Received: October 5, 2023. Revised: December 2, 2023.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Multimodal Generative Model Based Text-to-Image Synthesis

Nang Kham Htwe[1]* **Win Pa Pa[1]**

[1]*Natural Language Processing Lab, University of Computer Studies, Yangon, Myanmar*
* Corresponding author's Email: nangkhamhtwe@ucsy.edu.mm

Abstract: Text-to-image synthesis (T2I) is a challenging task because the model requires to create high-quality images that are semantically consistent and realistic. Therefore, the main objective of this paper is to improve the quality of generated images and the similarity level between text descriptions and these images. In this paper, we proposed deep-fusion generative adversarial networks (DF-GAN) with multimodal similarity model (MSM) to generate high-resolution images with better consistency between text and the generated images. In this work, MSM is pretrained using real images with captions in the dataset. This pretrained model is used to improve the visual-semantic consistency level during training of T2I. This paper investigates the improvement in the image generation process due to applying MSM model to the generator. The investigation is performed on two different datasets with different languages to prove our proposed model outperforms baseline DF-GAN. Firstly, the experiment is done on Caltech- Birds dataset and the evaluative results of the proposed model are compared with state-of-the-art models, StackGAN, StackGAN++, AttnGAN, DMGAN, DAE-GAN, TIME, DF-GAN. According to the comparative results, the proposed model outperforms the baseline state-of-the-art models in terms of Fréchet inception distance (FID) score and inception score. The improvements of the proposed model on synthesizing images over the baseline models are proved in terms of image quality and visual-semantic similarity in this work. Accordingly, the proposed model is applied on Myanmar text-to-image synthesis (Myanmar T2I) with Oxford 102 flowers dataset annotated in Myanmar to prove the effectiveness of the proposed model on different dataset and different language. To the extent of our knowledge, this is the first attempt of implementing generative adversarial networks with multimodal in Myanmar T2I and it got inception score of 3.54 ± 0.03 and FID score of 49.97. In Myanmar T2I, the proposed model also got better performance than the baseline DF-GAN because it achieved higher inception score and lower FID score than the baseline model. Drawing upon the findings derived from the outcomes of the experiments, it is evident that the proposed model demonstrates an enhancement in the quality of generated images within the context of the image generation.

Keywords: Generative adversarial networks, Multimodal similarity model, Text-to-image synthesis, Inception score, Fréchet inception distance (FID).

1. Introduction

All text-to-image synthesis has seen rapid progress in recent years after introducing of generative adversarial networks (GAN) [1]. It is an exciting area of research for both computer vision (CV) and natural language processing (NLP) communities. State-of-the-art models use attention mechanisms and multiple stages of image generation to produce high-quality images that are visually satisfying and align with semantic meaning of text [2-5]. But, the use of these multiple refinement stages brings unstable training process and higher computation costs. To address this issue, DF-GAN [6] introduces one stage T2I without using extra networks to evaluate the consistency level between the generated images and text.

Despite the progress that has been made by previous work [6], there are some artifacts that need to develop on the generated images such as color consistency, boundary and shape of the objects. Accordingly, pretrained MSM is applied to the generator side of the DF-GAN to more semantically understand textual descriptions and more correctly visualize the images from text descriptions. By

introducing additional knowledge to DFGAN with MSM, namely DFGAN+MSM, the generated image is more accuracy in shape and color, and more semantically consistent on T2I implemented with Caltech-UCSD Birds dataset annotated in English [7]. Moreover, DFGAN+MSM can create the bird images with different texture and color. As shown in the first input text of Fig. 3, the image generated from DFGAN+MSM has more colorful than the baseline (e.g. the wingbars have different texture and color).

