annotated image dataset that are well suited to implement this research. Therefore, building a dataset with text-image paired is the initial contribution in this work.

Data preprocessing is one of the essential step prior training of deep learning models because it can significantly increase the performance and accuracy of deep learning models. The tokenization of Myanmar sentence and encoding this tokenized sentence into numerical format is also taken as the contribution of this research.

The other contribution is investigation of text-to-image synthesis models (i.e. attention-mechanism with multiple refinement stages namely AttnGAN, single stage of image generation with fusing text and image at every block of the generator called DF-GAN) to highlight which method gives better impact for this research because of the complexity of Myanmar language.

The next contribution is building T2I synthesis using GANs that can visualize high-resolution images (256x256) from Myanmar input text. This model contains one stage of generator with multimodal similarity model (MSM) and discriminator.

To compute similarity score between text and the generated images during training stage, MSM is composed to generator site. Therefore, building MSM before implementing of T2I is also one of the contributions.

Moreover, the proposed multimodal generative model is also applied on different annotated with different languages to analyze the impact of improvements in the image generation stages.

1.5 Organization of the Research

This dissertation is organized with eight chapters. This chapter contains an introduction, the problem statements, motivations, objectives and contributions of the research work.

Chapter 2 contains the literature reviews on other text-to-image synthesis tasks related to this research. The theoretical methodologies used to conduct this search are described in Chapter 3. In Chapter 4, creation of Myanmar caption corpus, preprocessing of the dataset such as word segmentation, text encoding and image preprocessing have been discussed. The implementation of the first text-to-images-synthesis model with DCGAN is reported in Chapter 5. Chapter 6 contains implementation of AttnGAN and DFGAN and the comparative evaluation of these two models. The pretraining of MSM that contains text encoder based Bi-LSTM and image encoder-based CNN, details implementation of the proposed Myanmar T2I models are in Chapter 7. This chapter also describes the evaluation of the generated images on both quantitative and qualitative evaluation (i.e. evaluation of the generated image by human perception). The last chapter of this research concludes the research work and highlights the advantages and limitation of the research, the future works to progress this research.

CHAPTER 2

LITERTATURE REVIEW AND RELATED WORK

This chapter contains about the literature review on T2I synthesis techniques, the works related this research and previous research works on Myanmar T2I synthesis. The two popular evaluation metrics used to evaluate the performance of T2I models are also described in this chapter.

2.1 Text to Image Synthesis

It is a technique for translation of text descriptions into artificially created images with semantically consistent. Most recently proposed T2I synthesis methods based on GANs [20] has become popular. There are two networks in GANs: Generators and Discriminators. In this process, textual descriptions are first converted into semantic vectors representations by using language models. Second, the noise vectors are randomly sampled by using standard normal distributions. These noises and semantic vectors are fed as input to generator. Finally, these noises are generated into meaningful output or fake images semantically consistent with text. The discriminator task is to classify the synthesized images from generator is real or not.

There are many paradigms of T2I approaches. Among them, the first T2I approaches, followed by stacked architectures, stacked architectures with attention mechanisms and dynamic memory networks, and T2I with fusion module blocks are described in the following sections.

2.1.1 First Text to Image Synthesis

The first approach, GAN-INT-CLS [44], generates fake images conditioned on the whole sentence vectors passed from a pretrained text encoder. This approach has trained using a deep convolutional generative adversarial network (DC-GAN) conditioned on text features embedded by contributing with character-level CNN-RNN. In this GAN, three different inputs are passed to the discriminator: a real image with wrong text, a real image with matching text and a fake image with corresponding text. This model trained on both the generator and the discriminator by focusing on real images but also aligning with the input text.

2.1.2 GAN with Stacked Architectures

GAN-INT-CLS [49] enables to create low-dimension images with 64×64 . In order to enhance high-dimension images latter GANs proposed multiple stages of generators and discriminators.

In StackGAN [18], artificial images are generated from text embedding that contains many conditioning variables. This model contains two-cascaded GAN, each stage integrates a set of generator and discriminator. Stage I GAN takes a text description and transforms this input to a text features which contains several conditioning variables. Then the synthesized images with basic structure and colors are generated based on these text embeddings. Stage II GAN takes low-resolution 64×64 images from stage I and outputs high-resolution images by spatial replication of the text embedding used in stage I. This stage corrects defects in stage I and generates with dimension of 256×256 . Prior works used [49] fixed-length embedding, which contains static conditioned variables. They are converted from text descriptions and feed as input to the generator. This approach process like this and firstly creates Gaussian distribution from natural language descriptions. After this, it randomly adds selected variables from this distribution to conditioning variables during training. By impacting a small variation to the original text embedding, the trained models can generate more diverse images for the same input text.

StackGAN ++ [19] contains multiple stages of generators and discriminators and these are organized as ‘tree-like’ structure. At first stage, the generator combines conditioning variables and noise vectors and outputs low-resolution images. In the next following stages, generator transforms the resolution images, 64×64 into high-resolutions images. During training, the generators are joined to approximate multi-size, joint conditional and unconditional distributions. The discriminators are trained to differentiate how well the image match with text description and to distinguish the generated images look like realistic or not. In increasing image size at different stages of generating images, the synthesized images at various sizes should share similar shape and colors. Therefore, a color-consistency regularization was proposed to improve the artificially created images at different dimensions which are more accuracy in color and better in quality.

2.1.3 Attention Architectures

Most recently proposed text-to-image tasks generate the image based on the whole sentence features. These approaches have been acquired the impressive results but such kind of conditioning on the whole sentence vectors lacks of fine-grained information at word features. To address this issue, the author [53] proposed AttnGAN (Attentional GANs) that can create artificial images based on both word-level and sentence-level. This model consists of two contributions, attentional generative networks and Deep Attentional Multimodal Similarity Model (DAMSM). AttnGAN contains multiple stages of generators and discriminators. The generative network generates low quality images in the initial stage by combining the noise vector and the sentence vector. In the next two stages, it utilized the image vector in each sub-region of the generated images to query which word vectors is relevant to those sub-regions by using attentional layer and creates word-context feature. The features of its sub region and its correspondence word-context features are then combined to create higher quality images. The resolution of the generated images is 64x64 (at first generator), 128 x 128 (at the second stage) and 256 x 256 respectively (at the third stage). DAMSM computes the consistency level between the synthesized images and input text descriptions during training stages.

Frequently previous T2I models that employ on text condition would generate objects with unrealistic and deformed layouts. To solve this problems, they proposed SegAttnGAN, which applies the segmentation information to add global spatial attention. In this system, the segmentation attention module is utilized to advance the quality of images by referencing spatial features of the semantic-maps. They use CUB [47] and Oxford [28] datasets to implement their proposed method. The two quantitative measurements are used to evaluate generated images: (1) Inception Scores and (2) R-precision scores. SegAttnGAN generates the images with better quality and shape than the baseline model AttnGAN. This model synthesizes well and high-quality images, but this method needs to input segmentation data during the training phase.

The synthetic images generated from StackGAN and AttnGAN would be significantly varied from images outputted from the original descriptions when some words of a sentence are changed. This is inapplicable in real-word areas, when someone only wants to modify the generated image according their needs. Therefore, the author proposed ControlGAN [8] that can effectively disentangle various attributes and

exactly evaluate portions of the generated image without effecting diversity. The implementation results show that the proposed method outperforms existing state of the art [19, 49, 53]. It can be seen that when the text is changed, two compared methods are more likely to create new object, or modified some aspect values that are not correspond to the modified text. This model is able to exactly manipulate each region of synthesized images relevant to the modified text, while maintaining the visual aspects related to original text.

Previous methods usually generate image conditioned on sentence embedding at the initial stage and then refine this image using word-level features. Despite the remarkable results has achieved, but lack of aspect-level information in the sentence in this methods. This information can guide the important information in enhancing the quality of images. To address this problem, the author proposed DAE-GAN [45] that generate images conditioned on multiple levels of information, including word-level, sentence-level and aspect-level. Moreover, they also introduced an Attended Global Refinement (AGR) module and an Aspect-aware Local Refinement (ALR) module. The quality of images are enhanced by AGR module based on word-level features, while ALR gradually refines the images by adding details information with the used of the aspect-level embedding. Finally, they also contributed the matching loss function in order to ensure consistency between text and image at different levels.

2.1.4 Memory Architectures

There are two problems still remain despite multiple-stages image generation [19, 49, 53] has achieved remarkable progress. First, the quality of the synthesized images from refinement states is bad if the initially generated images quality is poor. Second, these models utilize the same word representation in the refinement process that leads to ineffective in translating semantic meaning of words and text-image consistency. Therefore, the authors [34] introduced DMGAN that contains memory mechanisms at every generator to generate high-dimension images even though the initially generated images are not well generated. They also introduced memory writing gate that enables to select the most important words by referencing the initial images and generate images upon these words in the refinement stages. Therefore, this model improves the IS from 4.36 to 4.75 and decreased FID from 3.98 to 16.09 on the CUB dataset compared with StackGAN ++.

2.1.5 Hierarchically Nested Architectures

To tackle the need of multiple generators, HDGAN [64] built with hierarchically-nested discriminators along the depth of a single stream generators that contains multi-scale intermediate layers (64x64, 128 x 128, 256 x 256 and 512 x 512 respectively). Therefore, it does not require multiple-stages of training and multiple-conditioning with text like StackGAN to generate high resolution images with dimension of 512 x 512. To guarantee diversity and visual-semantic consistency, the discriminators was paired with at each stage of the generator to simultaneously distinguish the real image or fake image not only measure the similarity between visual and semantic.

2.1.6 Hierarchically Fused Architectures

Most recently proposed [19, 34, 49, 53] methods required multiple stages of discriminators though they could generate high-resolution images. The use of many stages of discriminators leads to higher computational cost and unstable training process. Therefore, the author [55] proposed a HFGAN, which can synthesize high resolution images using three stages with only one discriminator. Compared to previous methods, it fully uses convolutional layers and avoids degradation quality of images.

2.1.7 Mirroring Architecture

The creation of images should perform as a mirror that accurately reflects semantics meaning of text. Based on these motivations, the author proposed a novel T2I-I2T (text-to-image-mage-to-text) framework called MirrorGAN [51] to improve T2I generation, which is extended T2I synthesis. It is composed with a mirror architecture by integrating both T2I and I2T. The proposed model includes a semantic text regeneration and alignment module (STREAM) to recreate he text description from the generated images that are semantically consistent with the given text description. This model improved the Inception Score 4.56 on CUB and from 26.47 on the more difficult COCO dataset.

2.1.8 Bridge Architecture

Generation of images based on text description with semantic consistent is still a challenging research areas. To address this issue, it is required to implement a transitional model with interpretable representation to relate image and text. The author proposed a Bridge-GAN (“Bridge-like Generative Adversarial Networks”) [33] to align the latent vectors based on input text in order to inspect representation learning. This methods exceeds the IS scores of 0.59 and 0.70 than HDGAN and StackGAN++. In comparative analysis with state-of-the-arts methods on CUB dataset, these model got the largest VS similarity than these methods, Compared with HDGAN, this approach can also improve the VS similarity scores of by 0.052. These comparative results on VS similarity results can prove the performance of this model on increasing consistency level on context.

2.1.9 Cycle Consistency Architectures

Previous T2I approaches prefer to translate the image with similar look for different text and overview the characteristics of the sentences. The translation of images with high-quality from text description is an important work and it can be applied in many real application areas, e.g. chatting online, graphic design, etc. Though advancement over generating images with best-quality, these models have mode collapse issue. To solve this issue, the author proposed SuperGAN [62] with cycle-consistency to generate the images with different shape from text descriptions. This approach is trained with text-to-image-to-text framework, and introduces a cycle-consistency loss to the text-to-image model to obtain more diverse modes. In order to generate image with more details correlation to text descriptions and avoid distortion, the initialize image from the previous stage is refined in the next stage. This proposed model is conducted using Oxford flowers dataset is to evaluate the model with cycle structure. The non-cycle structure model cannot obtain different modes from data distribution.

Previous methods emphasize the generation of images based on text description, and train the model for text-aligned image generation. Therefore, the author [17] pretrained GAN without inputting text to generate diversity and high-quality images before training of T2I models. And then, they build GAN inversion model to create

latent vector from the images. Thereby they introduced the cycle-consistency methods to discover more robust and coherent latent codes. Finally, a similarity model is constructed to evaluate the similarity between sentence representations and the inverted latent codes in order to optimize this codes. This optimized code is then input to the pretrained GAN generator and perform t text-to-image synthesis task. Most of the previous methods, such as AttnGAN [53], MirrorGAN [51] use the cGAN (conditional GAN) with the attention architecture to enhance text representation and further accelerate the process of image generation. In contrast, many existing T2I tasks are constrained on wrong interpretation of semantic text in some cases. However, this proposed method can not only create images which are relevant with the given text, but also synthesize high diversity and realistic images.

2.1.10 Single Stage Architectures

In T2I, the fake images need to be realistic, and semantically relevant with text. Although progressive results have been stated by previous researchers [8], there still remain three problems. To address the above issues, the author proposed DF-GAN (“Deep Fusion Generative Adversarial Network”) [32]. They proposed a novel one-stage backbone T2I that can generate high-quality images directly without using multiple stages of image refinement. With the use of Matching-Aware Gradient Penalty (MA-GP), it can significantly enhance the relevant similarity between text and images without applying alternative networks. They implement the proposed model on two different datasets, i.e., CUB bird and COCO. In the comparative analysis with state-of-the-art methods, the proposed model has achieved remarkable results in T2I without stacked structures. Compared with AttnGAN, DF-GAN increased IS score from 4.36 to 5.10 and declined FID scores from 23.98 to 14.81 on the CUB dataset. Moreover, DF-GAN can also generate the images with more fine-grained details information compared to these models.

The most previous methods are implemented with multi-stage of image refinement which generate low-resolution image from noise concatenated with sentence features and refines the information in detailed with attentional word-level in the next stages. The use of these multiple generators paired with discriminator tend to higher computational cost and unstable training stage. In addition, the quality of final images depend on the initially generated images by earlier generator. If the previous

outputted image is bad, the generators at the following stages cannot enhance its quality. To address these issues, the author proposed Semantic-Spatial Aware Generative Adversarial Network (SSA-GAN) [55] and one-stage framework for T2I task. In the comparison with the popular multi-stage network, one-stage network requires less computational time and can be trained more efficiently and stably. Compared to the methods which use word-level features, this approach is simple and less computation. They also introduced SSA framework to concatenate the sentence and image features effectively by forecasting semantic mask to guide the adaptive affine transformation in pixel level. SSA-GAN reports the significant improvements in IS (from 4.86 to 5.17) on CUB dataset compared to DF-GAN.

2.2 Evaluation Metrics

There are several evaluation metrics that can be used to determine the quality of artificially generated images. Among them, the most frequently utilized evaluation techniques for GANs are Inception score and Fréchet Inception Distance (FID).

2.2.1 Inception Score

This metric [52] is used to predict the quality of generated images from generative adversarial networks (GANs). Firstly, this score is computed by applying Inception-v3 network to classify the generated images and obtain their class probabilities. Then, the score is calculated as the expected value of the KL divergence between the class distribution of the generated images and the real images. The evaluation steps for this score is shown in Figure 2.1.

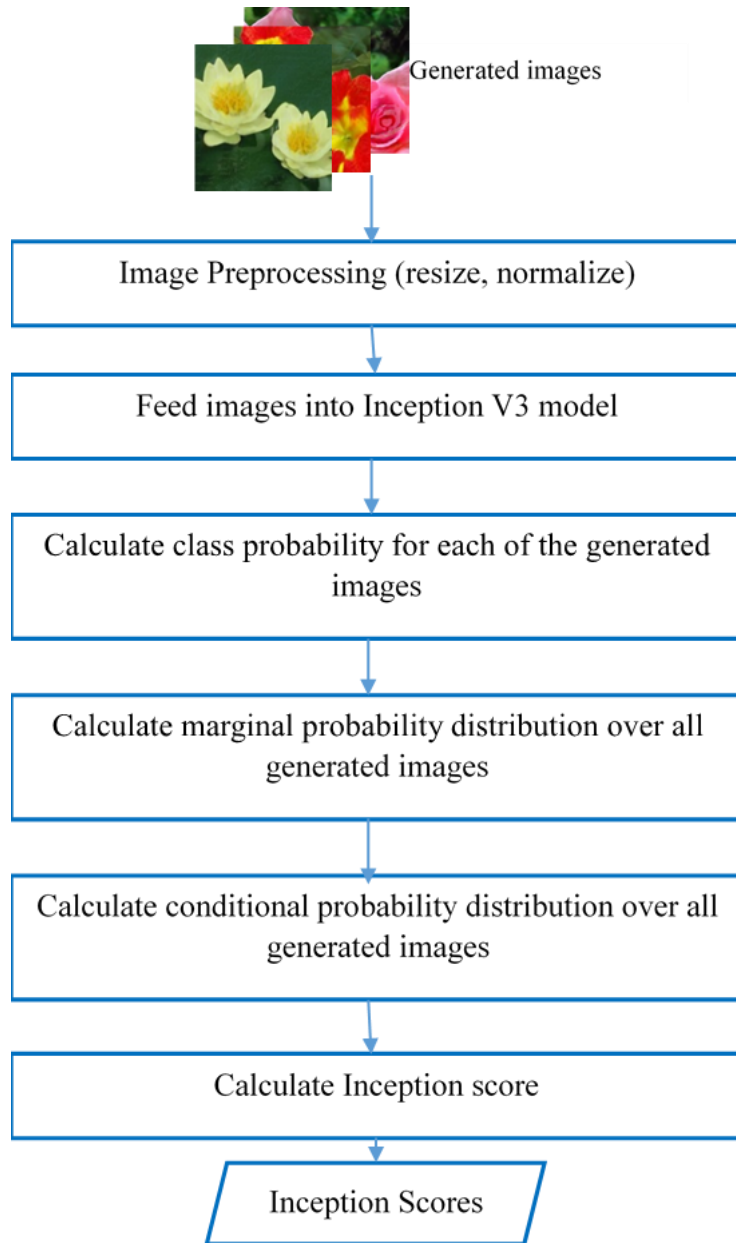


Figure 2.1 Steps of Evaluation of Inception Score

The IS score of the generated image is calculated by using the following equation:

$$I = \exp(\mathbb{E}_{x \sim p_g} D_{KL}(p(y|x) || p(y))) \quad (2.1)$$

Where,

x = the synthesized images

y = the labels predicted from model

$p(y|x)$ = the class probability distribution

$p(y)$ = the marginal class distribution

In simpler terms, IS measures both diversity and quality of the synthesized images, by predicting at how much the generated images capture the probability distribution of the real data. A higher score indicates higher quality and better diversity of the images.

2.2.2 Fréchet Inception Distance (FID)

This metric is [52] used to evaluate the quality of generated images compared to a set of real images. It is calculated by first feeding both the set of real images generated images Inception v3 model to obtain their features. Then, the mean and covariance of these images are built by using their extracted features. FID is then computed as the distance between these two distributions by using the following equations:

$$FID = \|\mu_r - \mu_g\|^2 + T_r(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g))^{1/2} \quad (2.2)$$

Where,

r = real image

g= generated image

T_r = trace of a matrix

Fréchet distance is the similarity measurement between two data distributions. A smaller FID indicates that the generated images are more similar to the real images in terms of their feature distributions.

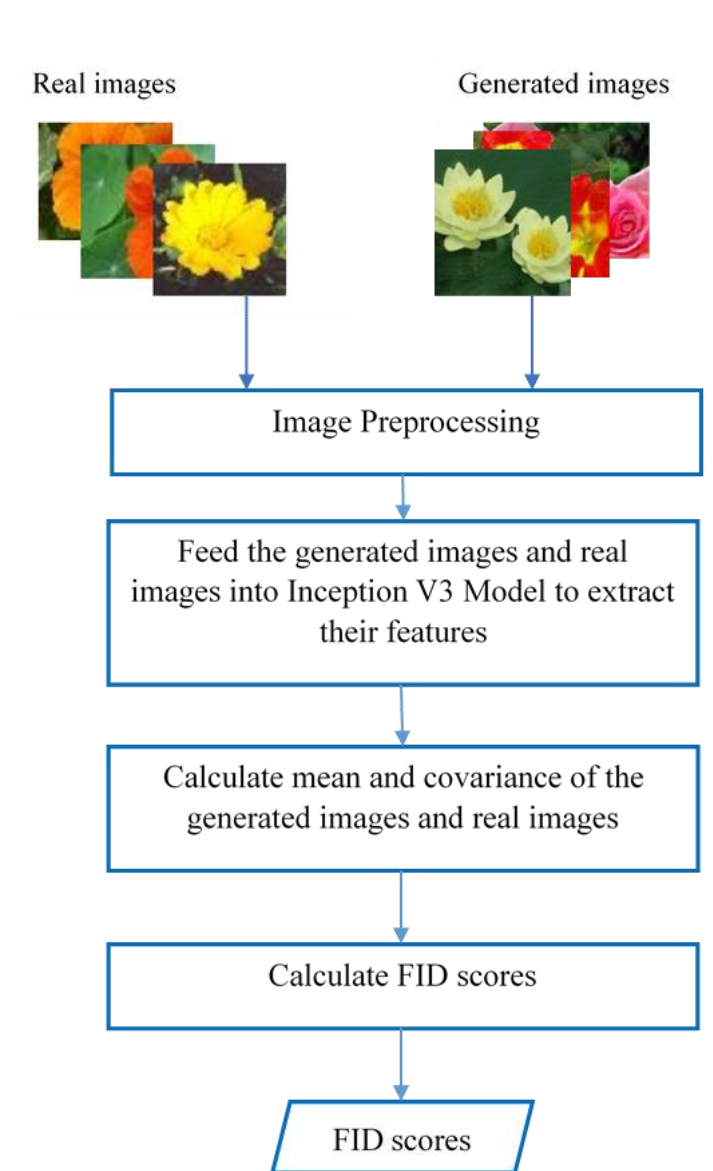


Figure 2.2 Steps of Evaluation of FID Score

2.3 Summary

The review of various research and methodologies related to Myanmar T2I is described in this chapter. According to these statements, there are many research works and improvement for other languages except for Myanmar language. Therefore, Myanmar T2I synthesis is implemented by applying Generative Adversarial Networks like other language. In addition, the two popular evaluation metrics: (1) Inception score and (2) Fréchet Inception Distance (FID) used to evaluate the quality of the generated images are also described.

CHAPTER 3

THEORETICAL BACKGROUND

The theoretical background applied in Myanmar T2I is described in this chapter. Deep learning is a subset of machine learning that applies artificial neural networks consisting of many layers to understand from large datasets. It involves training neural networks with multiple layers of connected nodes, artificial neurons, to recognize patterns in the data and make decisions based on that pattern. Among many algorithms of deep learning, Generative Adversarial Networks utilized to create images from Myanmar descriptions is mainly focused in this chapter. Besides this, other variations and applications of GAN, data representation techniques are also described. Finally, the two neural networks (Convolutional Neural Network and Bidirectional Long Short-Term Memory) that used to model image encoder and text encoder, and other methods used to encode Myanmar text descriptions are also discussed in this chapter.

3.1 Generative Adversarial Networks

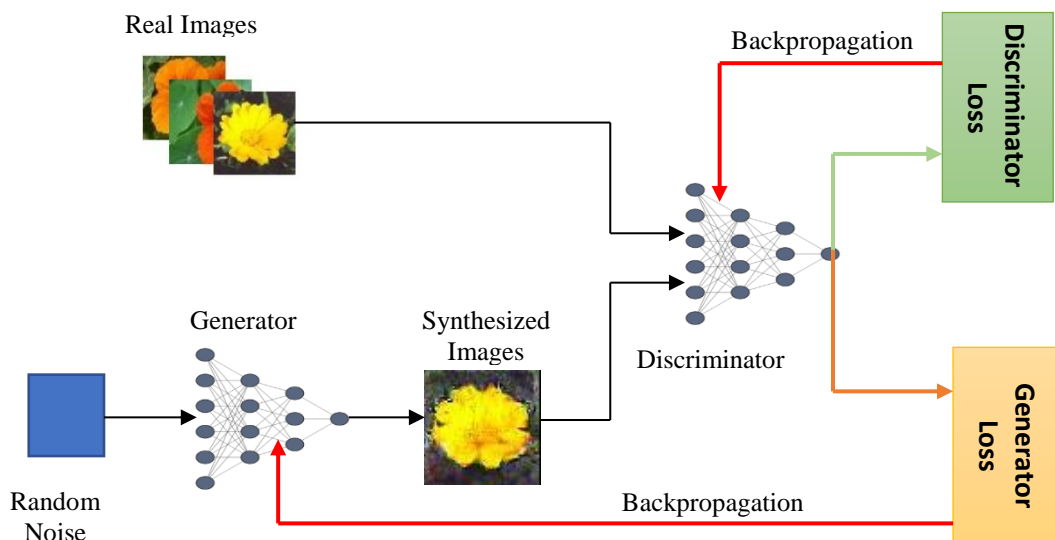


Figure.3.1 Framework of Generative Adversarial Networks

Generative Adversarial Network (GAN) is a deep learning model that involves separate models: generator and discriminator. The generator learns to create new data distribution similar to training data, while the discriminator tries to differentiate

between the generated samples and real data. The generator takes random noise as input and creates new samples, such as images, music, or text. The discriminator then evaluates the generated sample and determines whether it is real or fake. The generator is trained to generate more and more realistic images, while the discriminator is trained to become better at distinguishing real and fake samples.

