

# Generative Adversarial Networks for Myanmar Text to Image Synthesis

Nang Kham Htwe  
Natural Language Processing Lab  
University of Computer Studies, Yangon  
Yangon, Myanmar  
nangkhamhtwe@ucsy.edu.mm

Win Pa Pa  
Natural Language Processing Lab  
University of Computer Studies, Yangon  
Yangon, Myanmar  
winpapa@ucsy.edu.mm

**Abstract**— Generating synthesized images like realistic images from text descriptions has become popular in many application areas in combination of computer vision and natural language processing. With the introduction of (GANs), text-to-image synthesis has gained new heights of success and received much attention. Considering that, we build the first Myanmar text to image synthesis. Myanmar captions for Oxford-102 flowers dataset is manually prepared in this work. We conducted our experiments by using this annotated image dataset. In this study, we used two methods to synthesize high resolution images from Myanmar text: multiple-refinement stages of generators and generators with only one stage backbone. We compared and analyzed the quality of generated images from these two algorithms. The quality of generated images is evaluated and compared in terms of two evaluation metrics: Inception score and Frechet Inception Distance (FID))

**Keywords**—Generative Adversarial Networks, Inceptions Score, Frechet Inception Distance

## I. INTRODUCTION

The last few years, GANs (Generative Adversarial Networks) has remarkable progress in image and video generation. Among them, generation of synthesized images from text has become one the active research areas in computer vision and natural language processing communities. The first T2I methods enables to generate the image (64 x 64 dimension) conditioned on the whole sentence vector. In order to enhance high-resolution images, latter GANs [1] proposed multiple stages of generators and discriminators. These methods have effective in generating high resolution images. However, the use of multiple refinement stages brings unstable training process and higher computation.

To address this issue, the one-stage T2I back bone is introduced in [2] which contains one pair of discriminator and generator. However, the sentence vector is conditioned on every UPBlocks and used only one stage of generator to get high-resolution image with dimension of 256x256. Therefore, there have many remarkable progresses in T2I for English [1,2]. From this point of view, we also want to improve our language in this research areas. However, annotated image dataset with Myanmar language is not available to implement this research. Therefore, we manually constructed Myanmar Captions corpus for Oxford-102 flowers dataset [3] and implement the first Myanmar text to image synthesis.

In addition, we also want to analyze Myanmar caption corpus that has good enough quality or not to implement Myanmar text to image synthesis. However, there are many varieties of GAN to implement this study. For this reason, we have investigated the impact on our caption corpus in two areas. The first study is to evaluate the effectiveness of

attention with multiple refinement stages and the second one is fusing of text and image at every blocks. Therefore, we made our experiments by using AttnGAN and DF-GAN. And then we made evaluation and compared the quality of the generated images to picture which model is the best in implementing of Myanmar text to image synthesis.

## II. METHODOLOGY

This section contains about the methods applied in Myanmar text to image synthesis: (1) Attentional Generative Adversarial Networks (2) Deep Fusion Generative Adversarial Networks.

### C. Attentional Generative Adversarial Networks

The attentional generative adversarial network [1] can generate the images conditioned on both global sentence vectors and word vectors that are relevant to each sub-region of the images. 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 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:

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

Here,  $L_G$  is the sum of all losses of the generators and each generator has a corresponding discriminator.  $L_{DAMSM}$  is the loss from DAMSM model and this loss is used to measure the visual-semantic similarity. This model [6] contains two neural networks: text encoder (bi-directional Long Short-Term Memory) and image encoder (convolutional neural network).