Therefore, the proposed model is applied to another type of dataset with Myanmar text description to prove the performance of MSM model. But there is none of implementation about the visual representation of Myanmar textual description. In addition, there is no suitable a large dataset of images with corresponding Myanmar textual descriptions that are both accurate and varied. Therefore, manually constructed Myanmar captions corpus based on Oxford 102 flowers [8] is proposed to conduct the first Myanmar T2I. The proposed model also gives advancement in synthesizing images from Myanmar text which is a complex and morphological rich language because MSM adjusts the gaps of relationship between visual and semantic. As shown in the third and the fourth input text of Fig. 7, it is clearly seen that the image generated from DFGAN+MSM has more consistency in visual-semantic and color than the others (e.g. half and half of red and white, the black line).

In this paper, there are two main contributions:

We proposed multimodal similarity modal that is pretrained on real images with captions in the dataset. This model is used to compute the similarity between the generated images and textual descriptions during training of T2I.

In this work, Myanmar annotated image dataset based on Oxford 102 flower is prepared to implement Myanmar T2I.

Finally, we investigate the advancement of the image generation process due to applying MSM to the generators using two different datasets with different languages. Moreover, the effectiveness of this multimodal over T2I is evaluated by comparing with state-of-the-arts models.

The rest sections of the paper are organized as follows. The existing research works related to T2I is described in section 2. The background theory of this research works is examined in section 3 and the proposed methodology is introduced in section 4. The implementation of the proposed model is section 5 and the evaluation of the proposed models on two different datasets are in section 6 and section 7 respectively. Section 8 summarized the presented research works.

2. Related works

Generative adversarial networks introduced by Goodfellow has gained new heights of success in many computer vision problems including image-to-image translation and image generation. GANs composed of generator networks and discriminator networks, in which the generator attempts to create artificial images and the discriminator distinguish real images from fake.

The authors [2] proposed StackGAN with two-cascaded GANs to generate high-quality images with semantic consistency. They introduced conditioning augmentation (CA) to stabilize training process and to generate more diverse images for the same text description. However, this method suffers from mode collapse (produced similar images for different text descriptions) that leads to lack of diversity and less effective on semantic consistency in the image generation. To tackle these problems, StackGAN was improved to StackGAN ++ [3].

StackGAN++ contains multiple stages of generators and discriminators and they are organized as tree-like structure. They introduced conditional and unconditional feature at the discriminator side to control the quality of the generated image on conditioned variables. They also proposed color-consistency regularization at the generators to be more consistency in color from same input. This method has made more improvements over StackGAN to stabilize training process, tackle mode collapse, and improve the similarity between the generated images and text descriptions.

Previous methods [3] generates images conditioned only on the sentence-level and lack of fine-grained information at the word-level. Therefore, this leads to face challenges in generation high-quality images. To address this issue, the authors [4] proposed AttnGAN generates images conditioned on the sentence-level at the first stage. At the next two stages, the images are generated conditioned on the word-level by using attentional generative methods. As an alternative, DAMSM is used to evaluate image-text matching loss at both word-level and sentence-level during training of GAN. This work improved the quality of generated images with more semantically consistent by using word-level and sentence-level conditions.

There are two problems still remain despite multiple-stages image generation [2-4] has achieved remarkable progress. First, the quality of the generated images from refinement states is poor if the initially generated images quality is bad. 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 [5] introduced DMGAN that contains memory mechanisms at every generator to generate high-resolution images even when 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 ++.

Previous method [4] generate images on sentence-level and then refine images on word-level. However, they ignore "the aspect information" in the sentence-level that are very helpful in generating images that are realism and semantic-consistency. Therefore, the authors [9] designed DAE to generate images based on word-level, sentence-level and aspect-level. They introduced a novel Aspect-aware Dynamic Re-drawer (ADR) that contains two components for image refinement processes. Attended Global Refinement (AGR) module utilizes word-level features to improve previously generated images and Aspect-aware Local Refinement (ALR) that works on aspect-level features to refine the images in details by using local perspective. By applying these two components alternatively, this method has progressive results in generating images with more realism and image-text consistent.