GANs is designed to train its components simultaneously in a minimax game, where the generator tries to generate realistic data and the discriminator tries to differentiate between the real and fake data. These two models are trained together in a process called adversarial training. This training continues until the discriminator misclassify between real data and generated data. The training process is repeated to improve the performance of these two networks over time. Gradually, the generator network becomes better at generating fake samples that are indistinguishable from real data. The loss function of GANs is commonly known as the adversarial loss or the GAN loss, and it is defined as the following equations:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{x \sim p_z(z)} + [\log(1 - D(G(z)))] \quad (3.1)$$

Where,

x = a sample from the real data distribution,

z = a random noise vector input to the generator

$\mathbb{E}_{x \sim p_z(z)}$ = the expected values overall generated images

$\mathbb{E}_{x \sim p_{data}(x)}$ = the expected values overall real images

3.2 Types of Generative Adversarial Networks

Generative Adversarial Networks (GANs) were first proposed in 2014 by Ian Goodfellow and since then this model becomes an opened research area. With the introducing of this models, Generative Models had started showing best results in creating realistic images, text and audio. Therefore, GANs has shown many success in Computer Vision. In this section, some of the most popular GAN architectures are described.

3.2.1 Conditional GAN

In cGANs [30], an additional input variable is applied, meaning that both the generator and discriminator are conditioned on some extra information such as class labels from other modalities. The generator takes as input both the random noise vector and the conditioning information, and produces an output that is intended to minimal distribution of the training data given conditioning information. The discriminator takes both the real data with conditioning information or the fake data generated by the generator with conditioning information as inputs, and produces a binary output to predict whether generated images are real or not. Taking advantages with the use of the extra encoded information, it is able to execute both unimodal image datasets and multimodal datasets such as MSCOO that contains images paired with text descriptions. Even it enhances the discriminative ability of the discriminator due to applying encoded labels, some of the generated labels still lose relations with images.

3.2.2 Deep Convolutional GANs (DCGANs)

DCGANs [5] is built using convolutional neural networks. In this model, the generator contains transposed convolutional layers, batch normalization and ReLU activations. The generator network takes random sampled vectors and forward this input through a series of transposed convolutional layers to transform it into meaningful output. On the other side, the discriminator network which is built with convolutional layers takes both the fake images and real images in the dataset to determine the generated image is real or fake by comparing their probability distribution. The ConvNets are implemented by convolutional with stride operation instead of max pooling, which is in fact replaced by convolutional. The dimension of the generated images is a 3x64x64 RGB image. DCGANs have been used to create a varieties of images, including faces, bedrooms, and landscapes.

3.2.3 CycleGAN

It [23] is a powerful deep learning model that can be used for image-to-image translation without requiring training that contains paired images. It works by learning two generators and two discriminators. The two generators learn to map images from one object to the other object (e.g., horse to zebra), and the two discriminators learn to

distinguish between real images and generated images in each domain. The cycle consistency loss is used to show that the translated images are relevant with the input images. The main important point of this GAN is that it can do image-to-image translation on an unrelated image where there is no sophisticated relationship between the input image and output image. CycleGAN has been used in a variety of application areas, including style transfer, and image-to-image translation.

3.2.4 StyleGAN

It [50] can generate high-quality and diversity images with a controllable style such as the age, gender, and pose of a person and be used for image manipulation, such as style transfer and image morphing. There are a lot of advancements on state-of-the-art models that can create synthetic data including synthetic images. However, most of these advancements applied on the discriminator side which refines to improve the image generation ability of the generator. Style GAN makes many modifications in the generator portion which enables to synthesize high-quality realistic images. Style GAN makes improvements extended from the baseline progressive GAN and introduced some features on the generator side. Moreover, it gradually increases the resolution of the generated images from very low dimension (4×4) to high dimension (1024×1024). It built an alternative generator architecture with the use of adaptive instance normalization. In addition, it generates the images from a fixed value vector which is not randomly sampled latent variable as in regular GANs. The stochastically sampled latent variables are used as style vectors in the adaptive instance normalization at each resolution after being transformed by an 8-layer feedforward networks. Finally, it composes mixing regularization, which combines two style latent variables during training. The advantage of this style vector grants control over the characteristic of the generated image.

3.2.5 BigGAN

BigGAN [1] is designed to generate high-quality and diversity images by learning the features of large-complex datasets such as ImageNet. This model is trained on real samples at 128×128 resolutions of ImageNet dataset. The SA-GAN architecture with self-attention block is used as a baseline to build this model by adding spectral normalization allowing to control over the trade-off between sample fidelity and

diversity by reducing the amount of randomness in the generator's input. This model dramatically scaled up two to four times in parameters, 8 times of batch-size, and 50% at the width of each layer as compared to baseline. According these results, there has a big improvement on the IS score of the model. The other feature is the use of a conditional batch normalization technique, which allows the generator to produce images with specific attributes, such as the pose or background of an object. BigGAN is a powerful tool for image generation and has opened up new possibilities for applications such as art, design, and visual storytelling.

3.2.6 Super Resolution GANs, or SRGANs

SRGANs [10] plays an important area in the area of image processing because it can upscale the image with low-resolution into high-resolution while preserving and enhancing the image details. Recent generative adversarial networks (GANs) have achieved remarkable results on low-resolution images with small samples. Super Resolution GANs have shown impressive results in enhancing the resolution of images, and have been applied in a wide range areas, such as medical imaging, surveillance, and video compression. For example, in the medical field, generating higher-quality images can help doctors to be accurately detect diseases. However, they need large amount of high-quality data and are higher computational cost for training.

3.2.7 Generative Adversarial Text to Image Synthesis

Text-to-image synthesis [43] is the process of taking a text description as input and generates a corresponding image that matches the description. This technology has been rapidly evolving in recent years due to the significant advances in the field of deep learning and computer vision. GANs are a popular approach for text-to-image synthesis, where a generator network learns to produce images that match a given textual description, while a discriminator network tries to differentiate between real and generated images. The training of these models requires large datasets of images paired with text to learn the relationship between textual descriptions and visual features. One of the challenges is that it requires a large dataset of text-image pairs for training. Additionally, it can be difficult to capture the nuances of language in the text descriptions, which can generate the images that do not match with the text description.

With further advancements, this field has the potential to be deployed in various fields, such as content creation, design, and e-commerce.

3.2.8 Adversarial Autoencoder (AAE)

AAE [2] is a type of generative neural network architecture that combines the principles of autoencoders and adversarial training. An autoencoder is a type of deep learning that is trained to encode input data into representation with low-dimension and then decode it back to obtain its original shape, while minimizing the reconstruction error. In an AAE, the encoder and decoder components of the autoencoder are trained together with a discriminator network. The discriminator takes the encoded latent code z (fake data) from the autoencoder and a random vector z selected from the real data distribution and verifies the generated output is genuine or not. The ultimate goal of training an AAE is to create new data samples that are similar to the original dataset but also diverse and unique. AAEs have found applications in image synthesis, text generation, and anomaly detection.

3.3 Application Areas of Generative Adversarial Networks

GANs have been used to generate realistic samples such as images, videos, and music, and have shown promising results in other applications (e.g.; natural language processing and drug discovery). However, GANs can be challenging in training because it can be computationally expensive and requires large amounts of data. Despite these challenges, GANs have the potential to revolutionize many industries by creating new data that can be used for various purposes, such as training other machine learning models. GANs have found applications in a variety of fields, including:

- **Image synthesis:** It can be used to generate realistic images of objects, animals, and people. They have been also used to create high-quality images for video games, movies, and virtual reality applications.
- **Data augmentation:** GANs can generate new data that can be applied to augment the training data for machine learning models to improve the performance of these models, especially in cases where there is limited amount of training data are available.

- Style transfer: GANs can be implemented to transfer the style of one image to another image. This technique has been used to create artistic images, such as paintings and drawings.
- Text generation: GANs can create new text that is similar to the training data. They have been used to create realistic text for chatbots and other natural language processing applications.
- Video synthesis: GANs can generate new video frames that are similar to the training data. This can be used to create high-quality video content for movies, TV shows, and other applications.
- Drug discovery: GANs can be used to generate new molecules that have specific properties, such as being effective against a particular disease. This can help speed up the drug discovery process.
- Music generation: GANs can synthesize music that is similar to the training data. They have been used to create new music for video games and other applications.

3.4 Representations

Deep learning models typically require large amount of training data to learn complex patterns and make accurate predictions. However, raw data can be difficult for a machine learning algorithm to process, due to its high dimensionality and complexity. Therefore, the goal of representation learning is to extract the most relevant and useful features from raw data, which can then be used as input to a machine learning model. In this section, the most commonly used methods of (1) unimodal embedding (representation), and (2) multimodal embedding (representation) are described.

3.4.1 Unimodal Embedding

In deep learning, unimodal embedding [13] means the representation of data from a single modality (such as text, images, or audio) in a continuous vector space with a fixed dimension. This embedding process allows more efficient processing and analysis of the data, as well as facilitating comparisons and similarity calculations between different types of data. Unimodal embeddings are one type of deep learning

technique used to extract useful information from raw data in a single modality, such as images or text. Two types of unimodal embeddings used in this thesis are described among several types of unimodal embeddings. These two methods are (1) language representations (or) text embeddings and (2) visual representations (or) image embeddings.

3.4.1.1 Language Representations

Text embeddings refers to numerical representations of each word or sequence of words in a sentence. The goal of text embeddings is to capture the semantic and syntactic relationships between words, so that similar words or phrases are mapped to similar vectors. It can be extracted from neural network language model (NNLM) by estimating the probability of sequence of text using a chain rule. There are several RNN-based NNLMs that can be used for text embeddings such as simple RNNs, long short-term memory (LSTM) networks, and gated recurrent units (GRUs). Among them, LSTM and GRU networks are most commonly utilized methods for text embedding due to their ability to capture long-term dependencies in the input sequence.

In addition to NNLMs, Bag-of-Words (BoW) Embedding are used to encode each word in a vocabulary as one-hot vector, where each vector has a length equal to the size of the vocabulary. The value of each element in the vector is either 0 or 1, indicating whether the corresponding word is present or absent in a given text document. Word embedding is a technique used in natural language processing to represent words or phrases in a high-dimensional vector space, where the position of a word in the vector space is determined by its context and meaning. Word2Vec, GloVe, FastText are other commonly used methods for word embeddings.

Pretrained model-based text embeddings also called transformer-based models, such as BERT and GPT, use a self-attention mechanism to learn contextualized vector representations of words and sentences. These embeddings are powerful and can capture complex relationships between words, making them suitable for a wide range of NLP tasks, such as sentiment analysis, named entity recognition, and machine translation.

3.4.1.2 Visual Representations

Image embeddings [13] are vector representations of images in a high-dimensional space by using deep neural networks. These embeddings capture the underlying features of the images, such as edges, textures, shapes, and colors, and represent them in a compact and meaningful way that can be used for various image-related tasks. Image embeddings have been widely used in computer vision applications, such as image classification, object detection, image search, and image retrieval. They are also used in natural language processing (NLP) applications, such as image captioning and visual question answering, where images are combined with textual data.

One popular approach of image embeddings is convolutional neural networks (CNNs). They are designed to process and analyze images by convolving filters over the image and pooling the results to extract the most informative features. The activations from one of the intermediate layers of a CNN can be used as the image embedding. Another approach is to use pre-trained models which have been trained on large-scale image datasets to extract the images features and represent them as a vector of numbers.

3.4.2 Multimodal Embedding

Multimodal representation [13] refers to the process of representing and encoding data from multiple modalities, such as images, text, and audio, into a common vector space. The goal of multimodal representation is to capture the relationships between two different modalities and enable more effective analysis and processing of the data. Multimodal embeddings have a wide areas of applications, image captioning, visual question answering, multimodal sentiment analysis, and speech recognition. By combining information from multiple modalities, multimodal embeddings can improve the performance of these tasks and enable more natural and intuitive interaction between humans and machines.

3.4.2.1 Joint Representations

Joint representation [14] refers to a type of representation that combines information from multiple sources or modalities into a common feature space. The goal of joint representation is to learn a shared representation that captures the underlying structure

and relationships between the different sources of information. In the context of machine learning, joint representations are often used in multimodal learning, where data is represented using different multimodalities.

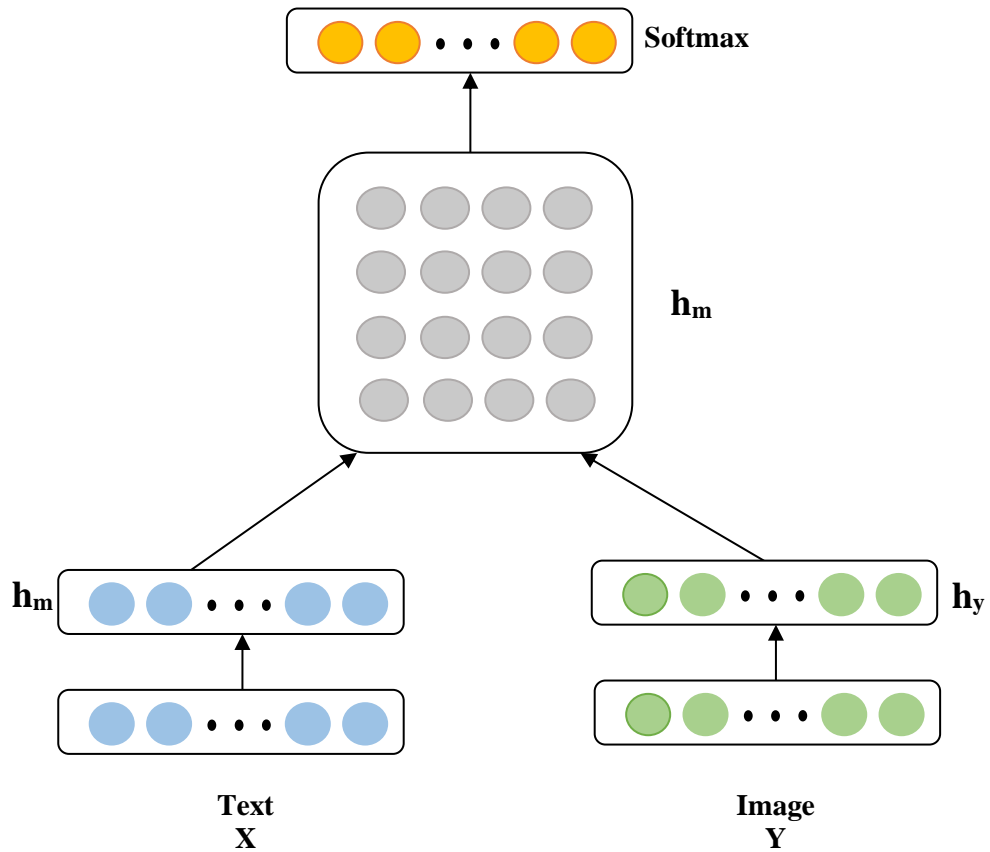


Figure 3.2 Joint Representations

3.4.2.2 Coordinated Representations

Coordinated multimodal representations [14] refer to a type of representation that captures the coordination and interaction between multiple modalities in a coordinated manner. Unlike joint representations, which combine information from multiple modalities into a common feature space, coordinated representations explicitly model the interaction and coordination between modalities. In the context of machine learning, coordinated multimodal representations are often used in tasks such as speech recognition, where the audio and visual modalities are highly coordinated. Dynamic fusion: In dynamic fusion, the weights used to combine the modalities are learned during training, and can be adjusted at runtime based on the input data. The weights can be determined using a gating mechanism, such as a sigmoid function or a learned

attention mechanism. The gating mechanism takes the input from each modality and generates a weight for that modality, which is then used to combine the modalities.

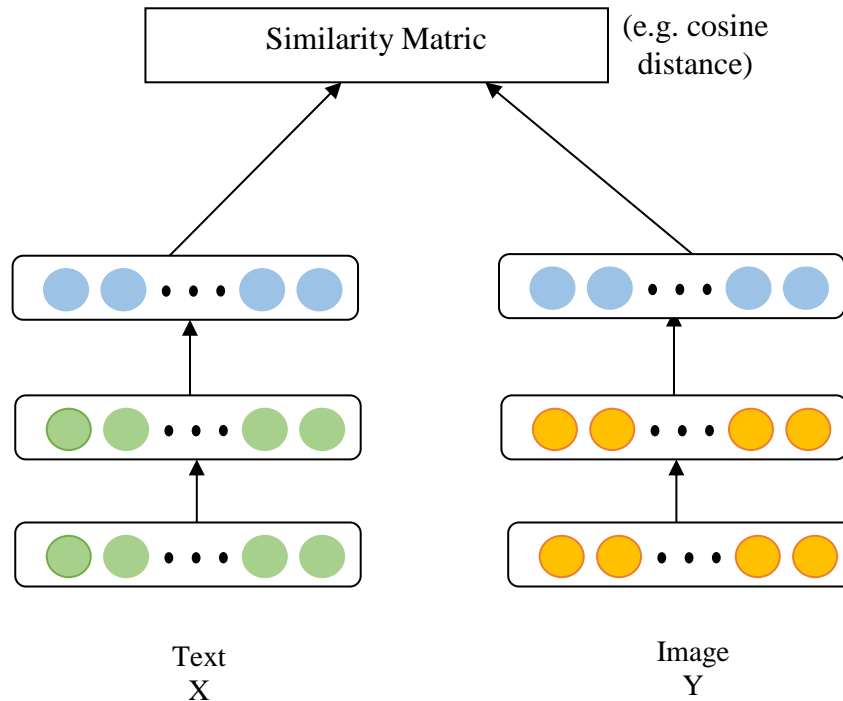


Figure 3.3 Coordinated Representation

3.5 Other Visual Language Generative Models

The other language generative models which generate images from text based on diffusion methods are described in this section.

3.5.1 DALL-E

DALL-E [6] tackles the visual generation (VG) problem by being able to synthesize images that capture the semantic of each words in the descriptions. DALL-E maps the images to tokens by applying a discrete variational autoencoder (dVAE [12]) that essentially utilizes a discrete latent space. Each words in a sentence is tokenized by using byte-pair encoding. The image and text tokens are concatenated and transformed to single data stream.

DALL-E operates this data stream by applying an autoregressive transformer in order to build the joint data distribution of images and text. In the transformer's

decoder, each image can attend to all text tokens. At inference time, the tokenized caption with a sample noise from the dVAE are concatenated, and passed this concatenated stream to the autoregressive decoder to create a novel token image.

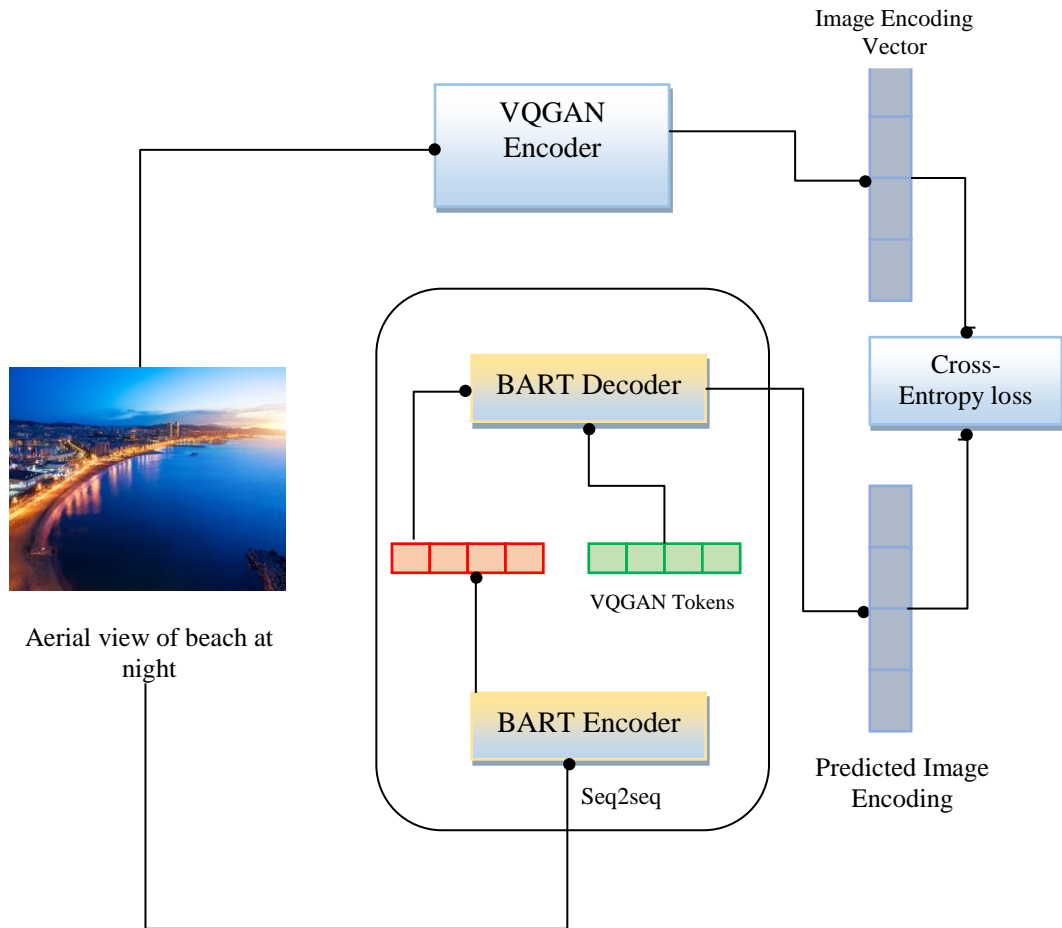


Figure 3.4 DALL-E

3.5.2 Guided Language-to-Image Diffusion for Generation and Editing (GLIDE)

GLIDE [7] is another generative model that seems to outperform recent T2I models and a diffusion model. As an alternative to creation images from text, it may be used to modify existing images by utilizing text descriptions prompts to insert new shapes, add shadows and reflections, inpainting, etc. It can change basic line drawings into photorealistic images, and enhances the quality of complicated objects.

Diffusion models implements by slowly introducing random noise to data in sequential structure obtained from Markov chain. Then, they learn to reverse the process in order to reconstruct realistic samples from the noise. They are sampled from a known data distribution created after a sequence of diffusion stages instead of

sampling from the unknown data distribution. This model takes images as inputs and can generate novel ones. In addition, it can also create image based on a specific text descriptions. It experiments with a variety of methods to “guide” the diffusion models. GLIDE results are even more impressive and more realistic than DALLE.

3.6 Text Encoder and Image Encoder

Convolutional Neural Network (CNN) [40] and Recurrent Neural Network (RNN)[34] are powerful deep learning techniques applied in various applications such as computer vision, natural language processing, and speech recognition. However, CNNs are primarily used for image classification, object detection, and segmentation tasks, while RNNs are more suitable for sequence modeling tasks such as language modeling, speech recognition, and machine translation. There are many variations of RNN to build language model such as Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM). Among them, text encoder is built using Bi-LSTM because Bi-LSTM can give more accuracy for large dataset. Therefore, we modeled CNN as image encoder and Bi-LSTM as text encoder to extract the features of Myanmar text description. The working principles of these two neural networks are discussed in this section.