### B. Deep Fusion Generative Adversarial Networks

In this model [2], there is only one discriminator and generator. The text descriptions are encoded by a pretrained encoder similar to Attention GAN. First, the noise vector is fed to Fully-connected layer. Then, the output is passed to a series of UpBlocks. 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. And the adversarial loss is calculated to evaluate the visual-semantic consistency. The whole formulation of loss function generator is:

$$L_G = -\mathbb{E}_{G(z) \sim p_g} [D(G(z), e)] \quad (2)$$

### III. IMPLEMENTATION DETIALS

#### C. Training of Myanmar Text to Image Synthesis

This section contains implementation of Myanmar text to image synthesis on two models. We manually constructed 5 Myanmar captions corpus [4] for each image (total of 8189 images) by focusing their features without directly using or translating English descriptions from the Oxford-102 flowers dataset [3] because the quality of translated sentence from machine translation is not accurate to use in this implementation. In this experiment, 7789 images are used for training while the remaining 400 images for testing. We pretrained DAMSM model that contains text encoder and image encoder. We embed Myanmar sentence by using bi-LSTM text encoder. We used this model to compute text and image similarity level during the training stages of AttnGAN. In AttnGAN, we generate the images conditioned on text using multiple refinement stages. The dimension of images at each stage of generator are 64x64, 128x 128 and 256x256 respectively. In DFGAN, we obtained the sentence features by using pretrained text encoder in AttnGAN. In this model, we generate the image (256x256 dimension) with only one stage backbone. We trained these two models 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, we compared these two models by using the best epochs of each model instead of using the same epoch.

#### D. Evaluation Metrics

**Inception Score** [5] is used to measure the quality of the generated images from the generative models. The class probabilities for each generated image are predicted using Inception v3 model. The larger inception score represents the higher quality of the generated images.

**Fréchet Inception Distance (FID)** [5] compares the distribution of generated images with the distribution of real images. The smaller FID score means the better quality of generated images.

#### E. Experiment Results and Discussion


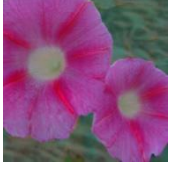


The two-evaluation metrics are used to evaluate quantitative results for generated images from Myanmar text descriptions. We computed inception scores and FID score of the generated images based on testing dataset. The comparisons score of these two GANs are shown in Table 1. In quantitative evaluation, DFGAN got higher scores on inception and lower FID scores than AttnGAN.

The qualitative evaluation is made by querying text descriptions to synthesize image. As shown in Figure [1], we compared the images generated by two methods with each text description and we select the best sample for comparison. 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. 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.

**Table 1.** Inception score and FID score of two models evaluate based on test data

Model	Inception Score	FID score
DCGAN[4]	<b>1.72 <math>\pm</math> 0.01</b>	<b>222.34</b>
AttnGAN	<b>3.35 <math>\pm</math> 0.03</b>	<b>66.92</b>
DFGAN	<b>3.38 <math>\pm</math> 0.03</b>	<b>51.86</b>

Text Descriptions	AttnGAN	DF-GAN
ခရမ်းရောင်ပွင့်ချပ်နှင့်ပန်းပွင့် တွင်အနီရောင်အမှတ်အသား များရှိတယ် <b>English:</b> The flower with purple petals has the red shade.		
ပန်းပွင့်တွင်အဖြူရောင်အဆင်းရှိ သောပွင့်ချပ်များနှင့်အဝါရောင် စင်တာရှိတယ် <b>English:</b> The flower has white color petals and yellow center.		

**Fig. 1.** The images generated from Myanmar Text descriptions

### CONCLUSION

In this paper, Myanmar captions are described for each image in Oxford-102 flowers dataset. The images are synthesized from Myanmar text using multiple-stages of generators and one-stage of generator. A comparative study has done on two models using qualitative and quantitative evaluations. According to experimental results, DF-GAN is better than AttnGAN for Myanmar to image synthesis.

### REFERENCES

- Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X. and He, X.: AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1316-1324 (2018)
- Tao, M., Tang, H., Wu, F., Jing, X.Y., Bao, B.K. and Xu, C.: DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16515-16525 (2022)
- Nilsback, M.E. and Zisserman, A.: Automated flower classification over a large number of classes. In: *IEEE Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp.722-729 (2008)
- Nang Kham Htwe, Win Pa Pa.: Building annotated image dataset for Myanmar text to image synthesis. In: *Proceedings of International Conference on Computer Application* (2021)
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. and Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: *Advances in neural information processing systems* (2017)