StackGANs and follow works are all depends on pretrained text encoder and image encoder for image-text consistency. Therefore, the authors proposed [10] TIME that do not require extra modules or pretraining of encoder. This model improves the performance of generators by jointly training with discriminators as a language model. They build Transformer by modeling cross-model relationships between the image and text features and using hinge loss to balance the learning between generator and discriminator. In addition, they proposed 2D positional encoding for image features to acquire advantage of coordinate signals. This jointly training of T2I and image-captioning module provides effectiveness for text-consistency without supporting extra networks.

In this work, we proposed multimodal generative model to investigate the progressive results of the generated images and text-image consistency. The investigation has done on two different languages: (1) English 2I and (2) Myanmar T2I. Previous works [2-5] used multiple image refinement stages and pretrained DAMSM model to improve image-text consistency. The use of these multiple stages leads to unstable training process and higher computation. Two-way mutual translations between text and images requires complex deep learning architectures and computational resources to advance image-text consistency and quality of the generated images. Therefore, we introduced the single stage architecture with MSM to advance the quality of generated images and image-text consistency as well as to save training time and higher computational cost.

3. Background: Deep fusion GANs (DF-GAN)

The objective of DF-GAN [6] is to generate high-quality realistic image with semantic consistency. This architecture is composed with one stage of generator and discriminator. The generator has two inputs: (1) sentence vector (text descriptions encoded with pretrained text encoder) (2) a noise vector sampled by using Gaussian distributions. First, the noise vector is feed forward to fully-connected layer. The output from this layer is passed through a series of UpBlocks to upsample the image features. Each of these blocks contains a residual block, upsample layer, and DFBlocks to fuse text and image features. The sentence vector is concatenated with the noise vector at every UpBlock. After passing through all UpBLocks, the image features are converted to images (256 x 256) by using convolutional layer.

The discriminator is built with several DownBlocks and the generated images from Generators are forwarded to these layers to extract image features. These image features are concatenated with replicated sentence features and the adversarial loss is evaluated to predict visual-semantic consistency. By differentiating fake images from real images, the discriminator guides the generator to output more realistic images with better semantic consistency. Moreover, matching-aware gradient penalty is applied on discriminator side to penalize the the gradient of the discriminator with respect to input (real images and captions) to be zero. MA-GP, the discriminator guides the generator to generated more realistic images with better consistency between visual and semantic without using extra-networks to compute similarity between text and image. The formulation loss of this model for discriminator and generator is as follows:

$$L_G = -\mathbb{E}_{G(z)\sim P_g}[\mathrm{D}(\mathrm{G}(\mathrm{z}),\mathrm{e})] \tag{1}$$

$$L_D = -\mathbb{E}_{\sim P_r}[\min(0,-1+D(x,e))] \tag{2}$$

$$-\frac{1}{2}\mathbb{E}_{G(z)\sim P_g}[\min(0,-1,-D(G(z),e))]$$

Table 1. Formal notations

Notation	Descriptions
e	The sentence vector
z	The noise vector sampled from the Gaussian distribution
p, k	Two hyper-parameters to balance the effectiveness of the gradient penalty
P_g	The generated data distribution
P_r	Real data Distribution
P_{mis}	Mismatching data distribution

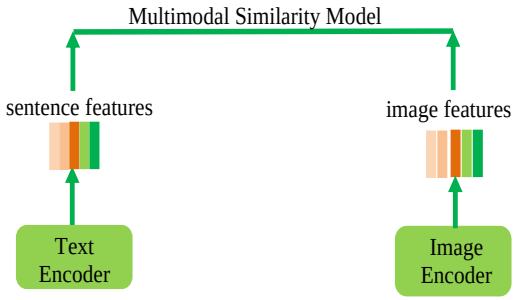

Figure. 1 Framework of multimodal similarity model

$$-\frac{1}{2}\mathbb{E}_{x\sim P_{mis}}[\min(0,-1,-D(x,e))]$$

$$+k\mathbb{E}_{x\sim P_r}[(||\nabla_x D(x,e)|| + ||\nabla_e D(x,e)||)^p]$$

4. Proposed methodology

This section contains the theorical concepts related to the proposed (1) multimodal similarity model (MSM) and (2) DFGAN+MSM to implement T2I.