3.6.1 Convolutional Neural Networks

Convolutional Neural Network is designed to extract features of data that has a grid-like structure, such as images, where each pixel represents a feature. CNNs have become the state-of-the-art method for various images and video-related task. The four main operation of convolutional neural network are:

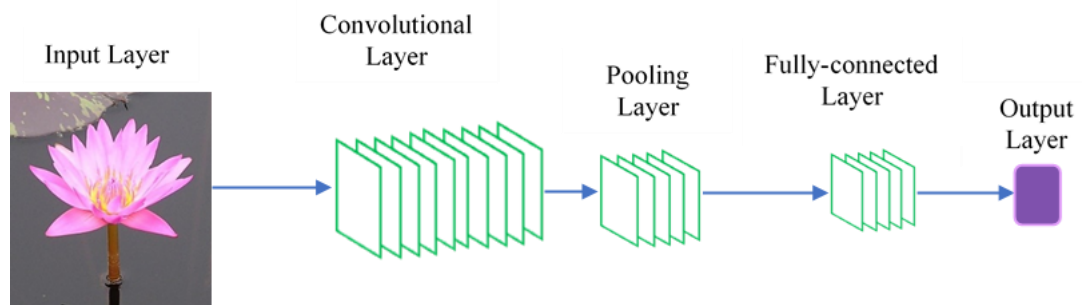


Figure 3.5 Simple Framework of Convolutional Neural Network

The four main operations of this network are:

- **Convolution:** This operation involves the application of a set of learnable filters to the input data. Each filter slides over the input data and produces a corresponding output feature map by performing the dot product between the weights and input the data. This operation enables the network to learn local patterns and features in the input data.

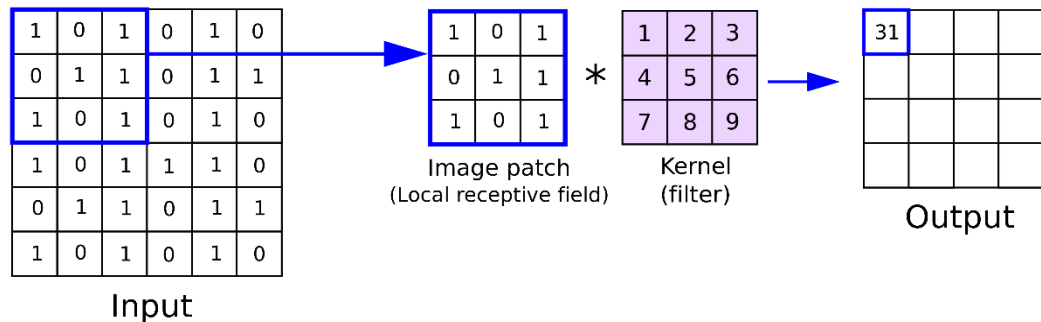


Figure 3.6 Convolution

- **Non-linearity (Activation):** After each convolution operation, a non-linear activation function is applied to the output feature map. This introduces non-linearities into the network and enables it to learn more complex and abstract features.
- **Pooling:** This operation involves down-sampling the output feature maps achieved from the convolution and activation operations. This reduces the dimensions of the feature maps while retaining the most critical features. The two main operations of pooling are (1) max-pooling, and (2) average pooling.

(1) Max-pooling

This operation is most commonly utilized operation and this operation is performed by selecting the maximum value within a local region of the feature map.

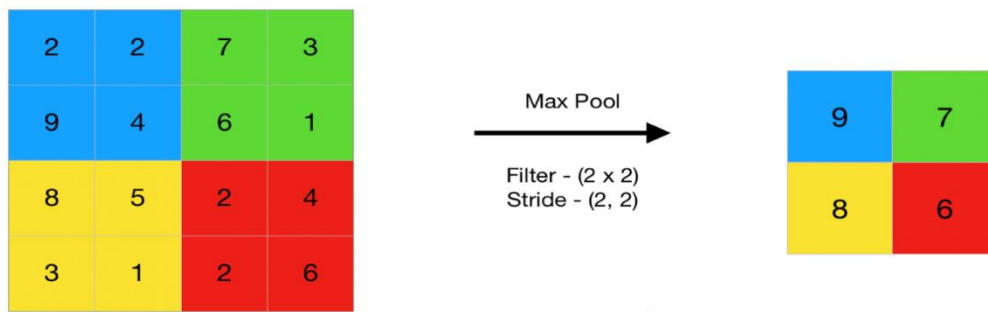


Figure 3.7 Max-Pooling

(2) Average-Pooling

In average pooling, a rectangular window of a fixed size is moved over the input feature map, and computed the average value of the elements inside that window and place the corresponding pixel in the output feature ma

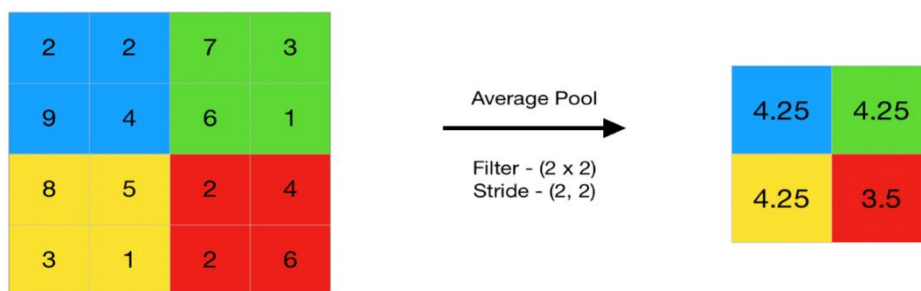


Figure 3.8 Average-Pooling

- Fully-connected layers: It is the final stage in the convolutional neural network and this layer takes the feature maps from the previous convolutional and pooling layers and flattens them into a vector, which is then passed through a group of neurons that are fully connected to the neurons in the previous layer. Each neuron in this layer receives inputs from the previous layer, and produces a single output value that represents the activation of that neuron. This layer is typically followed by one or more output layers that perform the final classification of the input. The output layer can have one or more neurons, depending on the number of classes to be classified. For example, in a binary classification problem, the output layer would have a single neuron that produces a value between 0 and 1, indicating the probability of the input belonging to one of the two classes.

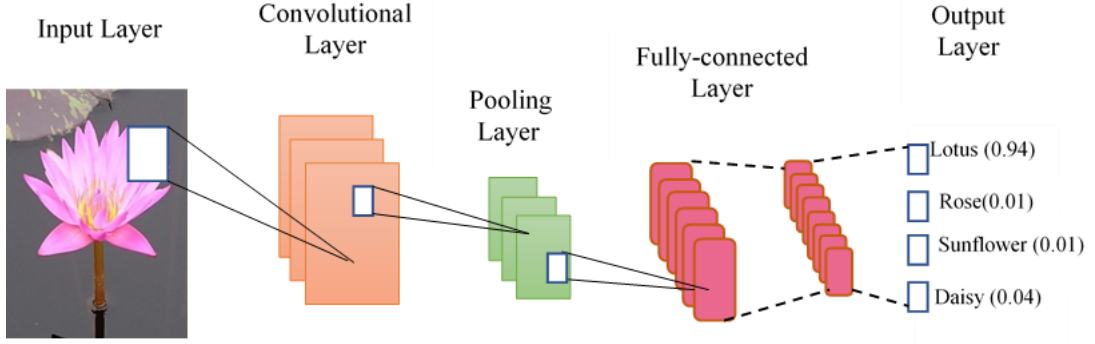


Figure 3.9 Fully-Connected Layers

3.6.2 Bi-directional Long Short-Term Memory (Bi-LSTM)

Bi-LSTM [31] is one type of recurrent neural network (RNN) that is capable of processing sequences of data in both forward and backward directions. The text encoded by one of the encoding-techniques are fed as input to Bi-LSTM encoder. The encoder processes the input sequences in both the forward ($\vec{h}_t = \vec{h}_1, \vec{h}_2, \dots$) and backward directions ($\overleftarrow{h}_t = \overleftarrow{h}_1, \overleftarrow{h}_2, \dots$) using two separate LSTM layers and produces a sequence of hidden states for each time step. The expression to compute the hidden states of Bi-LSTM are:

$$\vec{h}_t = \sigma(W_{\vec{h}} [\vec{h}_{t-1}, w_t] + b_{\vec{h}}) \quad (3.2)$$

$$\overleftarrow{h}_t = \sigma(W_{\overleftarrow{h}} [\overleftarrow{h}_{t-1}, w_t] + b_{\overleftarrow{h}}) \quad (3.3)$$

$$h_t = W_{\vec{h}} \vec{h}_t + W_{\overleftarrow{h}} \overleftarrow{h}_t + b_y \quad (3.4)$$

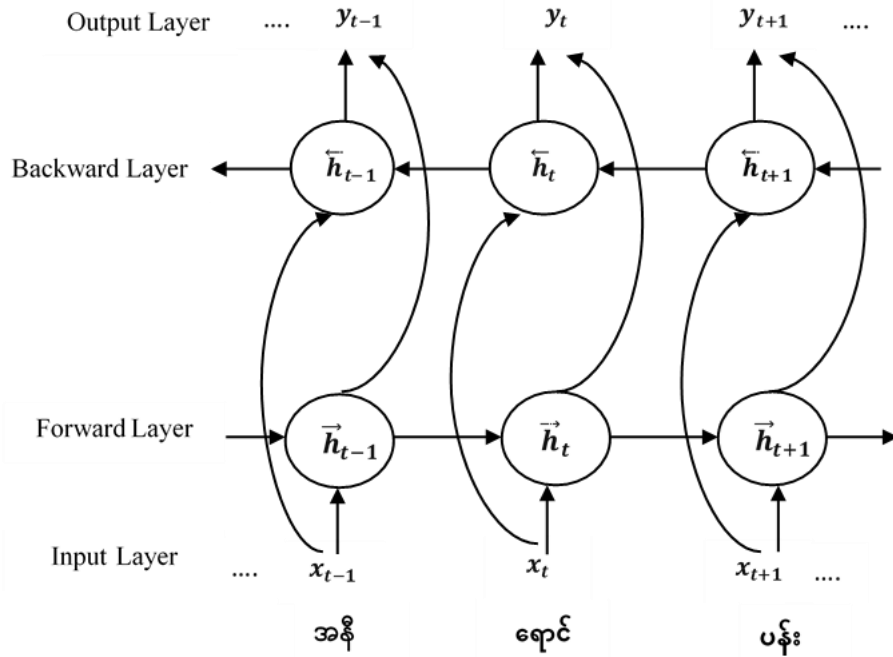


Figure 3.10 Framework of Bi-LSTM

The outputs of these two layers are concatenated at each time step to produce semantic meaning of a word for that time step. Meanwhile, the global sentence is output by concatenating the last hidden states of Bi-LSTM.

3.7 Text Encoding Techniques

Text encoding is often used as a preprocessing step in machine learning models, where the textual data needs to be converted into a numerical representation for further analysis or model training. It allows the model to work with structured data instead of raw text, enabling various mathematical operations and computations to be performed on the encoded sequences. Popular encoding methods are index-based encoding, one-hot encoding and more advanced techniques like word2vec, GloVe, and BERT, which capture semantic relationships between words and preserve contextual information within the encoded representations. These methods go beyond simple index assignments and leverage deep learning techniques to create rich vector representations of words or tokens. Among these encoding, the two encoding techniques used to experiment this research work is mainly focused on this chapter.

3.7.1 Indexed-Based Encoding

Indexed-based encoding is a technique used in various fields such as natural language processing, machine learning, and data compression. It involves representing data or text by assigning unique indices or numbers to individual elements, tokens, or symbols. In the context of natural language processing, indexed-based encoding is commonly used in language models and text classification tasks. It involves creating a vocabulary of unique words or tokens from a given corpus, assigning a unique index to each token, and representing text as sequences of these indices.

For example, consider the sentence: "I love to play soccer." Using indexed-based encoding, we can create a vocabulary consisting of unique words from this sentence: {I, love, to, play, soccer}. We assign an index to each word: {0, 1, 2, 3, 4}. Thus, the encoded representation of the sentence would be [0, 1, 2, 3, 4].

3.7.2 Word2Vec

Word2Vec [24] is a popular technique in natural language processing (NLP) used for generating word embeddings, which are dense vector representations of words. It was developed by Tomas Mikolov and his colleagues at Google in 2013. The basic idea behind Word2Vec is to learn word embeddings by predicting the context in which words appear in a given corpus. It operates on the assumption that words appearing in similar contexts are likely to have similar meanings. Word2Vec achieves this by training a neural network model on a large amount of text data and learning the word embeddings as a byproduct of the network's training process.

There are two main architectures for Word2Vec: Continuous Bag of Words (CBOW) and Skip-gram.

1. CBOW: In this architecture, the model predicts target words based on the context words surrounding it. The context words are used as input, and the target word is predicted as the output. CBOW is generally faster to train and works well for rare words.
2. Skip-gram: In contrast to CBOW, the skip-gram architecture predicts the context words given a target word. This model predicts the surrounding context-words by giving the target word as input. Skip-gram performs better on larger datasets and captures more information about infrequent words.

During the training process, Word2Vec model adjusts the word embeddings so that words with similar meanings are closer together in the embedding space. After training, these word embeddings can be used in various NLP tasks, such as semantic similarity, sentiment analysis, and document classification.

Word2Vec embeddings have several advantages. They capture semantic relationships between words, allow vector operations like addition and subtraction (e.g., "king" - "man" + "woman" \approx "queen"), and can handle out-of-vocabulary words by generalizing from similar words in the training corpus. Pretrained Word2Vec models are available for many languages and domains, allowing researchers and developers to leverage the power of word embeddings without training from scratch.

3.8 Summary

This chapter described the theoretical concepts applied in this research. Generative Adversarial Network is used to generate the images from Myanmar language text. The other types of GAN, useful areas of GAN, and data representation techniques used to encode sequence data are also described. In this implementation, two types of modalities (images and text) are used in training of GAN. Text encoder and image encoder are jointly trained using cross-modal attention (multimodal representation) before training of Myanmar T2I. This pretrained text encoder is used to extract the sentence vectors of Myanmar text descriptions. Image encoder is to extract the features of the generated images from GAN during the evaluation of the similarity between the generated image and input text description.

CHAPTER 4

DATASET PREPARATION

This chapter presents the two datasets with different languages used to implement text-to-image tasks. Oxford-102 flower dataset [28] annotated in Myanmar is to conduct Myanmar text-to-image. As an alternative experiment, Caltech-UCSD Birds-200-2011 dataset [47] annotated in English is used to prove the proposed model also gives improvement in the creation of images from different language.

4.1 Dataset for Myanmar T2I

This section contains annotation of Myanmar captions for each image in Oxford-102 flower dataset and preprocessing step of this caption corpus.

4.1.1 Building Annotated Myanmar Caption Corpus

To implement Myanmar text-to-image synthesis, images with labeled captions dataset are needed. There is the first Myanmar annotated images dataset extended from Flickr28k. Therefore, this research was implemented by using this dataset. However, the generated images are noisy and imprecise in shape due to its limitations in terms of diversity, object annotations, and spatial information. For this reason, this research is conducted using Oxford-102 flowers dataset. This dataset contains 102 categories, 8189 images. Firstly, the English captions corpus are translated to Myanmar captions by using machine translation. In this approach, the translated sentence is mismatch and inaccurate with the images because training dataset is not relevant with this caption corpus. Therefore, Myanmar captions corpus for each image are manually constructed by focusing their features without directly using or translating English descriptions from the original dataset. There are 5 annotated captions for each image in this dataset. The total number of sentences in Myanmar caption corpus is 40945. Example of images with annotated captions is shown in Figure 4.3.

4.1.2 Caption Preprocessing

Myanmar language is formed with syllables and which is unique and complex in nature. There are 33 consonants, 8 vowels (free standing and attached), 2 diacritics,

11 medials, a vowel killer or ASAT, 10 digits and 2 punctuation marks in Myanmar script. Basically, it does not contain boundary limiter between words or syllables. Each word is formed with syllables and each syllable contains sub sequent of syllables. For example, a Myanmar word “ပန်းပွင့်” (“flower” in English) has two syllables. In the first syllable “ပန်း”, the sub syllables are consonant (“ပ”) + consonant (“န”) + vowel killer (“်”) + diacritics (“း”). In the second syllable, “ပွင့်”, the sub syllables are consonant (“ပ”) + medial (“ွ”) + consonant (“င”) + vowel killer (“်”) + diacritics (“့”). In this research, Myanmar captions are segmented using word segmentation because text preprocessing is essential prior natural language processing. Word segmentation is performed by using python with pyidaungsu library. The samples of word segmentation are shown in Figure 4.1:

- ❖ ခရမ်း ရောင် ပွင့် ချပ် နှင့် ပန်း တွင် ပန်း ရောင် လိုင်း များ ရှိ တယ်
- ❖ ခရမ်း ရောင် ပန်းပွင့် တွင် အဖြူ ရောင်စင်တာ နှင့် အနားတွန့်ပွင့် ချပ် များ ရှိ တယ်
- ❖ ပန်းပွင့် ၌ တောက်ပ သော အနီ ရောင် ပွင့် ချပ် နှင့်အဝါ ရောင်စင်တာ ရှိ တယ်
- ❖ ပန်းပွင့် သည် အနားသတ် တွင် အနီ ရောင် ရှိ ပြီး အလယ် တွင် အဝါ ရောင် ရှိ တယ်
- ❖ ပန်းပွင့် တွင် သေးငယ် သော အနီ ရောင်အဝါ ရောင် ပွင့် ချပ် များ ရှိ တယ်
- ❖ ပန်းပွင့် ၌ တောက်ပ သော အနီ ရောင် ပွင့် ချပ် နှင့်အဝါ ရောင်စင်တာ ရှိ တယ်
- ❖ ပန်းပွင့် သည် အနားသတ် တွင် အနီ ရောင် ရှိ ပြီး အလယ် တွင် အဝါ ရောင် ရှိ တယ်
- ❖ ပန်းပွင့် တွင် သေးငယ် သော အနီ ရောင်အဝါ ရောင် ပွင့် ချပ် များ ရှိ တယ်
- ❖ ဤ ပန်းပွင့် တွင် အနီ ရောင် လိုင်း ပါ သော ပန်း ရောင် ပွင့် ချပ် ငါး ခု ရှိ တယ်
- ❖ ပန်း ရောင် ပွင့် ချပ် ငါး ခု ရှိ သော ပန်း တွင် အနက် ရောင်စင်တာ ရှိ တယ်

Figure 4.1 Sample of Segmented Word

Images

Captions



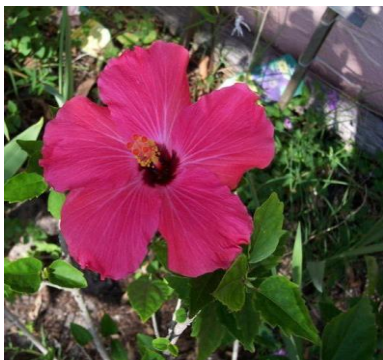
1. ခရမ်းရောင်ပွင့်ချပ်နှင့်ပန်းတွင်ပန်းရောင်လှိုင်းများရှိတယ်
2. ခရမ်းရောင်ပန်းပွင့်တွင်အဖြူရောင်စင်တာနှင့်အနားတွန်းပွင့်ချပ်များရှိတယ်
3. ပန်းပွင့်သည်ခရမ်းရောင်ရှိပြီး အလယ်တွင် အဖြူရောင် ရှိတယ်
4. ခရမ်းရောင်ပန်းပွင့်၏ပတ်လည်တွင်အစိမ်းရောင်အရွက်ရှိတယ်
5. ပန်းပွင့်တွင်ခရမ်းရောင်ပွင့်ချပ်အဖြူရောင်ဝတ်ဆံနှင့်ပန်းရောင်လှိုင်းရှိတယ်



1. ပန်းပွင့်၌တောက်ပသောအနီရောင်ပွင့်ချပ်နှင့်အဝါရောင်စင်တာရှိတယ်
2. ပန်းပွင့်ပေါ်တွင်အနီရောင်မှအဝါရောင်သို့ပြောင်းသွားသောပွင့်ချပ်များရှိတယ်
3. ပန်းပွင့်သည်အဝါရောင်အနီရောင်တို့နှင့်အရောင်စုံသောပွင့်ချပ်ရှိတယ်
4. ပန်းပွင့်သည်အနားသတ်တွင်အနီရောင်ရှိပြီးအလယ်တွင်အဝါရောင်ရှိတယ်
5. ပန်းပွင့်တွင်သေးငယ်သောအနီရောင်အဝါရောင်ပွင့်ချပ်များရှိတယ်



1. ပန်းပွင့်၌တောက်ပသောအနီရောင်ပွင့်ချပ်နှင့်အဝါရောင်စင်တာရှိတယ်
2. ပန်းပွင့်ပေါ်တွင်အနီရောင်မှအဝါရောင်သို့ပြောင်းသွားသောပွင့်ချပ်များရှိတယ်
3. ပန်းပွင့်သည်အဝါရောင်အနီရောင်တို့နှင့်အရောင်စုံသောပွင့်ချပ်ရှိတယ်
4. ပန်းပွင့်သည်အနားသတ်တွင်အနီရောင်ရှိပြီးအလယ်တွင်အဝါရောင်ရှိတယ်
5. ပန်းပွင့်တွင်သေးငယ်သောအနီရောင်အဝါရောင်ပွင့်ချပ်များရှိတယ်



1. ဤပန်းပွင့်တွင်အနီရောင်လှိုင်းပါသောပန်းရောင်ပွင့်ချပ် ငါးခုရှိတယ်
2. ပန်းပွင့်တွင်ပန်းရောင်ပွင့်ချပ်ငါးခုအနီရောင်အစင်းကြောင်း ရှိတယ်
3. ပန်းရောင်ပွင့်ချပ်ငါးခုရှိသော ပန်းတွင် အနက်ရောင်စင်တာ ရှိတယ်
4. ပန်းပွင့်တွင်ပန်းရောင်ပွင့်ချပ်ငါးခုအနီရောင်အစင်းကြောင်း ရှိတယ်
5. ပန်းရောင်ပွင့်ချပ်ငါးခုရှိသော ပန်းတွင် အနက်ရောင်စင်တာ ရှိတယ်

Figure 4.2 Sample of Flower Images with Annotated Captions

4.2 Dataset for English T2I

CUB-200-2011 is an extended version of CUB-200 and contains 200 bird species. The versions roughly doubles the number of images per speices and adds new part localization annotations. These images are annotated with part locations and attribute labels. The dataset contains 11,788 images and 10 annotated sentences for each image. Example of bird image and its related captions are described in Figure 4.3.