4.1 Multimodal similarity model (MSM)

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 measure the similarity loss between the generated images and the corresponding input text during training of T2I. The architecture of MSM is shown in Fig. 1.

Bidirectional long short-term memory(bi-LSTM) [11], text encoder, is used to extract sentence features. The input layer takes the sentence that are pre-encoded into numbers by using index-encoding methods. The embedding layer takes the sequence of word indices from the input layer and converts to word embedding features. Afterwards, the output from this layer is passed as input parameters to bi-LSTM layers. 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. We concatenated these last hidden states to extract the whole sentence features.

The image encoder is constructed using convolutional neural network [12] and built upon Inception-v3 model [13] 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 resized images are horizontally flipped with the probability of 0.5 to improve the performance, generalization, robustness of the model. The resulted outputs are then passed to the series of intermediate convolutional layers to extract the image features. The extracted image features are forward to "mixed-6e" layers of inception-v3 model. Finally, we get the global image feature by using average pooling layer of Inception-v3 model.

The dimension of the visual embedding is then resized to be the same size as the dimension of the sentence embedding and these two embeddings are projected into a semantic space. In this work, the dimension of text embedding is 256. 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:

$$C(I,T) = I * T / \| I \| * \| T \| \quad (3)$$

The objective of similarity learning is to further improve visual-semantic correlation and the quality of generated images during training of 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.

$$Loss_1 = \sum_{k=1}^{N} \log P(T_k|I_k) \quad (4)$$

$$Loss_2 = \sum_{k=1}^{N} \log P(I_k|T_k) \quad (5)$$

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

$$L_{MSM} = Loss_1 + Loss_2 \quad (6)$$

4.2 Architecture of DFGAN+MSM

The proposed model is built upon DF-GAN by adding MSM model to the generator side of DF-GAN