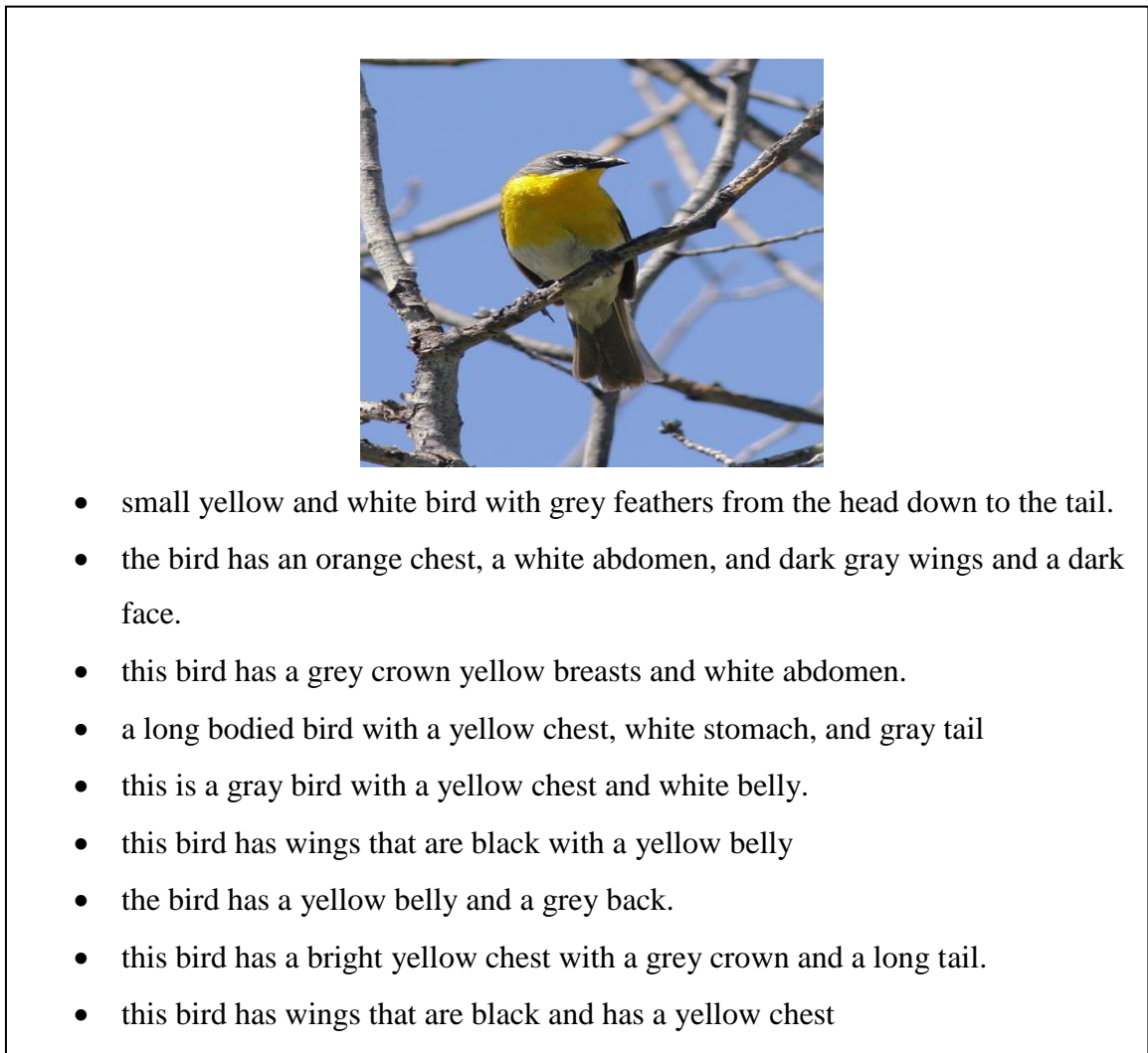


Figure 4.3 Sample of Bird Images with Annotated Captions

4.3 Image Preprocessing

Image normalization is an important preprocessing step in many image processing and computer vision tasks, as it helps to improve the quality and consistency of the data, and can ultimately lead to better results in the downstream analysis. It is a process of changing the range of pixel intensity values to make the image more familiar or normal to the senses. Normalizing images before processing is a good practice that can lead to more accurate and robust results for feature extraction or image segmentation. The training images are preprocessed by using Pillow (Image library) and NumPy library with python. After normalization, the normalized images are fed as input generative adversarial networks (GANs) during training. The original and preprocessed images are shown in Figure 4.5.

Original Images



Preprocess Images

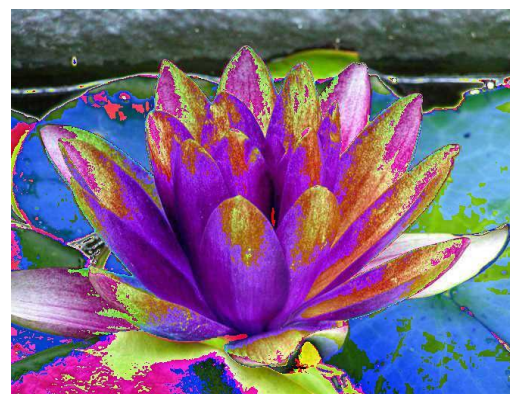


Figure 4.4. Sample of Original Images and Preprocessed Images

4.4 Summary

The steps for construction of annotated image dataset for Myanmar T2I are described in this chapter. In addition, the nature and the preprocessing of Myanmar language descriptions are also discussed. Furthermore, the ways of converting index number from segmented words that will give as input to LSTM encoder contains in this chapter. Finally, the preprocessing step of the real images in the training dataset is presented in this chapter.

CHAPTER 5

DEEP CONVOLUTIONAL GENERATIVE ADVERSARIAL NETWORKS FOR MYANMAR T2I

The implementation of Myanmar T2I using GANs and word2vec is described in this chapter. Text to image synthesis is the process of generating images from textual descriptions. It involves using artificial intelligence algorithms to convert a given text input into a corresponding image output. The goal is to create a visually realistic image that accurately represents the content of the input text. This is a challenging problem because it requires the AI system to understand the meaning of the text and translate it into visual content. This task contains two parts: (1) language understanding and (2) image generation.

There are various approaches to implement text to image synthesis. Generative Adversarial Networks (GANs) have become popular among state-of-the-art models because of their ability to generate high-quality realistic images. GANs can capture complex visual patterns and textures, which makes them well-suited for image generation tasks. They can learn from large datasets of images and can then generate new images that are visually similar to the training data. Therefore, the first Myanmar T2I was implemented using GANs to generate images from Myanmar text. For language understanding, word2vec model is used to extract features vectors of Myanmar sentence.

5.1 Deep Convolutional GANs (DCGAN) based Myanmar T2I

At the generator side, the noise-vectors are sampled by using Gaussian distribution and randomly selected one text descriptions from five captions paired with images in the training set are passed as input to this network during training. The word-vectors of these captions were already extracted using word2vec prior the training stages. These vectors are then forward to fully-connected layer to reshape to the dimension of 64x64. The leaky-relu is applied as activation function. The resulted output is then concatenated by noise vector. The output is then concatenated with the sampled noises. The concatenated then reshaped into 4 x 4x 512 dimension. The result is then pass to four deconvolutional layers. Each of these layers is composed with 2D batch norm layer and Relu. After passing through these layers, the resulted value is forwarded to the

output layer. The returned result is in the range of $[-1, 1]$ and feed as input to discriminator to classify the reality of the generated images.

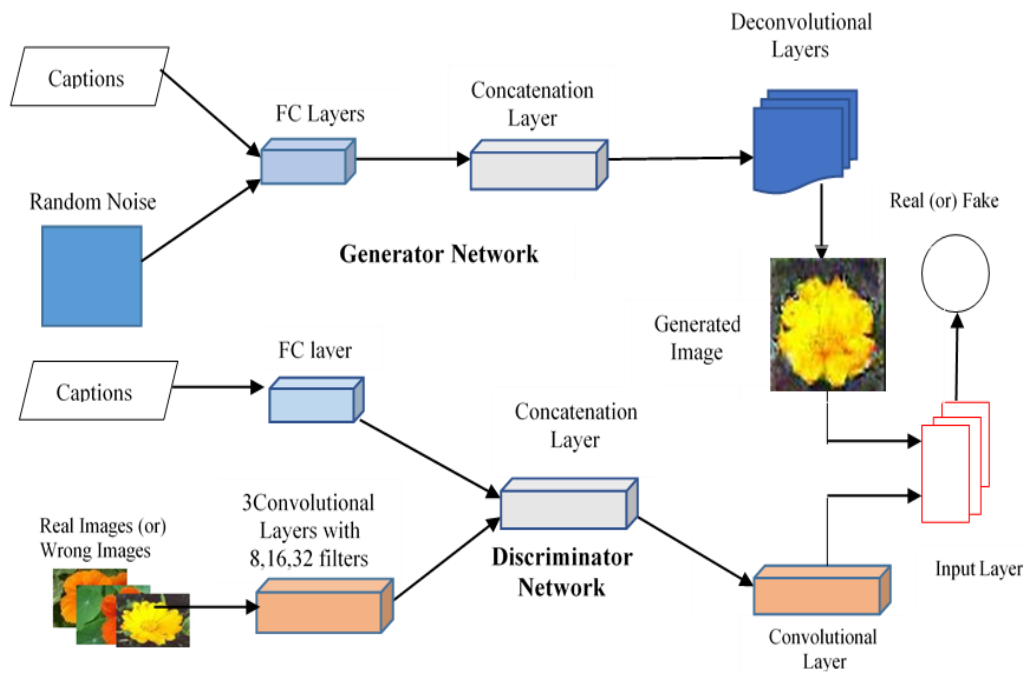


Figure 5.1. Framework of DCGAN based Myanmar T2I

The discriminator is composed with three convolutional networks layers and FC layers. The convolution function is performed by applying special normalization and LeakyRelu activation to all of its layers. There are three types of images conditioned Myanmar text on this network. These images are: (1) real images in the training dataset, (2) incorrect images created by shuffling the real one, and (3) synthesized images (the output from generator). The images are resized to 64×64 dimensions prior input to the discriminator. The resized images are concatenated with the randomly selected caption that is embedded by using Word2vec. The resulting output are then reshaped, convolved, and followed sigmoid activation function.

Finally, the discriminator evaluate the adversarial loss to classify how much the generated images look like realistic images. The weights of generator and discriminator networks are updated by using the loss evaluated by the discriminator. The generator gradually enhanced the quality of the fake images overtime by learning from its mistakes.

5.2 Implementation Details

This section described creation and preprocessing of the dataset, and the training stages of Myanmar T2I based DCGAN.

5.2.1 Dataset Creation

In order to implement Myanmar T2I models, a dataset of paired textual descriptions and corresponding images must be collected. Therefore, Myanmar caption corpus was built using Oxford-102 dataset. The captions are manually constructed descriptions for each image in the dataset by seeing their shapes, boundary and color. There are over 8000 images and 5 captions per image in this dataset. However, 1500 images were used in this training. The implementation of Myanmar T2I by using all of the images in these datasets are described in the next chapter.

5.2.2 Preprocessing the dataset

Text preprocessing is a crucial step in natural language processing as it helps to transform raw text into a more structured and meaningful format that can be easily analyzed by machine learning algorithms. Therefore, word segmentation and creating words vectors for each sentence are taken before training of GANs. There is no white space to limit boundary of words in Myanmar sentence. Therefore, sentence segmentation is done by using Myanmar syllable segmentation [11]. The segmented result are embedded by using word2vec algorithm to extract word-vectors. The extracted vectors of all sentences in Myanmar captions corpus are then saved in one hdf5 format. In order to speed up training process, this work is done before training of GANs.

5.2.3 Training Stages

Myanmar T2I is done by using DCGAN [5]. During training, the extracted word vectors of the sentence and random noises are fed are input to generator. The generator converts the images from the input sentence. The generated results from generator with captions and the real image with captions are fed to discriminator to evaluate real or fake. Based on the adversarial loss of the discriminator, the weights of these networks are updated. The ADAM solver is used for learning. The rate of learning, the batch-

size and the number of epochs are 0.0002, 64 and 600. The size of the fake output from the generator is 64x64.

5.2.4 Experimental Results

After training stages, the quantitative evaluation was done by querying descriptions to investigate the performance of the model. The query text is embedded by using word2vec model. Then, these word features are given as input to the pertained model to create fake images from this descriptions. The dimension of the images from this model is 64 x 64. In this work, the model can output a single specific flower images because the model is trained with the specific flower dataset. The generated results from this models is shown in figure [6]:

5.2.5 Evaluation

The experiments on the pretrained model is performed by querying text descriptions. The images generation process was performed by querying fifty text descriptions. The quality of the artificial images from generator is determined by applying IS and FID metrics. At this evaluation, IS score is used to assess the quantitative results of the artificially created images. The difference between real and fakes is performed by using FID. The evaluated results of these two scores are as shown in Table 5.1.

Table 5.1 FID and Inception Scores of the Generated Image from DCGAN

Scores	
Inception Score	FID
1.72 ± 0.01	222.34

Query Text	Generated Images
<p>အဖြူရောင်ပွင့်ဖတ်နှင့်ပန်းမှာအဝါရောင်ဝတ်ဆံ့ရှိတယ်</p> <p>The petal of flower is white and the stamen is yellow.</p>	
<p>ပန်းရောင်ပန်းပွင့်သည်ရေပေါ်တွင်ပွင့်နေသည်</p> <p>The flower in pink color is floating on water.</p>	
<p>ပန်းပွင့်သည်အညိုရောင်အလယ်ဗဟိုရှိပြီးအဝါရောင်ပွင့်ဖတ်ရှိတယ်</p> <p>This flower has the brown centre and yellow petals.</p>	
<p>ပန်းပွင့်မှာ အဖြူရောင်ပွင့်ဖတ်နှင့်အဝါရောင် မျိုးစေ့အိမ်ရှိတယ်</p> <p>The flower has white petal and yellow ovary.</p>	
<p>ပန်းပွင့်သည်လိမ္မော်ရောင်ဖြစ်ပြီးလိမ္မော်ရောင်ပွင့်ဖတ် ရှိတယ်</p> <p>The orange flower has the orange petal.</p>	
<p>ပန်းပွင့် ဟာ တောင်ပံပုံစံ အ သွင်အပြင် နဲ့ ခရမ်း ရောင် ရှိ ပါတယ်</p> <p>The flower with the wing-shaped is petal in color.</p>	

Figure 5.2 Images Generated from Myanmar Text Descriptions based on DCGAN

5.3 Summary

In this paper, Myanmar T2I synthesis model has been model using Generative Adversarial Networks and an annotated image dataset is created to conduct this implementation. In this research, an annotated image dataset is manually prepared by using Oxford-102 flowers dataset to implement this work. The sentences in Myanmar caption corpus are encoded using Word2vec before the training stages in order to speed up training of the models. The quantitative analysis was performed over the generated images by using inception v3 model and FID. The implementation was done by using over 1500 images in the training dataset. The training of Myanmar T2I with all of the images in Oxford-102 flowers is described in Chapter (6) and Chapter (7). Moreover, the enhancement over the quality of generated is also stated at these chapters.

CHAPTER 6

GENERATIVE ADVERSARIAL NETWORKS FOR MYANMAR T2I

Generative Adversarial Networks (GANs) have been used for various image synthesis tasks, such as generating the realistic images of people, animals, landscapes, etc. However, GAN-based text-to-image synthesis is still a challenging research area. There are also several challenges in Myanmar T2I, such as the lack of large-scale paired text and image datasets in the Myanmar language, the complexity of Myanmar script and language, and high variation in visual representations of text. New techniques and strategies are required to generate the images which can capture the nuances of the Myanmar language

Therefore, an annotated image dataset for Myanmar T2I was proposed to address the lack of image-text paired dataset in Myanmar. In addition, Myanmar language has a complex and diverse writing system, which includes various diacritic marks, ligatures, and fonts. This makes it challenging to create a model that can accurately synthesize images from text in Myanmar language. For this issue, investigation is made on two areas to highlight which techniques can accurately generate high-quality images from Myanmar input text. These techniques are (1) attention-based Myanmar T2I with multiple refinements stages and (2) deep-fusion text and images-based Myanmar T2I with only one stage of generator. The implementation and evaluation of these two techniques are described in this chapter.

6.1 Architecture of Attentional Generative Adversarial Networks

The attentional generative adversarial network [53] can generate the images conditioned on both global sentence vectors and word vectors that are relevant to each sub-region of the images. In this architecture, there are three generator networks to synthesize images from text descriptions. The noise sampled from Gaussian distribution with the global sentence is passed to the first stage of the generator to generate low resolution-images. In the following two stages, the word-context vector is computed by using attention layer which mainly focus on specific aspects of the sentence by weighting important contents. The word-context vector is dynamic representation of words that are relevant to sub-region of the image from previous generator. After computing word-context vector, the combination of the image features

and its corresponding word-context features are passed to the next generator to synthesize high-resolution images. To generate synthetic images based on sentence-level and word-level, the final objective function of attentional generative network is:

$$L = L_G + \lambda L_{DAMSM} \text{ where } L_G = \sum_{i=0}^{m-1} L_{G_i} \quad (6.1)$$

Here, L_G is the sum of all losses of the generators and each generator has a corresponding discriminator.

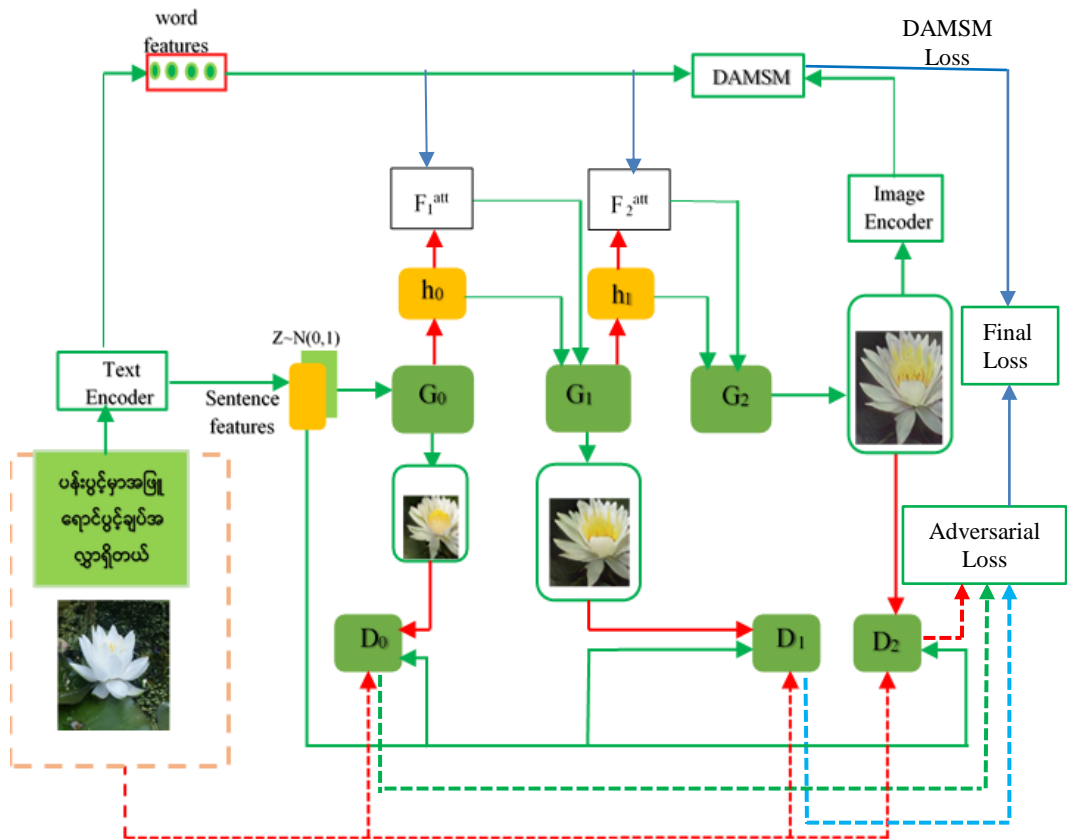


Figure 6.1 Frameworks of AttnGAN

6.2 Deep Attentional Multimodal Similarity Model (DAMSM)

This model [58] contains two neural networks: text encoder and image encoder. It is used to measure the visual-semantic similarity at the word-level to compute fine-grained loss for image generation. The text encoder is bi-LSTM (bi-directional Long Short-Term Memory) that extracts semantic vectors from the Myanmar text description. There are two hidden states for each word in bi-directional LSTM: forward hidden sequence and backward hidden sequence. Therefore, these two hidden states are

concatenated to represent the semantic meaning of words. The global sentence vector is obtained by concatenating the last hidden states of bi-LSTM.

The image encoder is composed of CNN (convolutional neural network) layer that extracts local image features and global image features. Both global and local image features passed through a perceptron layer to convert their dimensions as the same dimension of word-level and global sentence embeddings.

In this model, the matching score for image-text pair is measured based on attention model. The similarity scores between the images and Myanmar text descriptions are computed based on word-level and sentence-level. The cosine similarity is used to measure the visual-semantic similarity between each word in the sentence and sub-regions of the image. Moreover, the cosine similarity score for how well image Q matches with description D on sentence-level is computed by using with the following equations:

$$R(Q, D) = (\bar{v}^T \bar{e}) / (||\bar{v}|| ||\bar{e}||) \quad (6.2)$$

And then, the similarity scores of image Q matches with description D based on word-level and sentence-level are summed up to get the final DAMSM loss. This DAMSM loss is added with the loss of generators to get the final loss of generator.

6.3 Architecture of Deep Fusion GANs

In this model [32], there is only one stage of discriminator and generator. The text descriptions and a noise vector sampled from the Gaussian distribution are input to the generator. The text descriptions are encoded by a pretrained encoder similar to Attention GAN. First, the noise vector is fed to Fully-connected layer. The output is passed to a series of UpBlocks. The UpBlock contains three layers: (1) a residual block (2) upsample layer and (3) Deep Fusion blocks that is used to concatenate image features and text during the image generation stage. The image features are obtained by conditioning the sentence vector at each block. Finally, the resulted image features are passed through convolutional layer generate high-quality images.

The generated or synthesized images are passed to discriminator network. The discriminator extracts the image features and the output is downsampled at each of its layer. Then the sentence vector is concatenated with the image features. And the adversarial loss is calculated to evaluate the visual-semantic consistency. By using this

loss, the discriminator promotes the generator to output high resolution images that relevant to the natural language text description.

In text-to-image generation process, the model should be able to synthesize more realistic images with better semantic consistency. Therefore, it is very important to construct the proper loss function of the discriminator. In DF-GAN, the gradient penalty is applied to ensure the real images and captions are at the minimum points of the loss function and the vicinity of that data points is smooth in order to help the coverage of the generator to reach the minimum point easily. The whole formulation of loss function generator is:

$$L_G = -\mathbb{E}_{G(z) \sim P_g} [D(G(z), e)] \quad (6.3)$$

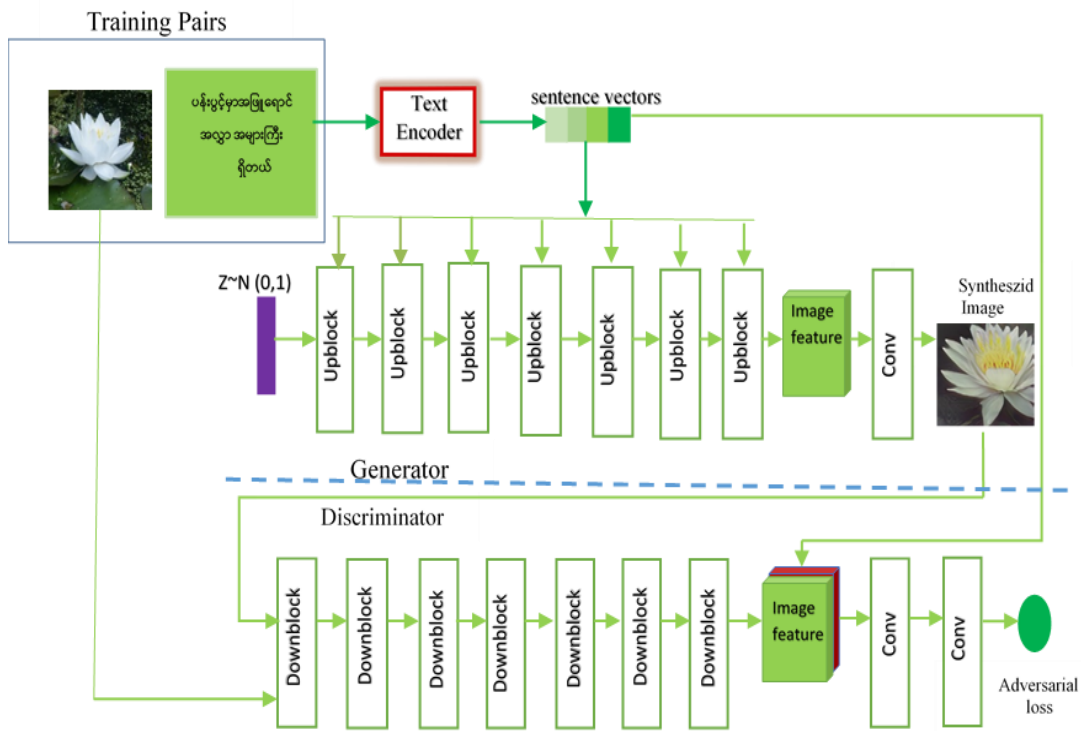


Figure 6.2 Frameworks of DF-GAN

6.4 Training of Myanmar Text to Image Synthesis

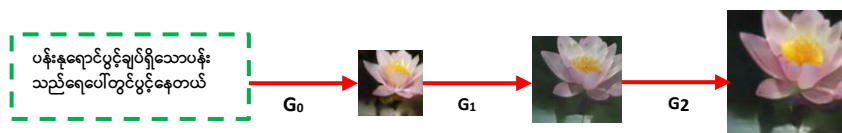
This section contains implementation of Myanmar text to image synthesis on two models. We pretrained DAMSM model that contains text encoder and image encoder. We embed Myanmar sentence by using bi-LSTM text encoder. We extract word features by concatenating two hidden states of bi-LSTM to represent the semantic meaning of each word in Myanmar sentence. The global sentence feature is obtained

by concatenating the last hidden states of bi-LSTM. This model is used to compute text and image similarity level during the training stages of AttnGAN.

In AttnGAN, sentence vectors are concatenated with the noise samples from Gaussian Distributions and fed to the first generator. In this stage, the images are synthesized with dimension of 64x64 based on sentence features. At the remaining two stages, the images are synthesized based on fine-grained word-levels. The dimension of synthesized images is 128x128 and 256x256 respectively.

In DFGAN, the sentence features are obtained by using pretrained text encoder in AttnGAN. This model generates the image with 256x256 resolution with only one stage backbone. Instead of implementing the extra network like DAMSM model to calculate similarity between text descriptions and images, the gradient penalty is applied to ensure the real image and text descriptions matching score are at the minimum point of loss function. The image generation process of these two models is shown in figure [5]. These two models were trained at maximum of 1000 epochs. But, the training results of AttnGAN become overfitting and degradation in the quality of images at over 600 epochs. Therefore, these two models were compared by using the best epochs of each model instead of using the same epoch. Table 6.1 shows the training parameters of these two models:

(a)



(b)

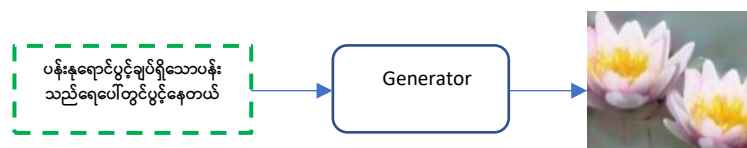


Figure 6.3 The Image Generation Process of (a) AttnGAN (b) DF-GAN

Table 6.1 Training Parameters of AttnGAN and DF-GAN

Training Parameters	AttnGAN	DF-GAN
Batch size	16	24
Epochs	600	1000
Loss function	Cross-Entropy loss	Hinge loss
Optimizer	Adam	Adam
Nos of generators	3	1
Generator Learning Rate	0.0002	0.0001
Generator Learning Rate	0.0002	0.0004
Noise dimension	100	100

6.5 Experimental Results and Comparison

This experiment used two evaluation metrics to evaluate quantitative results for generated images from Myanmar text descriptions. It also computed inception scores and FID score of the generated images based on testing dataset. The comparisons score of these two GANs are shown in Table 2. In quantitative evaluation, DFGAN got higher scores on inception and lower FID scores than AttnGAN. Therefore, DFGAN enables to generate better quality images than AttnGAN for Myanmar text to image synthesis in quantitative evaluation. For this point, fusing text-image at every block is more good impact on our corpus than focusing on specific words at every stage of GAN. Moreover, comparison on qualitative evaluation was also made based on human perception.

Table 6.2. Inception score and FID score of two models evaluate based on Test data

Model	Inception score	FID score
AttnGAN	3.5 ± 0.02	40.07
DFGAN	3.6 ± 0.03	39.68





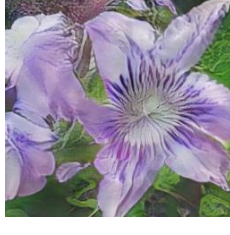

Text Descriptions	AttnGAN	DF-GAN
<p>ပန်းပွင့်တွင်အဖြူရောင်ပွင့်ချပ်နှင့်အဝါ ရောင်ဝတ်ဆံရှိတယ် In English: The flower has white petals and yellow stamens</p>		
<p>ပန်းပွင့်တွင်အဖြူရောင်ပွင့်ချပ်နှင့်ပန်း ရောင်ဝတ်ဆံရှိတယ် In English: The flower has white petals and pink stamens.</p>		
<p>ပန်းပွင့်တွင်အဖြူရောင်ပွင့်ချပ်နှင့်ခရမ်း ရောင်ဝတ်ဆံရှိတယ် In English: The flower has white petals and purple stamens.</p>		

Figure 6.4. The images generated from Myanmar descriptions while changing some words

Qualitative evaluation is made by querying text descriptions to synthesize image. As shown in Figure [5], the comparison of the images generated by two methods with each text descriptions has been made and the best sample is selected. In this evaluation, the generated images from DFGAN are sharper and clearer than the image generated from AttnGAN. And also, DFGAN can generate the images more precise in shape and brightness in color than AttnGAN. Moreover, artificially generated images from DFGAN are more realistic than those images from AttnGAN. DF-GAN enables to generate the images which features are more relevant to text descriptions than AttnGAN. Figure [6] shows the images generated from two models while changing some words in Myanmar text descriptions. In this testing, the synthesized images from DFGAN also got higher quality than those images from AttnGAN. In qualitative evaluation, DFGAN is also better than AttnGAN. DFGAN is better than AttnGAN in both quantitative and qualitative evaluation. DFGAN outperforms AttnGAN for

Myanmar text to image synthesis. Therefore, fusing text and image at every blocks highlights more good impact than attention with multiple refinements for Myanmar T2I











Text Descriptions	AttnGAN	DF-GAN
<p>ပန်းပွင့်တွင်အဖြူရောင်အဆင်းရှိသောပွင့်ချပ်များနှင့်အဝါရောင်စင်တာရှိတယ်</p> <p>In English: The flower has white color petals and yellow center.</p>		
<p>ပန်းပွင့်တွင်ရှည်လျားသောမက်မွန်ရောင်ပွင့်ချပ်များရှိတယ်</p> <p>In English: The flower has the elongated peach petals.</p>		
<p>ခရမ်းရောင်ပွင့်ချပ်နှင့်ပန်းပွင့်တွင်အနီရောင်အမှတ်အသားများရှိတယ်</p> <p>In English: The flower with purple petals has the red shade.</p>		
<p>ပန်းပွင့်ရဲ့အလယ်မှာရှိတဲ့အစိမ်းရောင်ဝတ်ဆံကိုဝန်းရံနေတဲ့ခရမ်းရောင်ပွင့်ချပ်တွေရှိတယ်</p> <p>In English: The green stamens in the center of flower are surrounded by the purple petals.</p>		
<p>ပန်းတွင်အနီတစ်ဝက်အဝါရောင်တစ်ဝက်ရှိသောပွင့်ချပ်ရှိတယ်</p> <p>In English: The half of petals of flower is red and the remaining part is yellow</p>		

Figure 6.5. The images generated from Myanmar Text descriptions

6.6 Summary

This chapter described building an annotated Myanmar captions for each image in Oxford-102 flowers dataset and modeling the first Myanmar text to image synthesis. High resolution images are generated from Myanmar text description by using AttnGAN and DF-GAN. We obtained the images with dimension of 256 x256 using multiple-stages of generators in AttnGAN and one-stage of generator in DF-GAN. In AttnGAN, visual-semantic similarity is evaluated using DAMSM in training stage. In DF-GAN, Gradient-Penalty Matching Aware is applied to Discriminator to generate image with matching description so that DF-GAN does not need extra network like DAMSM in AttnGAN. A comparative study has done in two models in both qualitative and quantitative evaluation. According to experimental results, DF-GAN is better than AttnGAN for Myanmar to image synthesis.

CHAPTER 7

MULTIMODAL GENERATIVE MODEL BASED T2I

While text-to-image synthesis technology has made the significant progress in recent years, there are still many challenges to overcome, such as generating high-quality and diverse images, as well as accurately capturing the nuances of complex textual descriptions. In the case of Myanmar T2I, there are several challenges in developing Myanmar T2I models, including the scarcity of large-scale datasets, the need for specialized models that can handle the unique features of the Myanmar language, and the requirement for high-quality images that accurately represent the input text.

For this issue, the comparison and analysis are made between Attn-GAN and DF-GAN to depict which model provides more effective results for Myanmar T2I. This work is described in chapter [6]. In this work, DF-GAN is better than Attn-GAN in generating images for Myanmar text. In this comparison, DF-GAN can generate high-resolution images (256x256) with only one stage of generator though Attn-GAN requires multiple generators to acquire the images with this resolution. In addition, DF-GAN does not require the extra pretrained model to evaluate similarity between text and image. However, there are some issues on visualization of Myanmar text inaccuracy in color, imprecision of its shape and boundary. Accordingly, pretrained MSM is applied to DF-GAN to more semantically understand Myanmar textual descriptions and more correctly visualize the images from these descriptions. Thus, we named DFGAN with our MSM as DFGAN+MSM. With DFGAN+MSM, the generated image is more accuracy in shape and color, and semantically consistency. Moreover, the proposed model is applied to another type of dataset with English text description to investigate the improvement of the image generation process to applying MSM to the generators. The detailed implementations for training DF-GAN+MSM is described in this chapter.

7.1 Architecture of Multimodal Similarity Model

Multimodal similarity model is a multimodal and consists of two sub-encoders (i) text encoder and (ii) image encoder. This model uses image encoder to extract visual features and text encoder to get text features. MSM is used to evaluate the similarity

loss between the generated images and text similarity loss during training of Myanmar T2I. The architecture of MSM is shown in Fig. 7.1.

The text encoder is built with bidirectional long short-term memory (bi-LSTM) [31]. It is used to extract sentence features from Myanmar sentence. There are two hidden states in bi-LSTM: forward layer and backward layer. The last hidden states are obtained by concatenating the hidden states of these two layers. These last hidden states are concatenated to obtain the whole sentence features of Myanmar sentence.

The image encoder is Convolutional Neural Network and built upon Inception-v3 model [11] pretrained on ImageNet. This encoder maps the image to image features. Image preprocessing is a critical part of the system and can impact the maximum accuracy during training of the model. At a minimum, images need to be decoded and resized to fit the model. Therefore, we resize the images into 299x299x3 pixels. After resizing, the image features are extracted by using “mixed-6e” layers of inception-v3 model. Finally, we get the global image feature by using average pooling layer of Inception-v3 model.

The visual features and the language features are then map into a semantic space with the same dimension. Finally, the similarity scores are calculated by using cosine similarity. The equation for calculation similarity scores between the image (I) and text description (T) is defined as follow:

$$C(I, T) = I * T / \| I \| * \| T \| \quad (7.1)$$

The objective of similarity learning is to further improve visual-semantic correlation and the quality of generated images during training of Myanmar T2I. Therefore, we compute the visual-semantic consistency loss upon two areas: the similarity loss upon (i) images given text descriptions and (ii) text descriptions given images. These two losses are computed by the following equations:

$$Loss_1 = \sum_{k=1}^N \log P(T_k | I_k) \quad (7.2)$$

$$Loss_2 = \sum_{k=1}^N \log P(I_k | T_k) \quad (7.3)$$

Finally, we sum these two losses to evaluate the final loss. Therefore, the loss of MSM is defined as

$$L_{MSM} = Loss_1 + Loss_2 \quad (7.4)$$

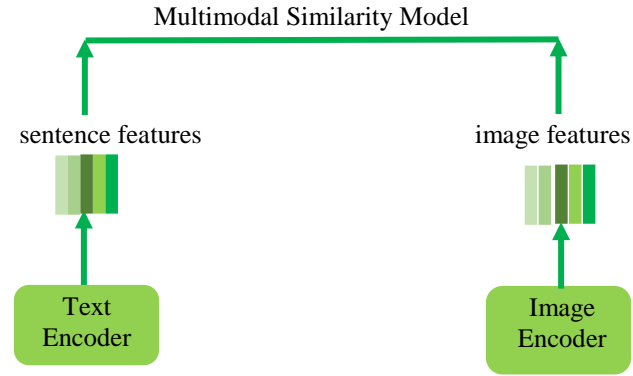


Figure 7.1 Architecture of MSM

7.2 Architecture of DF-GAN+MSM

This model is used to generate images conditioned on Myanmar textual descriptions. There are three processes in this network: (1) image generation, (2) evaluation of the similarity loss using MSM, and (3) calculation of the final loss for the generator. The architecture of this model is shown in Fig. 2.

In the image generation stage, the text encoder converts the textual input into sentence features. Random noise sampled from a Gaussian distribution is passed through a fully connected layer. The output from this layer is passed through a series of upblock layers of the generator. The sampled noise is concatenated with the textual input while forwarding this noise to these layers. Then, the generator transforms the resulting image features into a high-resolution image (256x256).

In the second stage, the generated image and the real image are passed to the downblock layers of the discriminator by concatenating with the replicated sentence features. Finally, the discriminator computes the adversarial loss to identify the data as realistic and semantically consistent with the inputs (generated image with caption and real image with caption in the dataset). At the same time, the discriminator is also learning to become more accurate in distinguishing between real and fake samples.

In the last stage, the generated images and the encoded textual input are fed as inputs to the MSM model to compute the similarity loss between these two inputs. Then, the final loss of the generator is obtained by summing the adversarial loss and the similarity loss, and this feedback is provided to the generator. This feedback allows the generator to learn from its mistakes and gradually produce better and more realistic samples with semantic consistency.

Therefore, the loss function of the generator is as follows:

$$L_G = -\mathbb{E}_{G(z) \sim P_g} [D(G(z), e)] + L_{MSM} \quad (7.5)$$

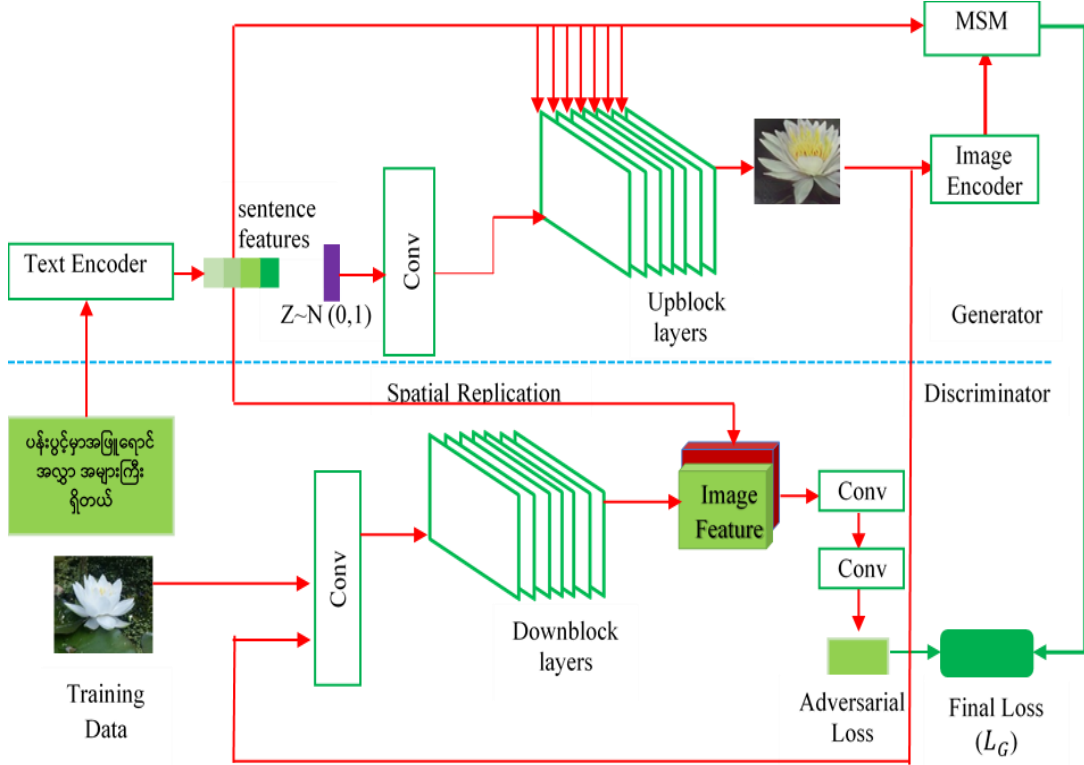


Figure 7.2. Framework of DF-GAN+MSM

7.3 Implementation Details

This section contains the processing steps for Myanmar T2I. The system flow diagram for implementing the system is shown in Fig.7.3.

7.3.1 Dataset

To conduct the first implementation of the proposed model, we used CUB birds dataset [17]. This dataset contains 11,788 images and 200 different kind of birds. Each image is paired with 10 text descriptions.

To implement Myanmar text to image synthesis, we conducted our experiments on Oxford-102 flowers dataset [28]. The dataset consists of 102 categories, 8189 images. Firstly, the English captions corpus are translated to Myanmar captions by using machine translation. In this approach, the quality of translation is not accurate to use in

text-to-image synthesis because training dataset is not relevant with this caption corpus. Therefore, we manually constructed Myanmar captions corpus for each image by focusing their features without directly using or translating English descriptions from the original dataset. There are 5 annotated captions for each image in this dataset. The total number of sentences in Myanmar caption corpus is 40945.

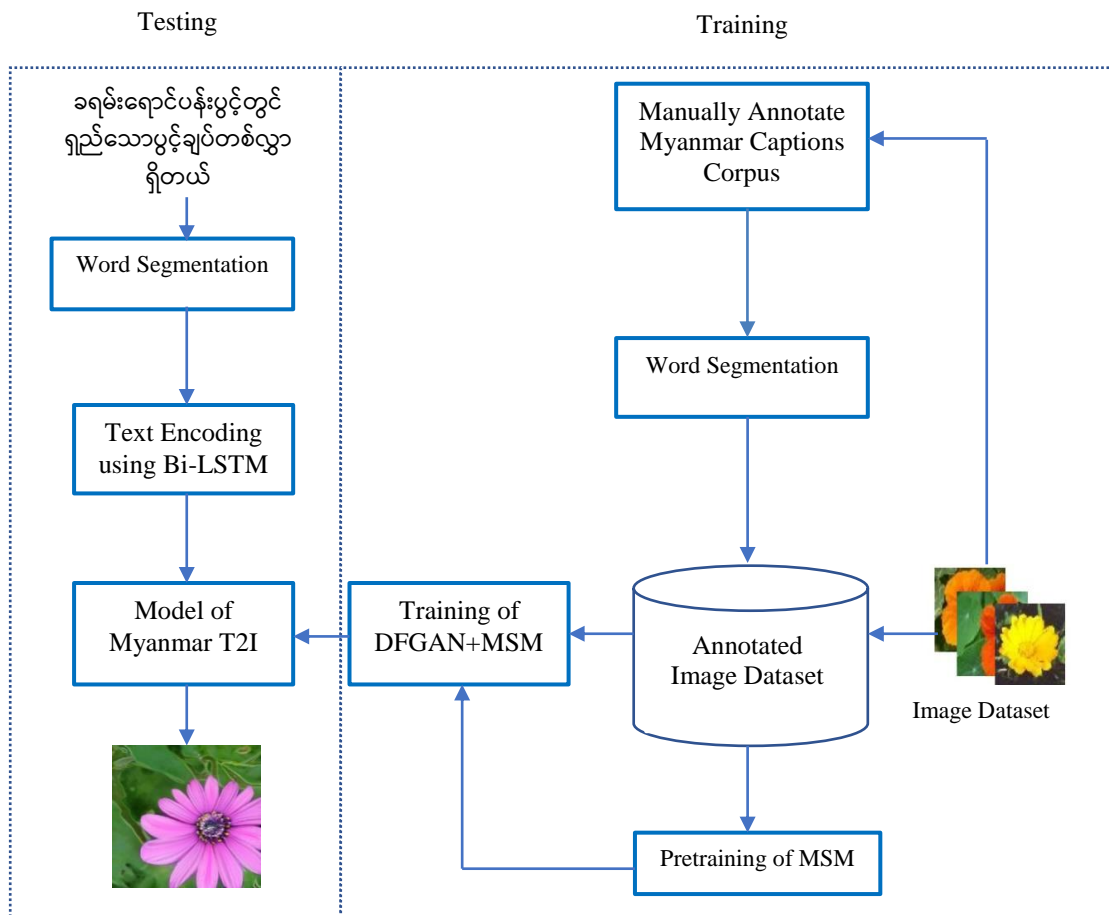


Figure 7.3. Flowchart of Myanmar T2I

7.3.2 Preprocessing Dataset

Myanmar sentences have been built with sequence of characters and do not contain white space to delimit word boundary like English. Therefore, sentence segmentation is an essential preprocessing step prior to language processing in Myanmar language. After segmentation, we got 1645 words in Myanmar caption corpus, and the maximum length of sentence is 25. After word segmentation, text encoding process is performed to convert meaningful descriptions into number representation. Myanmar sentences are encoded into numbers by using index-based encoding methods. This encoded result is

used in training of text encoder. The following is the segmented result of an annotated caption in our corpus.

7.3.3 Pretraining Multimodal Similarity Model

After pre-processing, MSM model was pretrained using real images-text pairs in dataset. In this training, the text encoder and image encoder are jointly trained by minimizing the visual-semantic consistency loss. The reasons for pretraining MSM is to increase the speed of training the other components. This pretrained model is added to the generator to evaluate the similarity loss of the generated image from generator and Myanmar text descriptions. Moreover, the text encoder is also used to extract the feature vectors of Myanmar text descriptions. The pretrained model with the minimum loss is used as MSM during training of T2I. The parameter settings on this model is shown in Table 1:

Table 7.1. Hyperparameters setting of MSM

Parameters	Values
Batch size	16
Number of epochs	300
Text embedding size	256
Encoder learning rate	0.002
Maximum text sequence length	25

7.3.4 Training DF-GAN+MSM

Myanmar text to image synthesis is implemented after training of MSM. In this training, 7789 images are used as training and 400 images are used as testing data. This testing data is used to evaluate and analyse the performance of the model. We used the pretrained text encoder in MSM as encoder to extract the feature vectors for Myanmar text descriptions. The image encoder and text encoder in our pretrained MSM are used to extract the feature vectors of the generated images from generator and text descriptions. The training system was stopped at the maximum of 1000 epochs because the system can generate the images with precise in shape and semantic consistency at

this epoch. We used Adam solver for learning and hinge loss to stabilize the training process. In this implementation, the dimension of the generated image is 256 x 256. The parameters used in this training are shown in Table 2.

Table 7.2. Hyperparameters setting of DF-GAN+MSM

Parameters	Values
Batch size	24
Number of epochs	1000
Text embedding size	256
Maximum text sequence length	25
Generator Learning Rate	0.0001
Discriminator Learning Rate	0.0004
Noise dimension	100

7.4 Experimental Results and Analysis on Myanmar T2I

Quantitative and qualitative evaluations are done to identify the quality of the generated images conditioned on Myanmar text descriptions. The quantitative results of DFGAN+MSM (proposed model) are compared to previous results of baseline DF-GAN and AttnGAN based Myanmar T2I published in [37] and DCGAN based Myanmar T2I published in [37]. Moreover, the results of qualitative evaluation are also compared with baseline DF-GAN. We trained all of these systems with the same training data and testing data.

7.4.1 Quantitative Analysis

This sections compared and analysed the inception score and FID score of the generated images from Myanmar sentences based on the proposed model with these cores of baseline DF-GAN and AttnGAN [53]. The quantitative analysis is investigated on testing data (total of 400 images) for our proposed model, DF-GAN and AttnGAN.

The comparative results of FID scores and inception scores for the proposed model, the baseline DF-GAN and AttnGAN are shown in Fig. 4 and Fig. 5. In this comparison, the result of AttnGAN is only compared up to 800 epochs because the model got lower inception scores and higher FID scores due to degradation in the quality of generated images over 600 epochs. Therefore, the training of Myanmar T2I based AttnGAN was stopped at 800 epochs. According to the results shown in these two figures, our model got lowest FID scores and highest inception scores at every epoch. Based on this result, our model outperforms other works. In addition, we made performance comparison of these two scores of our model to other models trained with the same dataset are also shown in Table 3. Our model got the highest inception scores and the lowest FID scores at 1000 epoch. Therefore, we used these two scores of this epoch in this comparison. According to the results, our model got the highest inception scores and smallest FID scores. Our model got the highest quantitative results compared with other models. Therefore, the proposed model can generate the best quality images from Myanmar sentence.