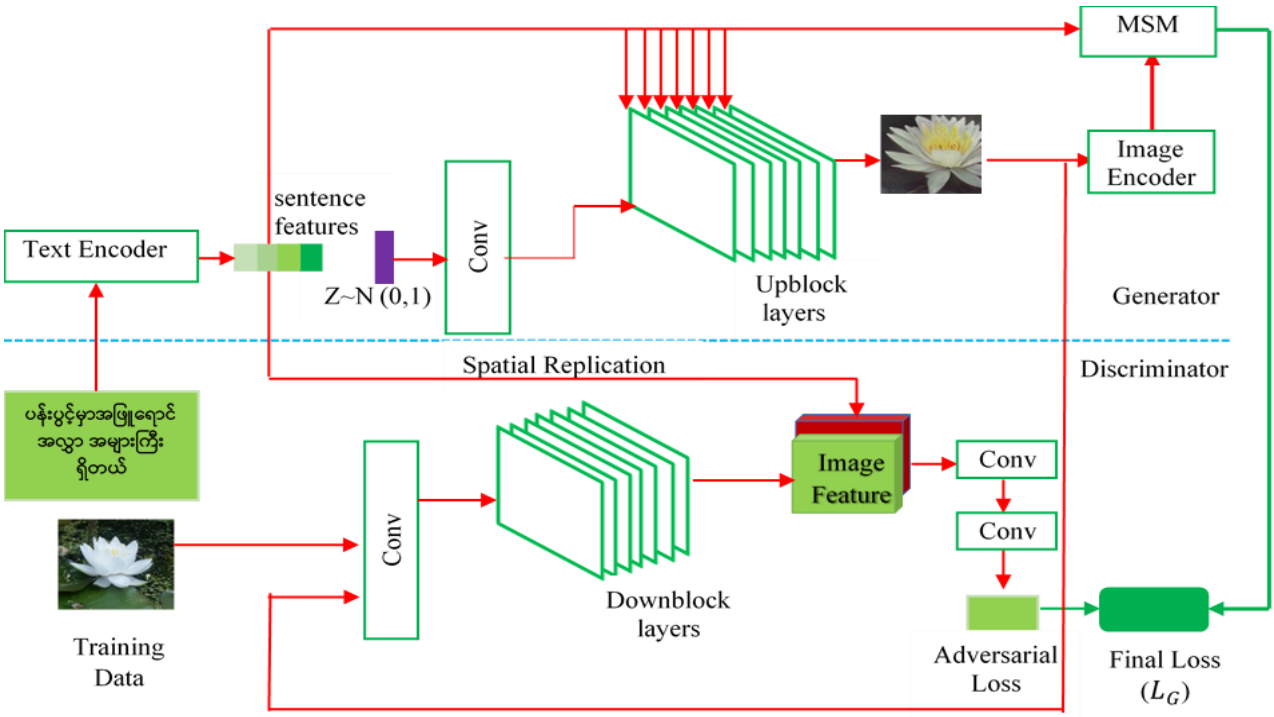

Figure. 2 Framework of DFGAN+MSM

to be more semantically understand the input language and to be able to generate high-quality images with more semantically correct with input text. There are three processes in this network (1) the image generation (2) evaluation 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, text encoder converts the textual input into sentence features. The random noise sampled by using Gaussian distribution is passed through 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 textual input while forwarding this noise to these layers. And then, the generator transforms the resulted image features into high resolution image (256x256).

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

In the last stage, the generated images and the encoded text input are fed as inputs to MSM model to compute the similarity loss between these two inputs. And then, the final loss of the generator is obtained by summing the adversarial loss and the similarity loss, and provides feedbacks to the generator. This feedback allows the generator to learn from its mistakes and gradually produce more and more realistic samples.

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

$$L_G = -\mathbb{E}_{G(z)\sim P_g}[\mathrm{D}(\mathrm{G}(\mathrm{z}),\mathrm{e})] + L_{MSM} \qquad (7)$$

5. Implementation details

This section contains about the two datasets, preprocessing steps of Myanmar corpus dataset, pretraining of MSM model, training details and evaluation metrics.

5.1 Dataset

To conduct the first implementation of the proposed model, we used CUB birds dataset [7]. This dataset contains 11,788 images and 200 different kinds of birds. Each image is paired with 10 text descriptions.

To implement Myanmar text to image synthesis for the proposed model, we conducted our experiments on Oxford-102 flowers dataset [8]. The dataset contains 102 categories, 8189 images. Firstly, the English captions corpus is 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 our Myanmar caption corpus is 40945.

Myanmar Input Text: ခရမ်းရောင်ပန်းပွင့်

English Translation: The purple flower

Segmented Sentence: ခရမ်းရောင် ပန်းပွင့်

(ခရမ်းရောင်= purple) (ပန်းပွင့်= The flower)

Myanmar sentence have built with sequence of characters and doesn't 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, we performed text encoding process to convert meaningful descriptions into number representation. We encoded Myanmar sentence 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.

5.2 Pretraining multimodal similarity model

After pre-processing, we pretrained MSM model 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 applied to the training of T2I to evaluate the similarity loss of the generated image from generator and text descriptions. Moreover, the text encoder is also used to extract the feature vectors of text descriptions. We use the pretrained model with the minimum loss as MSM during training of T2I.

5.3 Training DF-GAN+MSM

After training of MSM, we finally implemented Myanmar text to image synthesis. 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 analyze 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 in MSM is also used to extract features of the generated images to calculate the visual-semantic consistency loss. We stopped training of the system 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. We use Adam optimizer for learning, and the learning rate of generator and discriminator are 0.0001 and 0.004 respectively. The training batch size is 24.