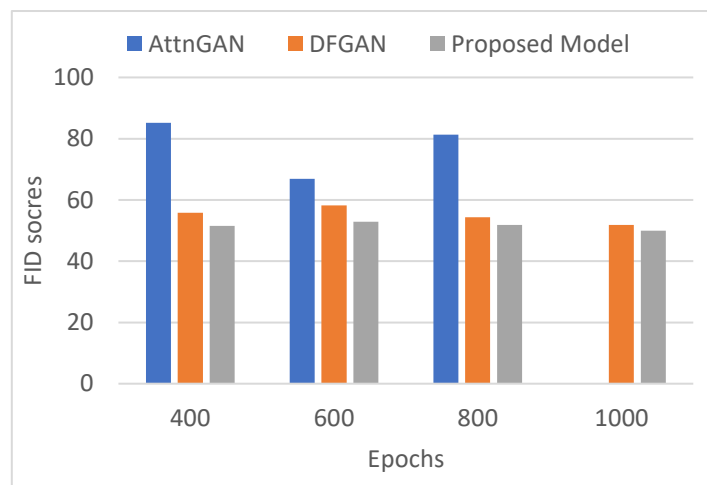


Figure 7.4. Comparison of FID Scores of AttnGAN, DF-GAN and Our Proposed Model

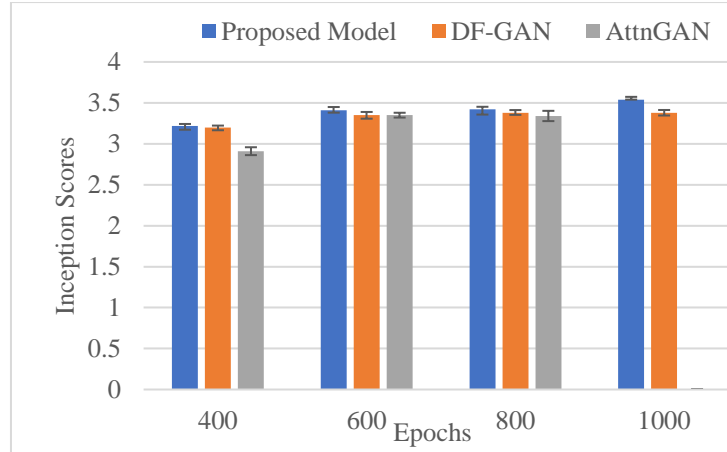


Figure 7.5. Comparison of IS Scores of AttnGAN, DF-GAN and Our Proposed Model

Table 7.3: Comparison of FID score and Inception Score

Model	Inception Score \uparrow	FID score \downarrow
DCGAN [36]	1.72 ± 0.01	222.34
AttnGAN [37]	3.35 ± 0.03	66.92
DF-GAN [37]	3.38 ± 0.03	51.86
Proposed Model	3.54 ± 0.03	49.97

7.4.2 Qualitative Analysis

The qualitative evaluation is done in two ways:(1) classifying quality of the generated images from querying text descriptions and (2) assessing the visual quality of the generated images based on human perception. The classification of image quality was performed depending on shape, colour accuracy and visual-semantic consistency. The comparison of the generated images per text descriptions with base-line DF-GAN and the proposed method are shown in Fig. 6.

As it is noted in this figure, the generated images from the proposed model is more precise in shape and brightness colour than those images from other models. In figure, the images generated from the third input text of DFGAN and AttnGAN failed in colour accuracy like the shape of flower does not illustrate with half and half of red and white colour. DF-GAN and AttnGAN cannot generate the image with semantic accuracy for the fourth input sentence because there is no black line over the generated flower like

the one from DFGAN+MSM. In the fifth input text, the proposed model can illustrate the whole features of flower but the other models cannot figure completely. Moreover, the generated images from DFGAN and AttnGAN are dull in colour and contains some noises (imprecision of shape and boundary) in the portion of petals. However, the image generated from our proposed model got more colour accuracy than the baseline model overall the generated images. Moreover, the proposed model can create the images with text-image consistent and brightness in colour than the other models for overall the generated images. Therefore, the proposed model outperforms the other models for Myanmar T2I.

The overall subjective evaluation of the generated images was also conducted to assess the quality of the generated images between baseline DF-GAN and DFGAN+MSM because DF-GAN is better than AttnGAN and DCGAN according to the results shown in Table 7.3. The number of 25 generated images from Myanmar text descriptions are used in this evaluation. The 25 non-expert native persons of age range from 20 to 45 years were rated the quality of generated images based on their visual appeal, realism, and overall quality. The scales to rate the quality of these images are 1(poor), 2(fair), 3(good), 4(very good) and 5(excellent).

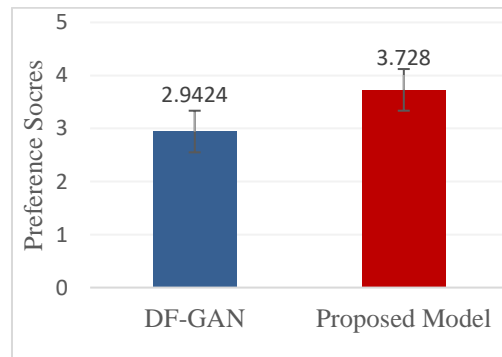


Figure 7.6 The Comparison of the Preference Scores

The evaluated preference scores of DF-GAN and the proposed model are shown in Fig. 7.6. According to these results, the scores of the proposed model is higher than the baseline DF-GAN. It can be observed that MSM with DF-GAN gives improvements for MyanmarT2I in terms of image quality and visual-semantic consistency. The proposed model got higher performance than DF-GAN in both quantitative and qualitative analysis. The proposed model is better than DF-GAN for Myanmar T2I. Therefore, MSM guides DF-GAN to generate high quality image with better semantic consistency for Myanmar T2I.

Input Text Descriptions	Generated Images with AttnGAN	Generated Images with DF-GAN	Generated Images with DF-GAN+MSM
<p>In Myanmar: ပန်းပွင့်တွင် ရှည်လျားသောမက်မွန်ရောင်ပွင့်ချပ်များရှိတယ်</p> <p>In English: The flower has the elongated peach petals.</p>			
<p>In Myanmar: ပန်းနုရောင် ပွင့်ချပ် ရှိသော ပန်းသည် ရေပေါ်တွင်ပွင့်နေတယ်</p> <p>In English: The flower with the light pink petals is blooming on the water.</p>			
<p>In Myanmar: ပန်းပေါ်တွင်အနီရောင်နှင့်အဖြူရောင် တစ်ဝက်စီရှိသောပွင့်ချပ်ရှိတယ်</p> <p>In English: The half of petals of flower is red and the remaining part is white.</p>			
<p>In Myanmar: အနက်ရောင် လိုင်းပါသော ခရမ်းရောင်ပန်းပွင့်</p> <p>In English: The purple flower with the black line</p>			
<p>In Myanmar: ပန်းတွင် ပန်းရောင်တောက်တောက် ပွင့်ချပ်များနှင့်အညိုရောင်ဝတ်ဆံဖိုများရှိတယ်</p> <p>In English: This flower has the bright petals and the brown stamens.</p>			

Figure 7.7 The Generated Images from Myanmar Textual Inputs

7.5 Experiment Results and Analysis on English T2I

This section contains the quantitative analysis and quality data analysis the images generated from the proposed model implemented on CUB birds dataset annotated in English.

7.5.1 Quantitative Analysis

The researcher compare and analyses the quantitative results of the proposed models with state-of-the-arts models. The comparative results of FID score and Inception score are shown in Table 2. In this comparison, the baseline DF-GAN got the largest score compared with the state-of-the-art models. According to comparative results, FID score of the baseline DF-GAN greater than TIME. However, the proposed model got smallest FID score and highest IS score in this analysis. Compared with DF-GAN, our proposed model decreases the FID score from 14.81 to 13.05 and has a smaller significance increase on IS score. Therefore, it is obvious that our proposed model outperforms the baseline DF-GAN for T2I conducted on CUB birds dataset.

Table 7.4. Comparison of FID and Inception Scores of the Proposed Model with others on CUB dataset

Models	IS \uparrow	FID \downarrow
StackGAN [18]	3.70	-
StackGAN++ [29]	3.84	-
AttnGAN [53]	4.36	23.98
DM-GAN [34]	4.75	16.09
DAE-GAN [45]	4.42	15.19
TIME [9]	4.91	14.30
DF-GAN [32]	5.10	14.81
DF-GAN+MSM(Our)	5.12	13.05

7.5.2 Qualitative Analysis

The researcher also compared the visual aspects of the synthesized image from the baseline DF-GAN and the proposed model. The output results of these two models conditioned on English text descriptions are shown in Figure 7.7. As we noted in this result, the synthesized image by the proposed model is more different texture and colour than those image of the baseline model from the first input text. In the second and third input text, the synthesized images of the proposed model is more accuracy in shape and

better realistic than the baseline model. The image generated from the fourth input sentence by the proposed model is more semantically correct than the baseline DFGAN. Because the visualization from our proposed model can create more exactly than the baseline model for the words “a black layering”. As it is noted in this result, the images generated by the proposed model achieves better performance than the baseline model in terms of reality and semantically consistence.

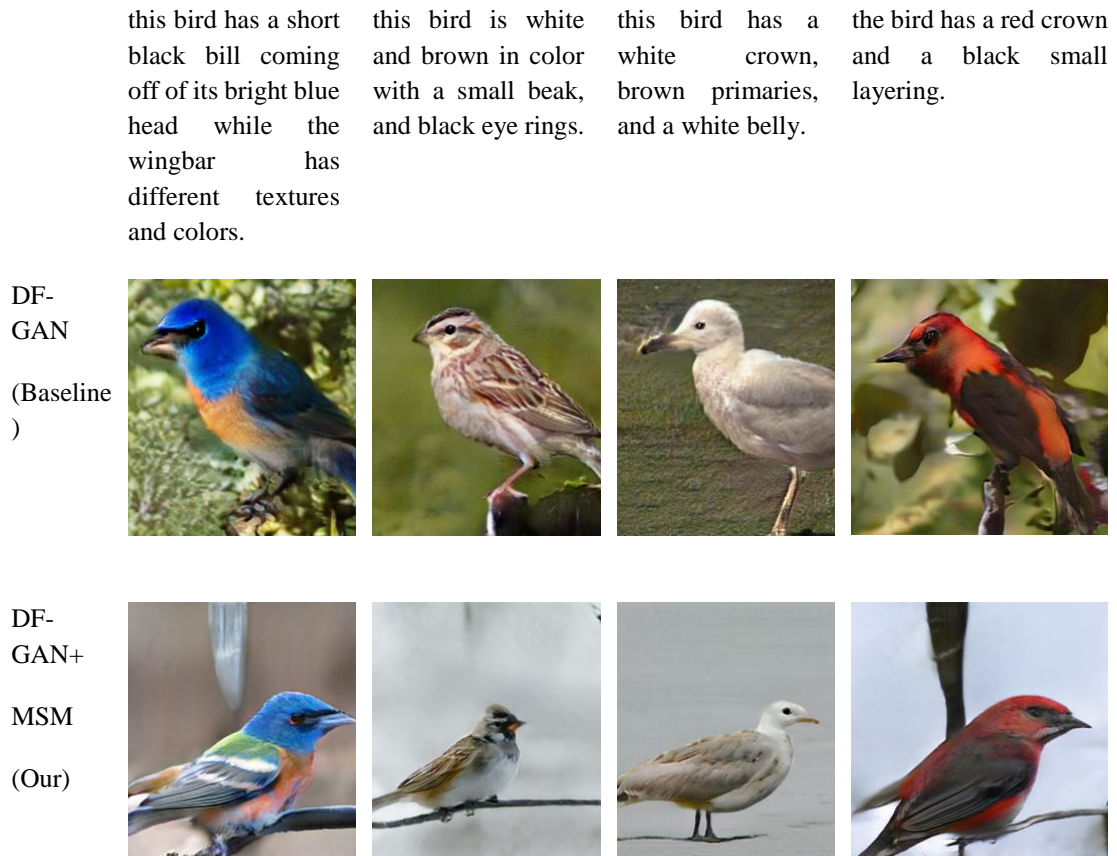


Figure 7.8. The Bird Images Generated from English Text Description

7.6 Summary

In this work, the improvement results in image generation process were examined due to applying multimodal similarity model to the generator side using two different datasets annotated on different languages. Moreover, the researcher manually prepared Myanmar captions for each image in Oxford 102 flowers dataset to implement the first Myanmar T2I. Before training of DFGAN+MSM, MSM model is pretrained using real image text pairs in the dataset for faster training of T2I. The effectiveness of the proposed model is investigated by comparing with state-of-the-arts models. This investigation is performed on both quantitative and qualitative evaluation. In quantitative evaluation, the proposed model obtained higher

inception score and smaller FID score than baseline model. In addition, the proposed model got higher scores on human perception evaluation than the baseline DF-GAN on Myanmar T2I. Therefore, multimodal generative models give better effectiveness than generative models without multimodal.

CHAPTER 8

CONCLUSION AND FUTURE WORKS

This paper reports the results of research on DF-GAN+MSM model based Myanmar text-to-image synthesis. All contributions in multimodal and generative models meet the objectives of this dissertation described in chapter (1).

The main contribution of this research is building the first GANs based Myanmar text-to-image synthesis. To train Myanmar T2I model, a large dataset of Myanmar text descriptions is manually annotated for their corresponding images based on Oxford-102 flowers. The annotation and preprocessing of text descriptions and images in this dataset is described in chapter (4).

Firstly, Myanmar T2I is proposed using Deep Convolutional GANs (DCGANs) with word2vec. This work is implemented to investigate the quality and reliability of this dataset is effective to implement this research. The implementation of DCGANs with this dataset is described chapter (5).

One of the challenges in generating images from Myanmar text is lack of large-scale annotated image datasets. Additionally, the Myanmar script is complex and has many character. Therefore, it may be difficult to generate realistic images that accurately reflect the meaning of the given text description. Therefore, the researcher has investigated on both attention mechanisms with multiple refinements states and deep fusion with one stage of generative to depict which method gives better impact on Myanmar T2I. In this comparison, deep fusion method is better than attention mechanisms for Myanmar T2I. The implementation and evaluation of Myanmar T2I about these two models is described in chapter (6). The first DCGAN-based Myanmar T2I can only generate low resolution images with dimension of 64x 64. However, these two models can generate high-resolution images with 256x256.

In the second work, deep fusion of text with image at every upblocks of generative gives more effectiveness for Myanmar T2I. Despite the progress has achieved by DFGAN, there are some artifacts that need to enhance on the quality of synthesized images such as color accuracy, shape, etc. For reason, DF-GAN with multimodal similarity model is proposed in order to advance the quality of generated images. MSM model is applied to the generator in order to evaluate the similarity loss of visual-semantic during training stage. This loss is also summed to adversarial of the

discriminator to obtain the final loss of generator. The training stages of DF-GAN+MSM is described in chapter (7).

The quantitative results of models-based Myanmar T2I is compared and analyzed using Inception score and FID scores. In this evaluation, DF-GAN+MSM got the highest inception score and the smallest FID score than other models. Moreover, the qualitative evaluation is performed by classifying the quality of these images based on human perception. In this evaluation, DFGAN+MSM outperforms baseline model DF-GAN. Therefore, DF-GAN with MSM gives better performance than other models for Myanmar T2I. The comparative results of qualitative and quantitative is expressed in chapter (7).

8.1 Advantages and Limitations of the Proposed System

Text-to-image synthesis can help artists, designers, and content creators to generate new and unique images based on their textual descriptions, allowing them to be more creative and efficient in their work. This is especially relevant for small businesses, startups, or organizations with limited budgets. It can also save time by generating images quickly and automatically, allowing content creators to focus on other aspects of their work.

In addition, text-to-image synthesis can be an important tool to help people with visual impairments to better understand and visualize information that they would otherwise be unable to see. Moreover, it can help to localize content by generating images that are specific to Myanmar culture and language.

The annotated dataset-based Oxford 102 flowers in this research can also be used in other multimodal learning such as image captioning, and other text-to-image synthesis tasks. The multimodal similarity model is also useful for similarity analysis in other text-to-image synthesis.

As a limitation, GANs are trained on a specific dataset and designed to create the new realistic images that are similar to the ones in training dataset. For this reason, this system can only generate flower images because the annotated images dataset used in this research only contains flowers images. While GANs have the potential to generate diverse images for the same input text, achieving consistent diversity can be challenging due to factors such as model capacity, training data, training procedure, and hyperparameter settings. Addressing these challenges and improving diversity in text-

to-image generation remains an active area of research in the field of GANs. As a result, the generation of different images with the same text is a challenging task in this research. In addition, Myanmar captions corpus are manually constructed using font that are based on Unicode system. Therefore, the users need to query by using this type of font while generation of images from Myanmar text descriptions.

8.2 Future Works

In this work, an annotated caption image dataset has been constructed based on Oxford-102 flowers and implemented Myanmar T2I. In this work, dimension of the generated images is 256x256. In the future, Myanmar T2I will be extended to enable to generate the images with higher resolution than the current result. Another kind of annotated image dataset will be built and trained with this current method.

In addition, more comparative analysis will be performed by modelling this system with the current trending technology such as Variational Auto Encoder (VAE), diffusion models. Furthermore, the similarity score between the generated images and text description will also be evaluated by using the large pretrained multimodal such as Contrastive Language-Image Pre-Training (CLIP), Contrastive Captioner (CoCa) developed by Google to improve the quality of realistic images with better semantic consistency.

AUTHOR'S PUBLICATIONS

- [P1] N. K. Htwe, W. P. Pa, “Building Annotated Image Dataset for Myanmar Text to Image Synthesis”, Proceeding of the IEEE 19th International Conference on Computer Applications (IEEE-ICCA 2021), pp. 164-169, Yangon, Myanmar.
- [P2] N. K. Htwe, and W. P. Pa, “Generative Adversarial Networks for Myanmar Text to Image Synthesis”, In: Proc. of International Conf. on Communication and Computer Research, Seoul, Korea, 2022.
- [P3] N. K. Htwe, and W. P. Pa, “Multimodal Generative Models based Text to Image Synthesis”, International Journal of Intelligent Engineering and Systems (**IJIES**), Japan, 2023. ISSN: 2185-3118 (**Scimago index**)

BIBLIOGRAPHY

- [1] A. Brock, J. Donahue, K. Simonyan, “Large scale GAN training for high fidelity natural image synthesis”, arXiv preprint arXiv:1809.11096. sep 2018.
- [2] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow and B. Frey, “Adversarial autoencoders”, arXiv preprint arXiv:1511.05644.
- [3] A. N. Amalia, A. F. Huda, D. R. Ramdania and M. Irfan, "Making a Batik Dataset for Text to Image Synthesis Using Generative Adversarial Networks," 2019 IEEE 5th International Conference on Wireless and Telematics (ICWT), Yogyakarta, Indonesia, pp. 1-7, 2019.
- [4] A. Radford, J. W. Kim, C. Hallac, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, and G. Krueger, “Learning transferable visual models from natural language supervision”, International conference on machine learning, pp. 8748-8763, 2021.
- [5] A. Radford, L. Metz, L. and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks.”, arXiv preprint arXiv:1511.06434.
- [6] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, “Hierarchical text-conditional image generation with clip latents”, arXiv preprint arXiv:2204.06125. 2022 Apr 13;1(2):3.
- [7] A.Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, Mishkin, B. McGrew, I. Sutskever, and M. Chen, “GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models”, In International Conference on Machine Learning (pp. 16784-16804). PMLR, June 2022.
- [8] B. Li, X. Qi, T. Lukasiewicz, and P. Torr, “Controllable text-to-image generation”, Advances in Neural Information Processing Systems, 2019.
- [9] B. Liu, K. Song, Y. Zhu, G. D. Melo, and A. Elgammal, “TIME: Text and Image Mutual-Translation Adversarial Networks”, In: *Proc of the AAAI Conf. on Artificial Intelligence*, pp. 2082-2090, 2021.
- [10] C. Ledig et al., "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," 2017 IEEE Conference on Computer

- Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 105-114, 2017.
- [11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision", In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818-2826, 2016.
 - [12] C. Zhang, M. Zhang, and I.S. Kweon, "Text-to-image Diffusion Model in Generative AI: A Survey." arXiv preprint arXiv:2303.07909., 2023.
 - [13] C. Zhang, Z. Yang, X. He and L. Deng, "Multimodal Intelligence: Representation Learning, Information Fusion, and Applications," in IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 3, pp. 478-493, March 2020.
 - [14] E. A. Kenan, Y. Sun and J. H. Lim, "Learning Cross-Modal Representations for Language-Based Image Manipulation," 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, pp. 1601-1605, 2020.
 - [15] H. Dong, J. Zhang, D. McIlwraith and Y. Guo, "I2T2I: Learning text to image synthesis with textual data augmentation," 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, pp. 2015-2019, 2017.
 - [16] H. Lee, U. Ullah, J. -S. Lee, B. Jeong and H. -C. Choi, "A Brief Survey of text driven image generation and manipulation," 2021 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), Gangwon, Korea, pp. 1-4, 2022.
 - [17] H. Wang, G. Lin, S. C. H. Hoi, and C. Miao, "Cycle-consistent inverse GAN for text-to-image synthesis", 29th ACM International Conference on Multimedia (MM '21), 630-638, 2021.
 - [18] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks", In: ICCV, 2017.
 - [19] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "StackGAN++: Realistic image synthesis with stacked generative adversarial networks", IEEE transactions on pattern analysis and machine intelligence, 41(8), pp. 1947-1962, 2018.

- [20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, X. Bing, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets", *Communications of the ACM*, Vol. 63, pp.139-144,2022.
- [21] J. Agnese, J. Herrera, H. Tao, and X. Zhu, "A survey and taxonomy of adversarial neural networks for text-to-image synthesis", *WIREs Data Mining and Knowledge Discovery*, 10(4).
- [22] J. Pennington, R. Socher, and C.D. Manning, "Glove: Global vectors for word representation", In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532-1543, October 2014.
- [23] J. -Y. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 2242-2251, 2017.
- [24] K.W. Church, "Word2Vec. *Natural Language Engineering*", 23(1), pp.155-162 ,2017.
- [25] L. Mei, X. Ran and J. Hu, "Weakly Supervised Attention Inference Generative Adversarial Network for Text-to-Image," 2019 IEEE Symposium Series on Computational Intelligence (SSCI), Xiamen, China, pp. 1574-1578, 2019.
- [26] L. S. Hanne, R. Kundana, R. Thirukkumaran, Y. V. Parvatikar and K. Madhura, "Text-To-Image Synthesis Using Modified GANs," 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2022, pp. 1-7, 2022.
- [27] L. Xiaolin and G. Yuwei, "Research on Text to Image Based on Generative Adversarial Network," 2020 2nd International Conference on Information Technology and Computer Application (ITCA), Guangzhou, China, pp. 330-334, 2020.
- [28] M. E. Nilsback, and A. Zisserman, "Automated flower classification over a large number of classes: IEEE Sixth Indian Conference on Computer Vision, Graphics & Image Processing, pp.722-729, 2008.
- [29] M. Heusel, H., Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium", *Advances in neural information processing systems*, 2017.

- [30] M. Mirza, and S. Osindero, "Conditional generative adversarial nets., arXiv preprint arXiv:1411.1784.
- [31] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks", IEEE Transactions on Signal Processing, Vol. 45, No. 11, pp. 2673-2681, 1997.
- [32] M. Tao, H. Tang, F. Wu, F. X. Y. Jing, B. K. Bao and C. Xu, "DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis", In: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16515-16525, 2022.
- [33] M. Yuan and Y. Peng, "Bridge-GAN: Interpretable Representation Learning for Text-to-Image Synthesis," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 11, pp. 4258-4268, Nov 2020.
- [34] M. Zhu, P. Pan, W. Chen, and Y. Yang, "DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-To-Image Synthesis", In: Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5802-5810, 2019.
- [35] M.Kratky, V.Snasel, J.Pokorny, P.Zezula, "Efficient Processing of Narrow Range Queries in Multi-dimensional Data Structures". International Database Engineering & Applications Symposium, December 2006.
- [36] N. K. Htwe and W. P. Pa, "Building Annotated Image Dataset for Myanmar Text to Image Synthesis", In: Proc. of IEEE Conf. on Computer Applications (ICCA), Yangon, Myanmar, pp. 194-199, 2023.
- [37] N. K. Htwe, and W. P. Pa, "Generative Adversarial Networks for Myanmar Text to Image Synthesis", In: Proc. of International Conf. on Communication and Computer Research, Seoul, Korea, 2022.
- [38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge", IJCV, 115(3), pp. 211–252, 2015.
- [39] P. Sumi, S. Sindhuja and S. Sureshkumar, "A Comparison between AttnGAN and DF GAN: Text to Image Synthesis," 2021 3rd International Conference on Signal Processing and Communication (ICPSC), Coimbatore, India, pp. 615-619, 2021.