Table 2. 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

5.4 Evaluation metrics

We used the two-evaluation metrics (1) Inception score and (2) Fréchet inception distance (FID) are used to evaluate the quality of the generated images from our proposed model.

5.4.1. Inception score

This metric [14] is used to evaluate the quality of the images generated from generative models. The probability distribution of each image is predicted using Inception v3 model. These probabilities are then summarized into scores to investigate how much each image look like a known class and how diverse set of images are across the known classes. The larger inception score presents the higher quality of images. The inception score is computed by using the following equation:

$$I = \exp\left(\mathbb{E}_{x\sim p_g} D_{KL}\left(p(y|x)||p(y)\right)\right) \quad (8)$$

5.4.2. Fréchet inception distance (FID)

It [15] is used to compare the probability distribution between the synthesized images from

generators and the real images in the training dataset. The smaller FID score represents the better quality of the synthesized images. The features map of these two data are extracted by using Inception v3 model. Then, the multivariate Gaussian distribution is built to predict distribution of the feature maps. The following equation is used to compute the FID score between the generated images, g and the realistic images, r.

$$FID = \|\mu_{r-}\mu_g\|^2 + T_r(\sum_r + \sum_g - 2(\sum_r \sum_g))^{1/2} \quad (9)$$

6. 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.

6.1 Quantitative analysis

We compare and analyse 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 3.

In multiple refinement stages [2-5, 9], DAE got lowest FID score of 15.19 but its IS score is lower than DMGAN. DM-GAN can generate diverse images with high IS score by incorporating some noises. These images are visually reality but not close to the real data distribution. This condition affects the FID score because this metric is used to measure the similarity between the generated images and real images. DF-GAN got the largest IS score of 5.10 compared with the state-of-the-art models, but its FID score is greater than TIME. This may be due to the fact that the image generated from DF-GAN doesn't capture the overall distribution of the real data. This leads to get higher FID score despite this model can generate the image with high diversity that reflects the IS score.

Compared with highest IS score and smallest FID score in multiple refinement work [2-5, 9], our proposed model can decrease FID scores from 15.19 to 13.05 and increases IS scores from 4.75 to 5.12. However, the proposed model got smallest FID score of 13.05 and highest IS score of 5.12 compared with state-of-the-art model. 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. Meanwhile, it can prove that the proposed model can advance the quality of images in terms of reality and diversity.

Table 3. Comparison of FID and inception scores of our proposed model with others on CUB dataset

Models	IS↑	FID↓
StackGAN [2]	3.70	-
StackGAN++ [3]	3.84	-
AttnGAN [4]	4.36	23.98
DM-GAN [5]	4.75	16.09
DAE-GAN [9]	4.42	15.19
TIME [10]	4.91	14.30
DF-GAN [6]	5.10	14.81
DF-GAN+MSM(Our)	**5.12**	**13.05**

6.2 Qualitative analysis

Compared with the baseline model, 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. We 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 Fig. 3. 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 images generated from our proposed model is more accuracy in shape and better realistic than the baseline model. The image generated from the fourth input sentence by our 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 we noted in this result, the images generated by our proposed model achieves better performance than the baseline model in terms of reality and semantically consistence.

7. Experiment results and analysis on Myanmar T2I

In this section, quantitative and qualitative evaluations are done to identify the proposed model gives improvements on the quality of the generated images conditioned on Myanmar text descriptions. Myanmar T2I is implemented by the proposed model using Oxford 102 flowers annotated in Myanmar to prove the proposed model get the progressive result on synthesising images on different dataset and different languages. Moreover, the quantitative results of DFGAN+MSM (proposed model) are

Figure. 3 The images generated from DF-GAN and proposed model implemented on CUB Birds dataset

compared to previous results of baseline DF-GAN and AttnGAN based Myanmar T2I published in [17] and DCGAN based Myanmar T2I published in [16]. Moreover, the results of qualitative evaluation are also compared with baseline DF-GAN. The same training data and testing data are used to implement all of the systems.