- [40] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network", International Conference on Engineering and Technology (ICET), pp. 1-6,2017.
- [41] S. Frolov, T. Hinz, Fi. Raue, J. Hees, and A. Dengel, "Adversarial text-to-image synthesis: A review", Neural Networks: Official Journal of the International Neural Network Society, 144, pp.187-209,2021.
- [42] S. Naveen, M.S.R. Kiran, M. Indupriya, T.V. Manikanta and P.V. Sudeep, "Transformer models for enhancing AttnGAN based text to image generation", Image and Vision Computing, 115, p.104284.
- [43] S. Reed, Z. Akata, B. Schiele, and H. Lee, "Learning deep representations of fine-grained visual descriptions", CVPR, 2016.
- [44] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele and H. Lee, "Generative adversarial text to image synthesis: In: Proc of International conference on machine learning, pp. 1060-1069, 2016.
- [45] S. Ruan, Y. Zhang, K. Zhang, Y. Fan, F. Tang, Q. Liu, and E. Chen, "DAE-GAN: Dynamic Aspect-aware GAN for Text-to-Image Synthesis", In: *Proc of IEEE/CVF International Conf. on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 13940-13949, 2021.
- [46] S. S. Baraheem and T. V. Nguyen, "Aesthetic-Aware Text to Image Synthesis," 2020 54th Annual Conference on Information Sciences and Systems (CISS), Princeton, NJ, USA, pp. 1-6, 2020.
- [47] S. Wah, P. Welinder, P. Perona, S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset", Technical report CNS-TR-2011-001, California Institute of Technology.
- [48] T. Baltrušaitis, C. Ahuja and L. -P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423-443, 1 Feb. 2019.
- [49] T. Hinz, S. Heinrich, and S. Wermter, "Semantic object accuracy for generative text-to-image synthesis", *IEEE transactions on pattern analysis and machine intelligence.*, 44(3), 1552-1565, 2020.
- [50] T. Karras, S. Laine and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 12, pp. 4217-4228, 1 Dec. 2021.

- [51] T. Qiao, J. Zhang, D. Xu, and D. Tao, "MirrorGAN: Learning text-to-image generation by redescription", In: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1505-1514, 2019.
- [52] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans", In NIPS, pages 2234–2242, 2016.
- [53] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks", In: Proc. of the IEEE conference on computer vision and pattern recognition, pp. 1316-1324, 2018.
- [54] W. Li, S. Wen, K. Shi, Y. Yang and T. Huang, "Neural Architecture Search with a Lightweight Transformer for Text-to-Image Synthesis," in *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 3, pp. 1567-1576, 1 May-June 2022.
- [55] W. Liao, K. Hu, M. Y. Yang, and B. Rosenhan, "Text to image generation with semantic-spatial aware GAN", In: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18187-18196, 2022. December 2014.
- [56] X. Huang, M. Wang, and M. Gong, "Hierarchically-fused generative adversarial network for text to realistic image synthesis", *IEEE 16th Conference on Computer and Robot Vision (CRV)*, pp. 73-80, 2019.
- [57] X. Wu, H. Zhao, L. Zheng, S. Ding and X. Li, "Adma-GAN: Attribute-Driven Memory Augmented GANs for Text-to-Image Generation", In: *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1593-1602, 2022.
- [58] Y. Gou, Q. Wu, M. Li, B. Gong, M. Han, "Segattngan: Text to image generation with segmentation attention", arXiv preprint arXiv:2005.12444., 2020.
- [59] Y. Zhang, S. Liu, C. Dong, X. Zhang and Y. Yuan, "Multiple Cycle-in-Cycle Generative Adversarial Networks for Unsupervised Image Super-Resolution," in *IEEE Transactions on Image Processing*, vol. 29, pp. 1101-1112, 2020.

- [60] Y. Zhou et al., "Towards Language-Free Training for Text-to-Image Generation," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, pp. 17886-17896, 2022.
- [61] Y. Zhou, "Generative adversarial network for text-to-face synthesis and manipulation", In Proceedings of the 29th ACM International Conference on Multimedia (pp. 2940-2944).
- [62] Z. Chen and Y. Luo, "Cycle-Consistent Diverse Image Synthesis from Natural Language," 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Shanghai, China, pp. 459-464,2019.
- [63] Z. Zhang, C. Fu, J. Zhou, W. Yu and N. Jiang, "Text to Image Synthesis based on Multi - Perspective Fusion," 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 2021, pp. 1-8, 2021.
- [64] Z. Zhang, Y. Xie and L. Yang, "Photographic Text-to-Image Synthesis with a Hierarchically-Nested Adversarial Network," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 6199-6208,2018.

APPENDICES

This section contains about the various testing of this research with visual interface. Good graphical interface makes the user to easily understand the process and makes the system more interesting and approachable. Therefore, GUI has been designed by using Flask (web application) framework with python to show the user creativity.

Figure. 8. 1 shows the home page of Myanmar text to image system using generative adversarial networks. Firstly, the user needs to input the desire captions and click generate button to create images based on Myanmar text descriptions. As a limitation, the model is only trained on a specific annotated images dataset. Therefore, the user needs to enter the descriptions relevant to the features of flowers such as color, shape, style and boundary. The example of various images generated from Myanmar text descriptions using DFGAN+MSM (proposed model), DF-GAN (baseline model 1) and AttnGAN (baseline model 2) are shown in the following figures. As a result, the proposed model can generate the high-quality images with better semantic consistency than baseline models.

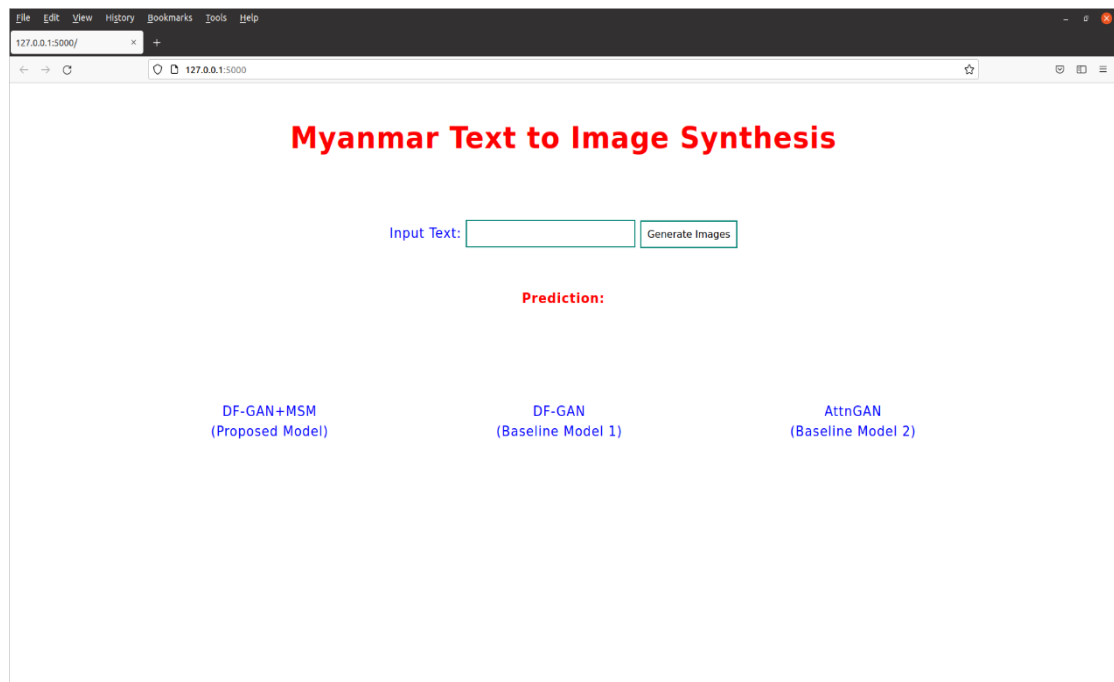


Figure. A.1 Home Page of Myanmar T2I

In figure A.1 and figure A.3 shows the impact of replacing some words in the sentence “input text 1” with “input text 2”. As a result, there were some changes in the shape of petal of flowers. Therefore, the result of the generated images is depend on each word of the sentence.

Table A.1. Query Input 1

Input Text 1	<p>ပန်းရောင်ပန်းပွင့်</p> <p>the pink flower</p>
Input Text 2	<p>ပန်းရောင်ရှိသောပန်းပွင့်</p> <p>the flower with the pink color</p>

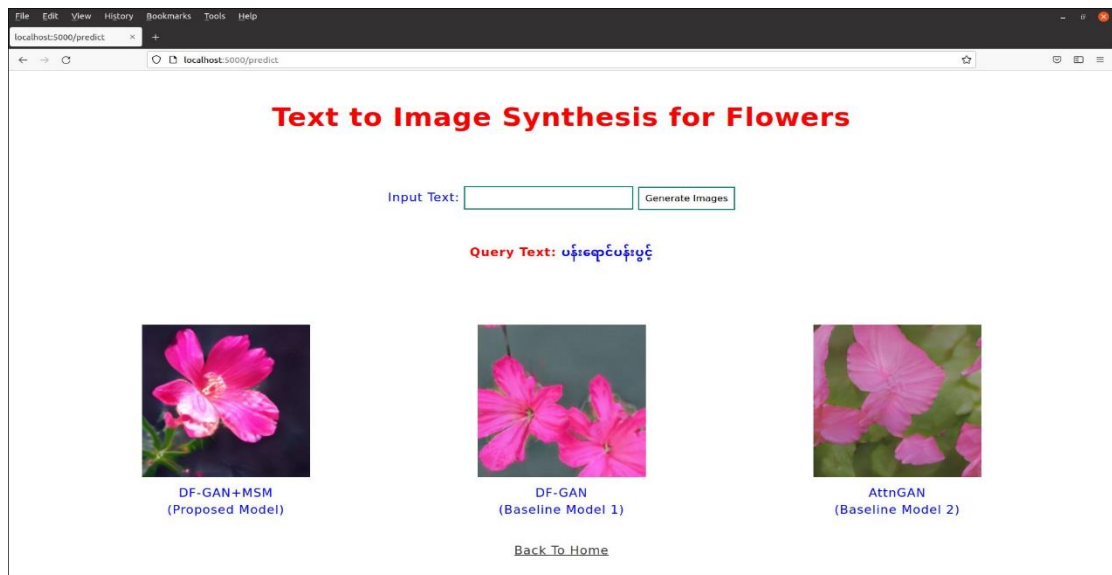


Figure A.2 The Generated Images from Query Text 1

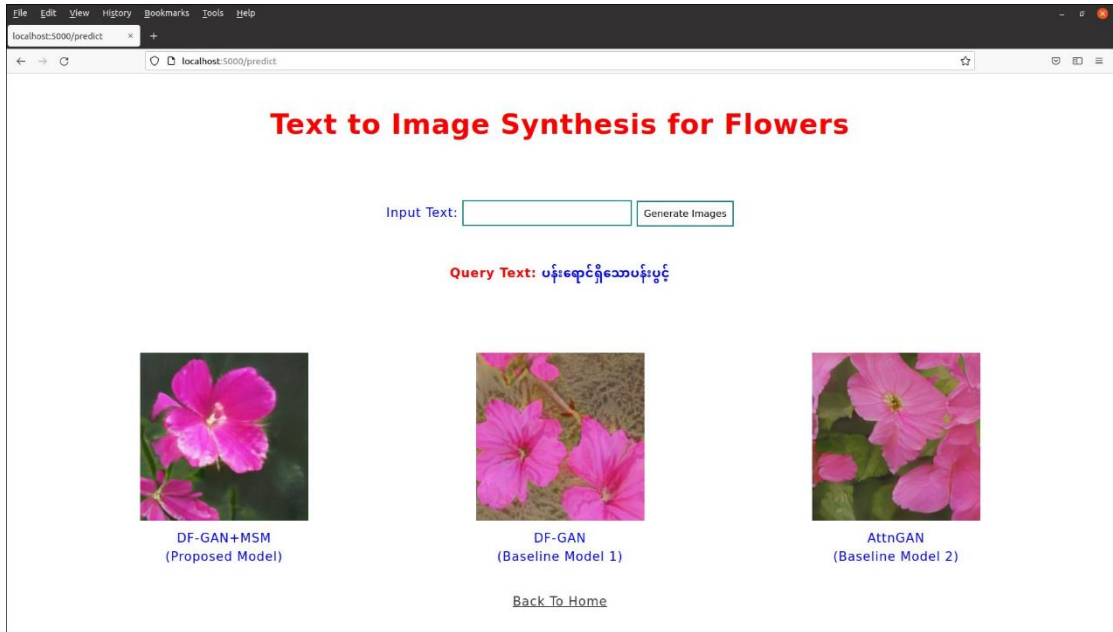


Figure A.3 The Generated Images from Query Text 2

The image generated from “**Input Text 3**” (“The flower with the large pink petal”) and “**Input Text 4**” (“The flower with the large red petal”) is shown Figure A.4 and Figure A.5. In this results, the image are generated based on the shape of petal such as large or small petal.

Table A.2. Query Input 2

Input Text 3	ပွင့်ချပ်ကြီးကြီးပါသောပန်းရောင်ပန်းပွင့် the pink flower with large pink petal
Input Text 4	ပွင့်ချပ်ကြီးကြီးပါသောအနီရောင်ပန်းပွင့် the flower with the large red petal

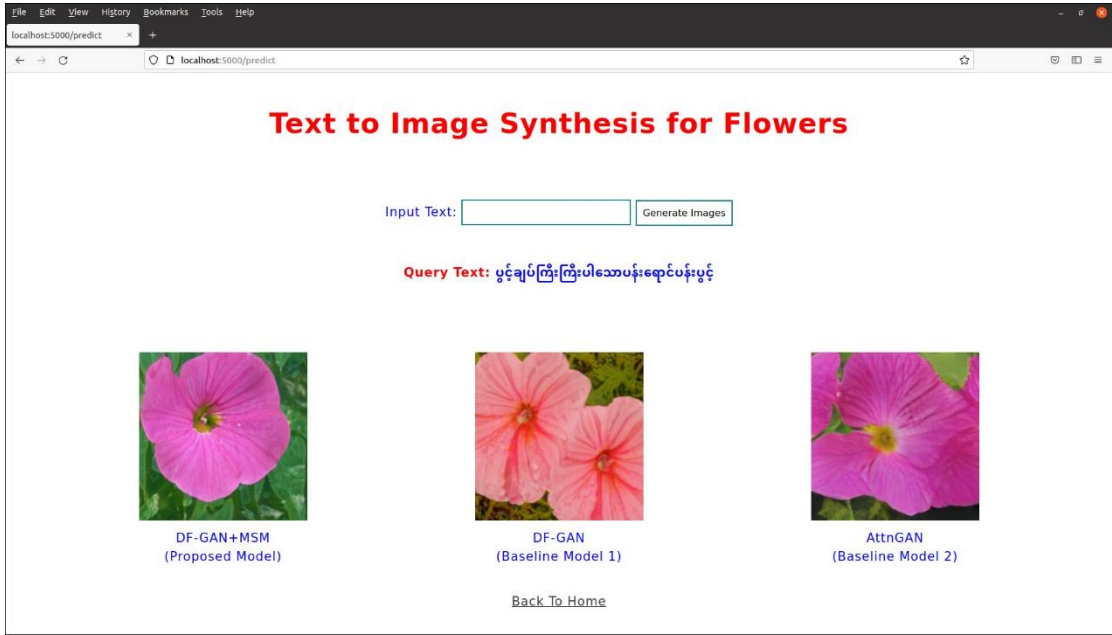


Figure A.4 The Generated Images from Query Text 3

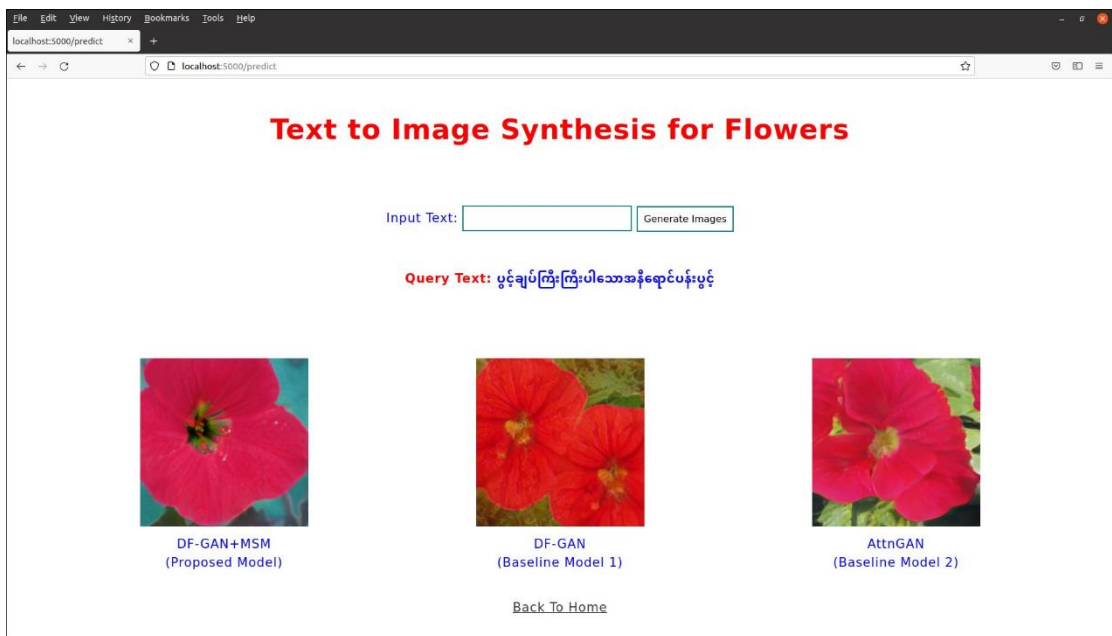


Figure A.5 The Image Generated from Input Text 4

This can also generate the flower with two colors based on the text descriptions that contains two color combination words such as “the flower with the red and white petals”. The results of the generated images for “input text 5” and “input text 6” are shown in Figure A.6 and Figure A.7.

Table A.3. Query Input 3

<p>Input Text 5</p>	<p>ပန်းပွင့်ပေါ်တွင်အနီရောင်နှင့်အဖြူရောင်တစ်ဝက်စီရှိသောပွင့်ချပ်ရှိတယ် The flower has half and half of red and white petals.</p>
<p>Input Text 6</p>	<p>ပန်းပွင့်တွင်အဝါရောင်နှင့်အဖြူရောင်ရှိသောပွင့်ချပ်များရှိတယ် The flower has the yellow and white petals.</p>

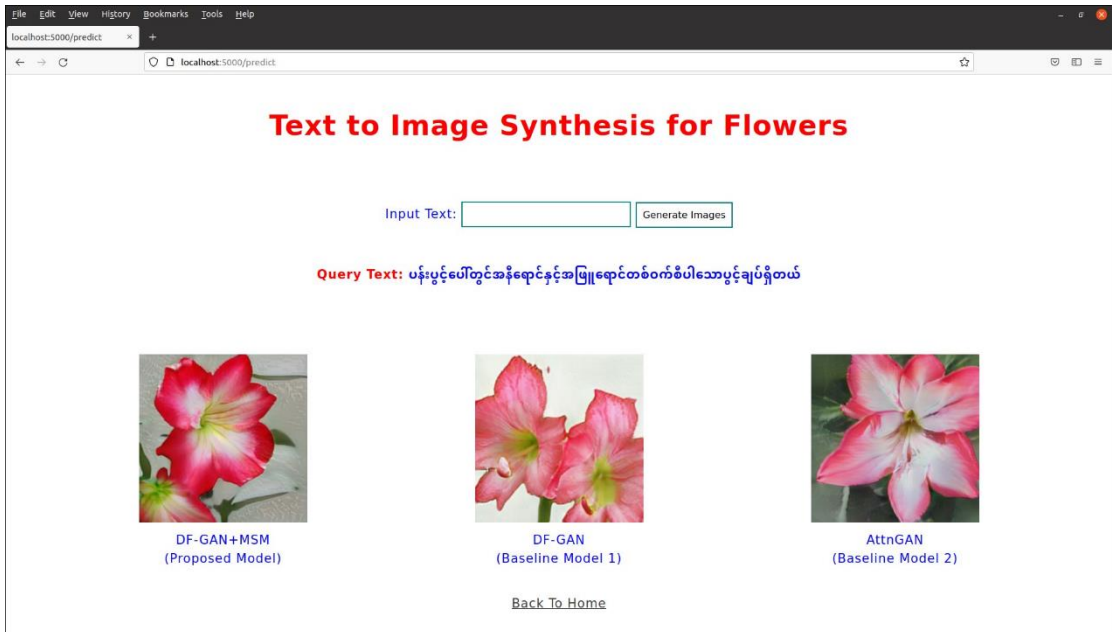


Figure A.6 The Generated Images from Query Text 5

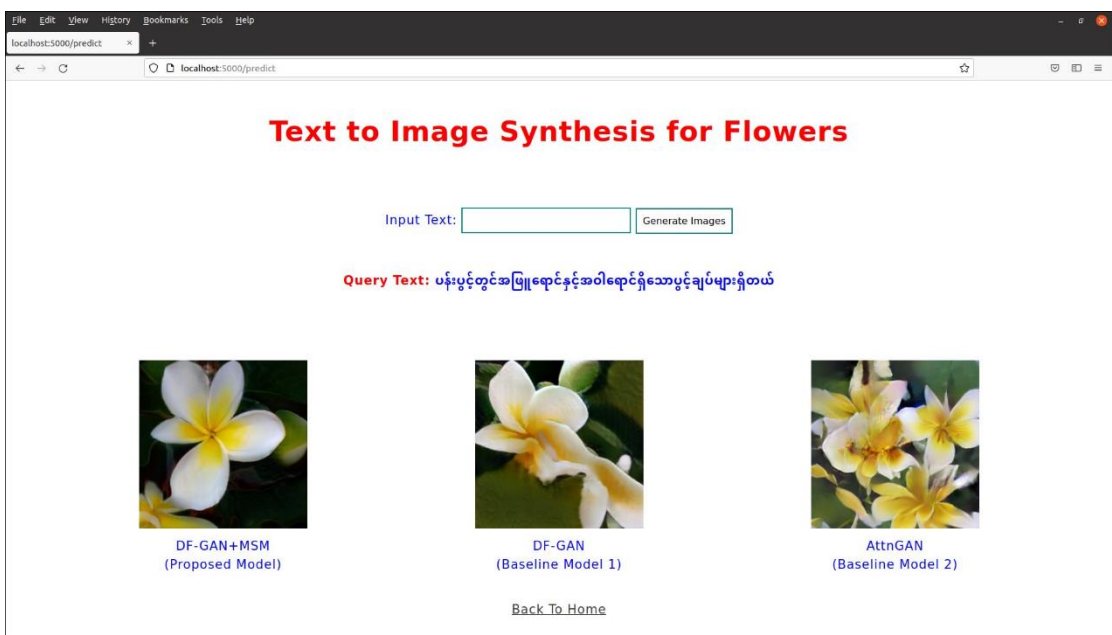


Figure A.7 The Generated Images from Query Text 6

Figure A.8 and Figure A.9 shows output from “input text 7” and “input text 8”. These two images are generated based on these words such as shape, color and spot of the petals.