7.1 Quantitative analysis

In this section, we compared and analysed the inception score and FID score of the generated images from Myanmar sentences based on our model with these cores of baseline DF-GAN and AttnGAN [17]. 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 Figs. 4 and 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, we stopped training of the AttnGAN for Myanmar T2I 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

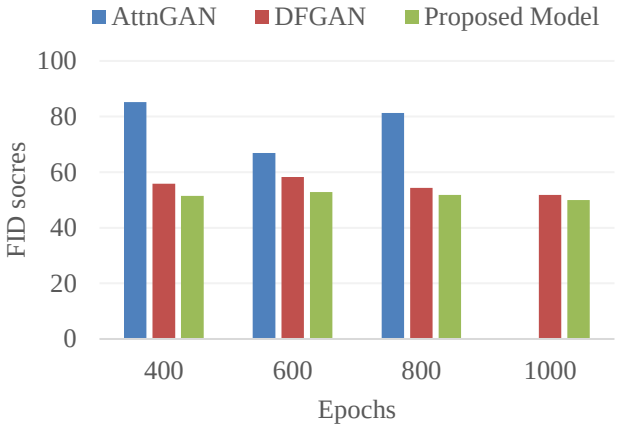

Figure. 4 Comparison of FID scores of AttnGAN, DF-GAN, and our proposed model

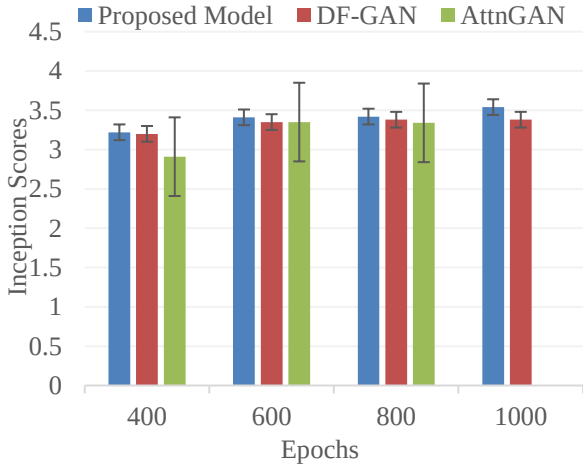

Figure. 5 Comparison of inception scores of AttnGAN, DF-GAN and our proposed model

Table 4. Performance comparison of FID and inception scores of our proposed model with others

Model	Inception Score ↑	FID score ↓
DCGAN [16]	1.72 ± 0.01	222.34
AttnGAN [17]	3.35 ± 0.03	66.92
DF-GAN [17]	3.38 ± 0.03	51.86
Proposed Model	**3.54 ± 0.03**	**49.97**

dataset are also shown in Table 4. 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, our proposed model advanced the quality of the generated images from Myanmar sentence.

7.2 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 Myanmar text descriptions with AttnGAN, the base-line DF-GAN and the proposed method are shown in Fig. 7.

As we noted in this figure, the images generated from our proposed model is more precise in shape and brightness in colour than 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 doesn't illustrate with half and half of red and white colour. DF-GAN and AttnGAN can't 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, our 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, our model can generate the images with better semantic consistency and brightness in colour than the other models for overall the generated

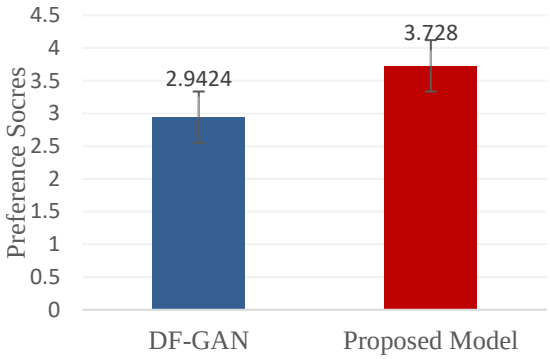

Figure. 6 Comparison of preference scores between DF-GAN and our proposed model

images. Therefore, our proposed model outperforms the other models for Myanmar T2I.

The overall subjective evaluation of the generated images was only 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 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).

The evaluated preference scores of DF-GAN and the proposed model are shown in Fig. 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 progressive results 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. Therefore, DFGAN+MSM can generate high quality image with better semantic consistency for Myanmar T2I.

8. Conclusion and future works

In this work, we examined the improvement results in image generation process due to applying multimodal similarity model to the generator side using two different datasets annotated on different languages. Firstly, we made experiments English T2I using UCSD-Caltech Birds dataset. The comparisons are done on the proposed model and state-of-the-art models, StackGAN, StackGAN++, AttnGAN, DAE-GAN, and TIME. According to the comparative results, the proposed model increased the quality images compared with the baseline model. Secondly, the performance of our proposed model was proved

Figure. 7 The images generated from DF-GAN, AttnGAN and our proposed model implemented on Oxford 102 flowers dataset

by implementing Myanmar T2I using Oxford-102 flowers dataset. In this work, we manually prepared Myanmar captions for each image (about 40945 sentences) in Oxford 102 flowers dataset to implement the first Myanmar T2I. The quality of the generated images is compared with DCGAN, AttnGAN, and the baseline DF-GAN based Myanmar T2I. As a result, the proposed model outperforms all of the model-based Myanmar T2I. Therefore, the proposed model enhanced the quality of images in both English T2I and Myanmar T2I. The MSM model is only trained on sentence-level therefore that lacks of fine-grained details at the word-level. In the future, MSM that is pretrained on both word-level and sentence-level will be added to the generator to investigate the progressive results on synthesizing images.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

The paper background work, methodology,

dataset collection and preprocessing, system implementation, visualization, the comparative and analysis, the preparation and editing draft have been done by the first author. The supervision, system review, and project administration have done by the second author.

References

[1] I. Goodfellow, J. P. Abadie, M. Mirza, B. Bing, D. W. Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets", In: *Proc. of Advances in Neural Information Processing Systems*, Montreal, Quebec, Canada, pp. 2672-2680, 2014.

[2] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks", In: *Proc*. of *IEEE International Conf. on ComputerVision (ICCV)*, Venice, Italy, pp. 5908-5916, 2017.

[3] 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*, Vol. 41, No. 8, pp. 1947-1962, 2019.

[4] 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 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 1316-1324, 2018.

[5] 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 Conf. on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 5795-5803, 2019.

[6] M. Tao, H. Tang, F. Wu, X. Jing, B. Bao, and C. Xu, "DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis", In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, pp. 16494-16504, 2022.

[7] 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, 2011.

[8] M. Nilsback and A. Zisserman, "Automated Flower Classification over a Large Number of Classes", In: *Proc. of IEEE Sixth Indian Conf. on Computer Vision, Graphics & Image Processing*, Bhubaneswar, India, pp. 722-729, 2008.

[9] 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.

[10] 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.

[11] M. Schuster and K. Paliwal, "Bidirectional Recurrent Neural Networks", *IEEE Transactions on Signal Processing*, Vol. 45, No. 11, pp. 2673-2681, 1997.

[12] S. Albawi, T. A. Mohammed, and S. A. Zawi, "Understanding of a Convolutional Neural Network", In: *Proc. of International Conf. on Engineering and Technology (ICET)*, Antalya, Turkey, pp. 1-6, 2017.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision", In: *Proc. of IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 2818-2826, 2016.

[14] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved Techniques for Training GANs", In: *Proc. of Advances in Neural Information Processing Systems*, Barcelona, Spain, pp. 2234–2242, 2016.

[15] 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", In: *Proc. of Advances in Neural Information Processing Systems*, Long Beach, CA, USA, pp. 6626–6637, 2017.

[16] 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.

[17] 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.