Table A.4. Sample Test 4

Input Text 7	<p>အနက်ရောင်လိုင်းပါသောခရမ်းရောင်ပန်းပွင့်</p> <p>The purple flower with the black line</p>
Input Text 8	<p>ပန်းပွင့်ပေါ်တွင်အနက်ရောင်အစက်ပြောက်ပါသောအဖြူရောင်ပွင့်ချပ်ရှိတယ်</p> <p>The flower has the white petals with dark spots.</p>

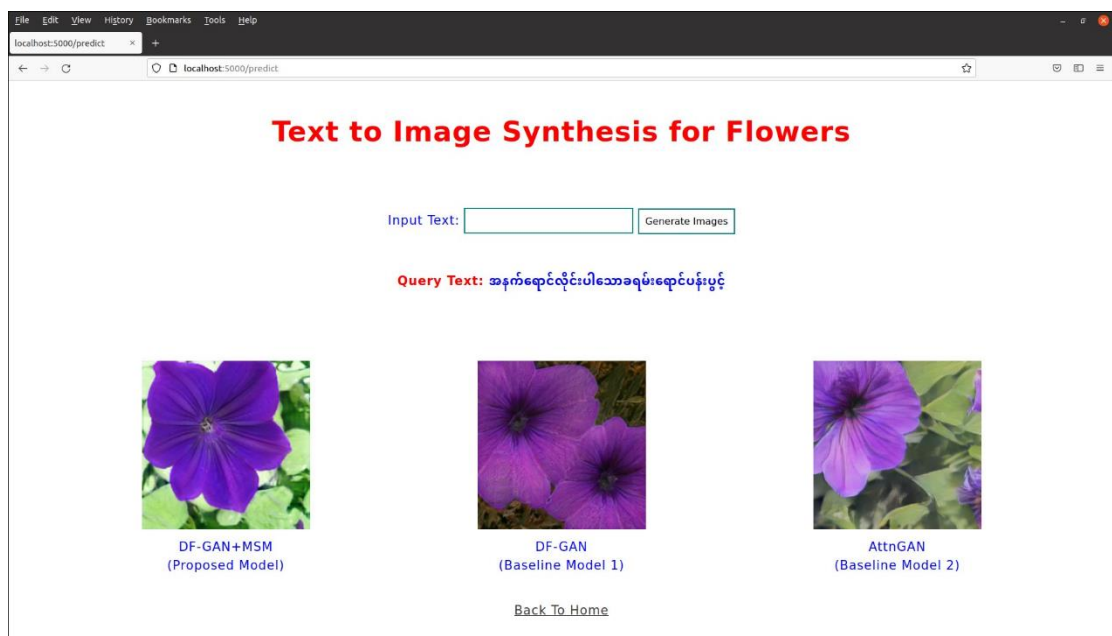


Figure A.8 The Generated Images from Query Text 7

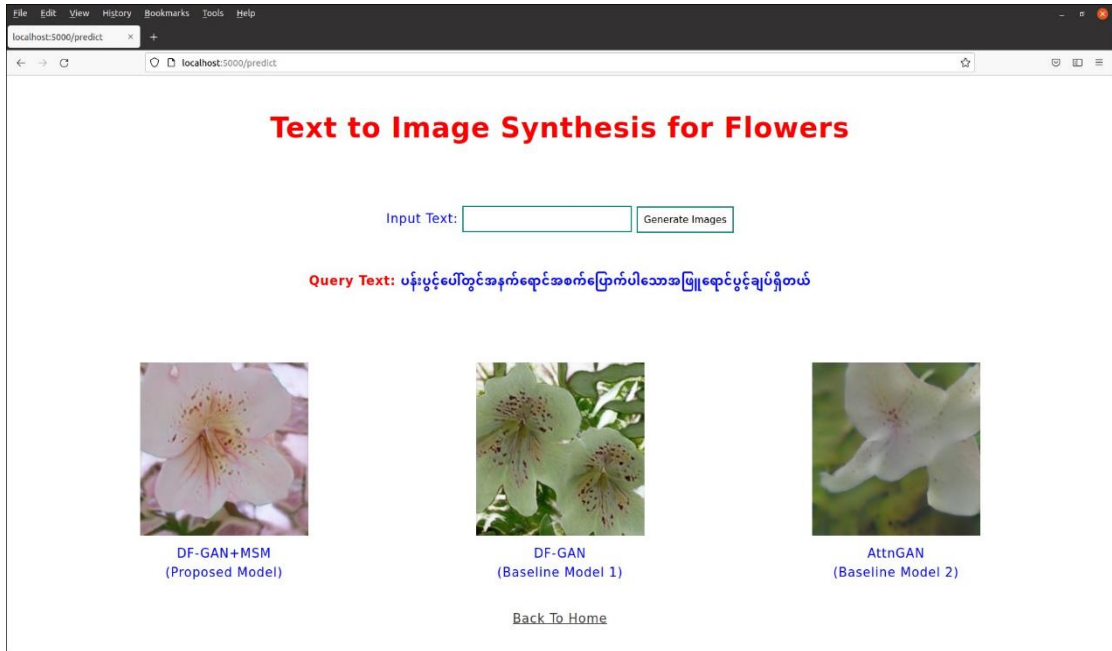


Figure A.9 The Generated Images from Query Text 8

The images generated based on the layer of flower is shown in Figure A.10 and Figure A.11. The first image consists of the flower with single layer of petal and the flower with multiple layers of petal is in the second.

Table A.5. Query Input 5

Input Text 9	အဖြူရောင်ပွင့်ချပ်တစ်လွှာရှိသောပန်း The flower with single layer of white
Input Text 10	ပန်းပွင့်ပေါ်တွင်ရှည်လျားပြီးထပ်နေသောခရမ်းရောင်ပွင့်ချပ်များရှိတယ် The flower has the long and overlap purple petals.

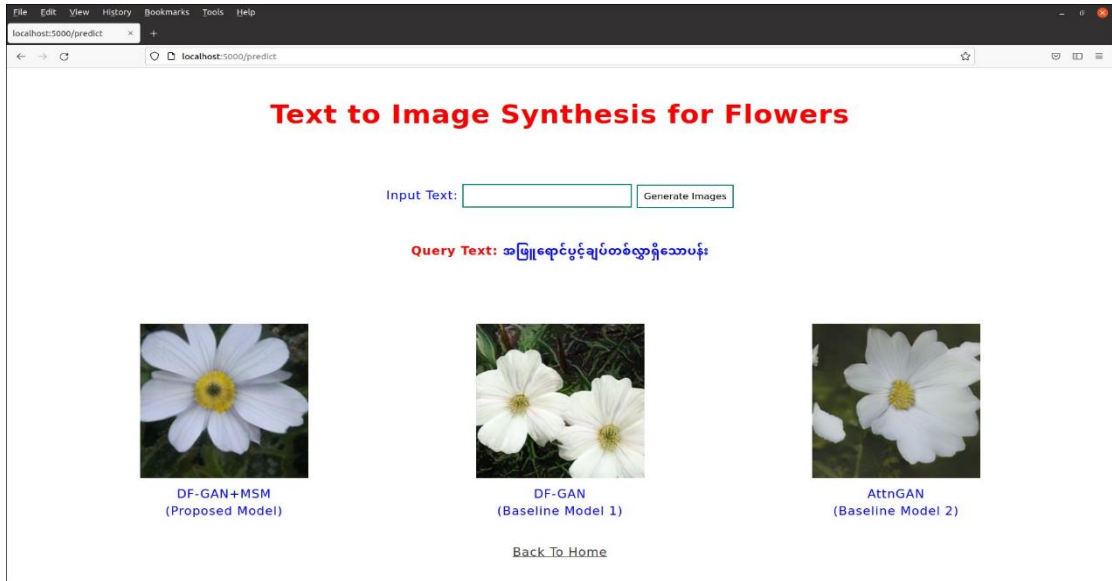


Figure A.10 The Generated Images from Query Text 9

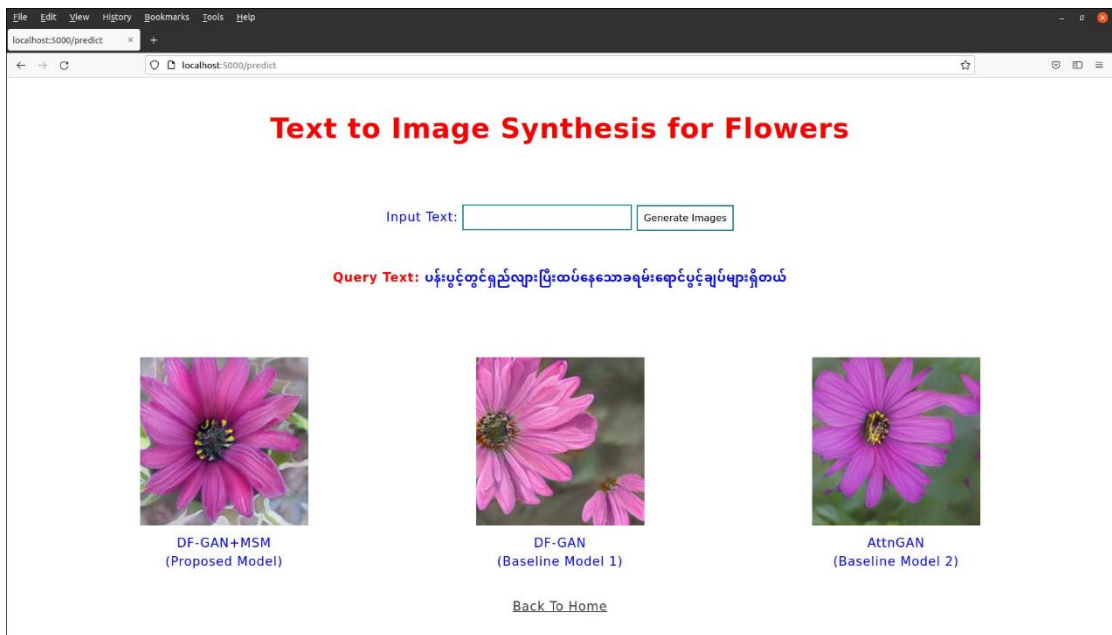


Figure A.11 The Generated Images from Query Text 10

The following section describes the visual testing of English text-to-image that is conducted using Caltech Birds Dataset. As a limitation, this model only trained on an annotated image dataset that contains only bird species. Therefore, this model enables to generate the bird images and the users need to input English text descriptions. Table A.6 shows the various query input text conditioned on shape, each part with different colors of the bird.

Table A.6 Query Input Text on English T2I

Input Text 11	this bird is brown and white in color, and has a brown beak.
Input Text 12	this bird has wings that are black and has a white belly
Input Text 13	this bird is white with grey and has a long, pointy beak.
Input Text 14	a small brown bird with yellow along the top of the feathers.
Input Text 15	this bird is white with brown and has a very short beak.
Input Text 16	a white body bird with a regularly sized head in comparison to the body.
Input Text 17	this bird has wings that are black and has a red belly

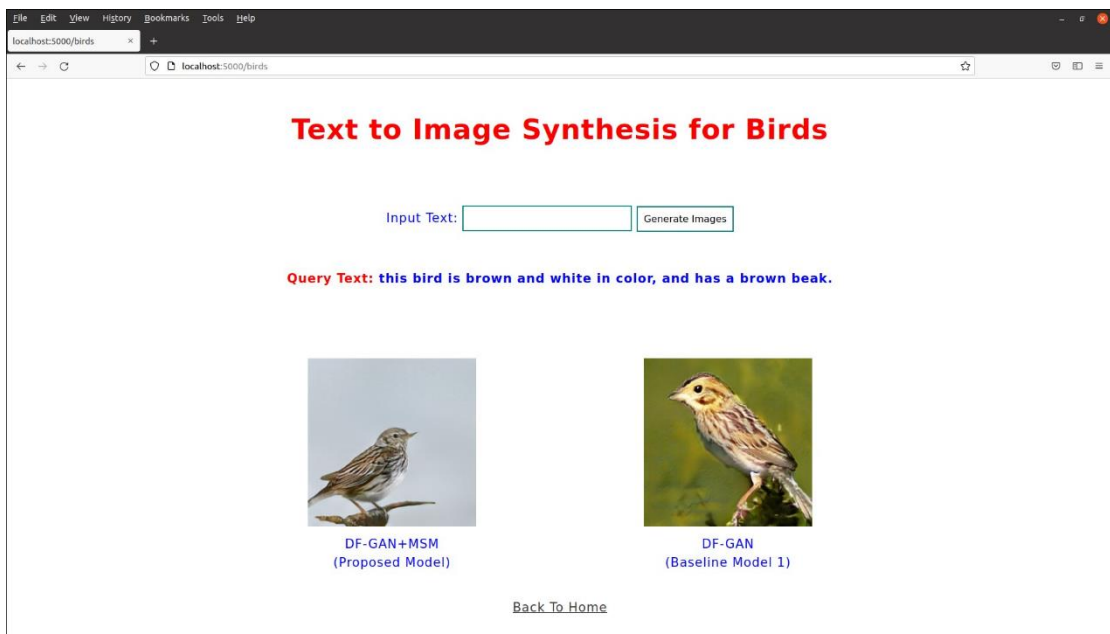


Figure A.12 The Generated Images from Query Text 11



Figure A.13 The Generated Images from Query Text 12

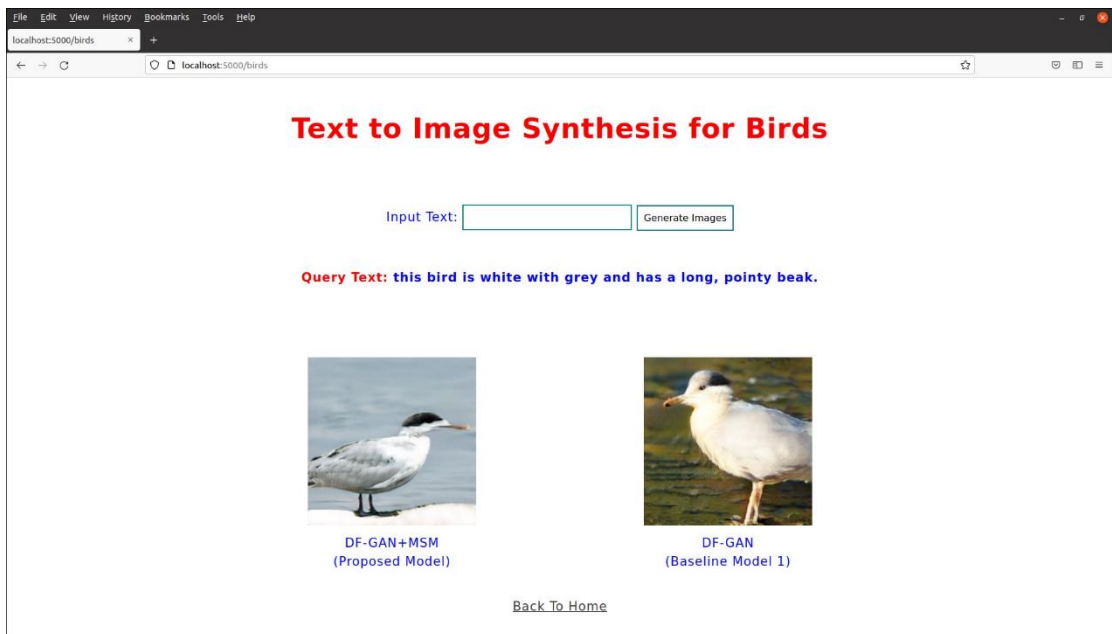


Figure A.14 The Generated Images from Query Text 13

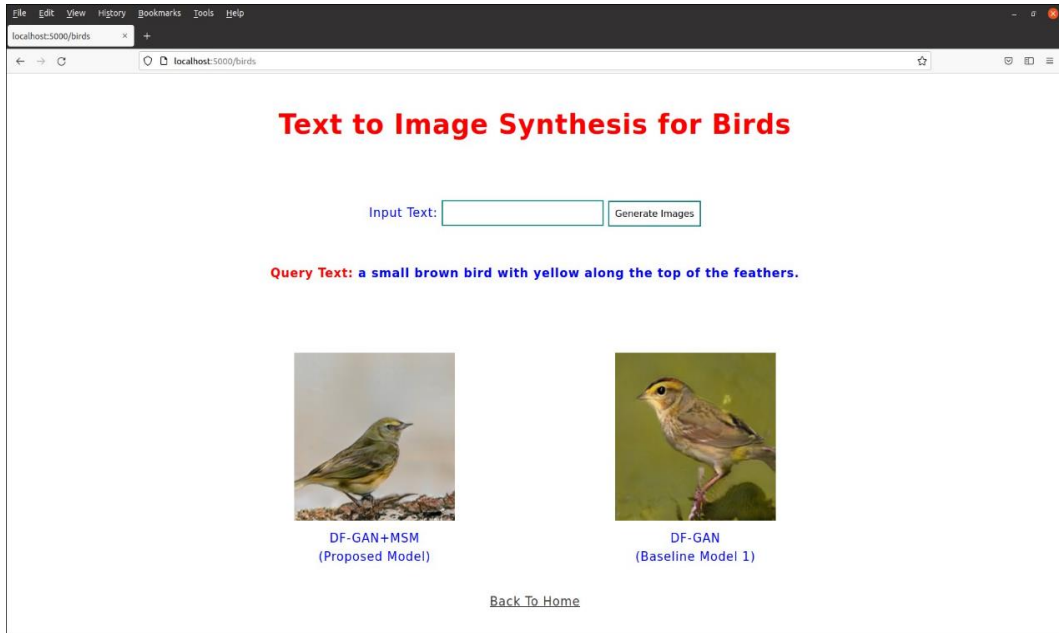


Figure A.15 The Generated Images from Query Text 14

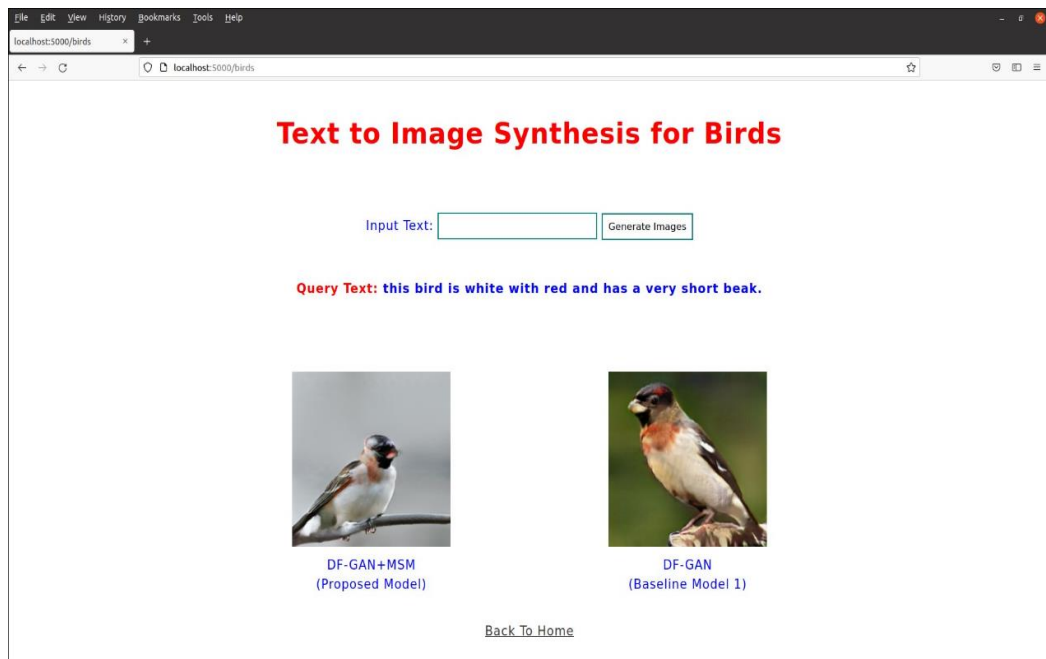


Figure A.16 The Image Generated from Input Text 15

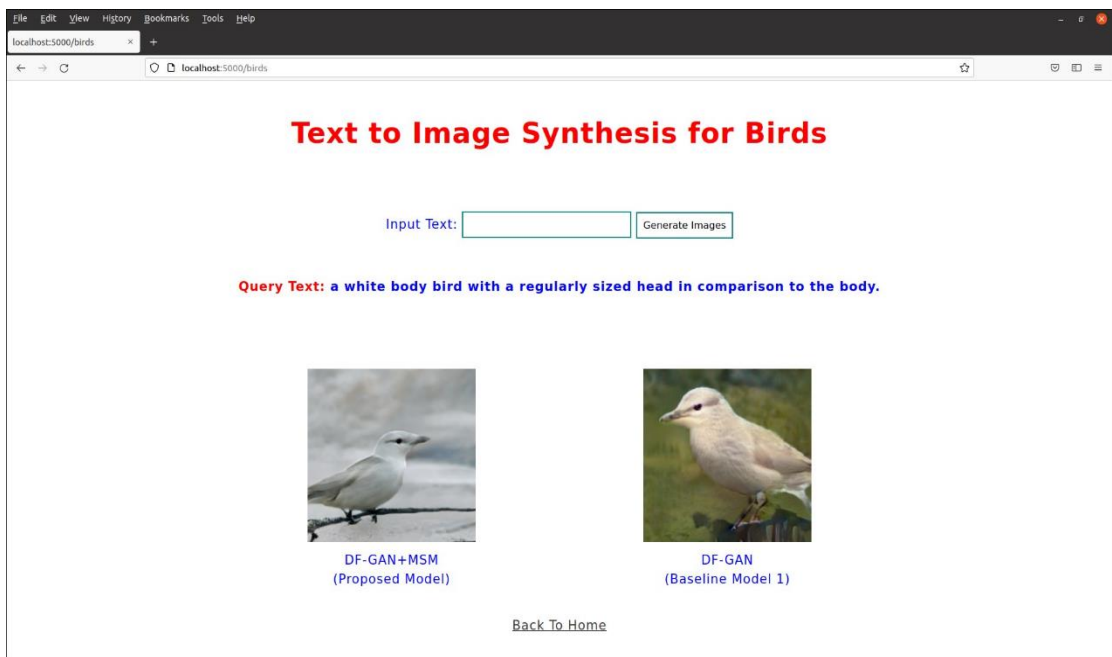


Figure A.17 The Generated Images from Query Text 16

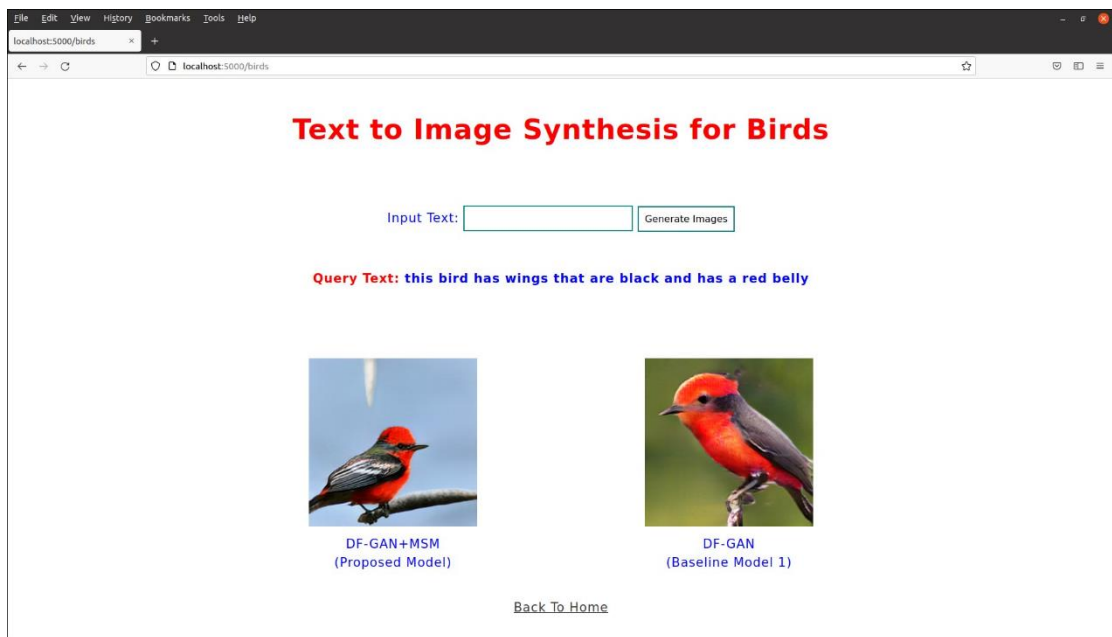


Figure A.18 The Generated Images from Query Text 17

LIST OF ACRONYMS

AAE	Adversarial Autoencoder
ADR	Aspect aware Dynamic Re-drawer
AGR	Attended Global Refinement
AttnGAN	Attentional Generative Adversarial Networks
BERT	Bidirectional Encoder Representations from Transformers
BoW	Bag-of-Words
cGANs	Conditional Generative Adversarial Network
CNN	Convolutional Neural Network
CV	Computer Vision
DAE-GAN	Dynamic Aspect-aware Generative Adversarial Networks
DAMSM	Deep Attentional Multimodal Similarity Model
DCGAN	Deep Convolutional Neural Network
DFGAN	Deep Fusion Generative Adversarial Network
DM-GAN	Dynamic Memory Generative Adversarial Networks
FID	Fréchet Inception Distance
GAN	Generative Adversarial Networks
GLIDE	Guided Language-to-Image Diffusion for Generation and Editing
GPT	Generative Pre-trained Transformers
LSTM	Long Short-Term Memory
MA-GP	Matching-Aware Gradient Penalty
MSM	Multimodal Similarity Model
RNN	Recurrent Neural Network
NLP	Natural Language Processing

SSA-GAN	Semantic-Spatial Aware Generative Adversarial Network
STREAM	Semantic Text Regeneration and Alignment module
T2I	Text to Image