다중 스케일 그라디언트 조건부 적대적 생성 신경망을 활용한 문장 기반 영상 생성 기법

Nguyen P. Bui*, Duc-Tai Le**, 추현승**
*성균관대학교 수퍼인텔리젼스학과
**성균관대학교 소프트웨어대학
email : *phuocnguyen@skku.edu, **{ldtai, choo}@skku.edu

Text-to-Face Generation Using Multi-Scale Gradients Conditional Generative Adversarial Networks

Nguyen P. Bui*, Duc-Tai Le**, Hyunseung Choo**
*Dept. of Superintelligence, Sungkyunkwan University
**College of Computing. Sungkyunkwan University

Abstract

While Generative Adversarial Networks (GANs) have seen huge success in image synthesis tasks, synthesizing high-quality images from text descriptions is a challenging problem in computer vision. This paper proposes a method named Text-to-Face Generation Using Multi-Scale Gradients for Conditional Generative Adversarial Networks (T2F-MSGGANs) that combines GANs and a natural language processing model to create human faces has features found in the input text. The proposed method addresses two problems of GANs: model collapse and training instability by investigating how gradients at multiple scales can be used to generate high-resolution images. We show that T2F-MSGGANs converge stably and generate good-quality images.

1. Introduction

Since GANs's introduction by Goodfellow et al. [1], they have seen huge success in image synthesis tasks. The success of GANs comes from the fact that they do not require manually designed loss functions for optimization, and can therefore learn to generate complex data distributions without the need to be able to explicitly define them. Generative Adversarial Nets [1] consist of two 'adversarial' models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. Both G and D could be a non-linear mapping function, such as a multi-layer perceptron. Both G and D are trained simultaneously: we adjust parameters for G to minimize $\log(1-D(G(z))$ and adjust parameters for D to minimize $\log D(X)$, as if they are following the two-player min-max game with value function $V(G,D)$:

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

Conditional GANs (cGANs) [2] is an extended version of GANs if both the generator and discriminator are conditioned on some extra information y, y could be any kind of auxiliary information, such as class labels or data from other modalities. We can perform the conditioning by feeding y into both the discriminator and generator as an additional input layer. The objective function of a two-player min-max game would be:

$$\min_G \max_D V(D,G) = E_{x \sim p_{data}(x)}[\log D(x|y)] + E_{x \sim p_z(z)}[\log(1-D(G(z|y)))]$$

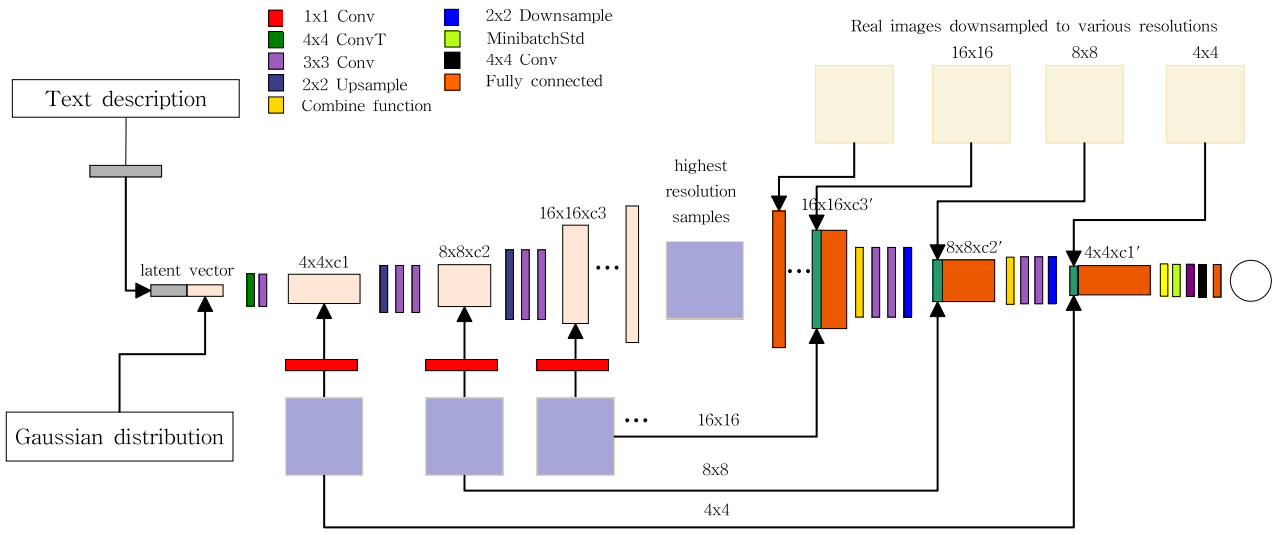

Figure 1. Illustration of our proposed T2F-MSGGANs

2. Paper Structure

In this part, we summarize our paper as follows: In Part 3, we focus on our proposed method, T2F-MSGGANs, and explain our algorithm. In Part 4, we describe the dataset, implementation details, and experimental results.

3. Text-to-Face Generation Using Multi-Scale Gradients for cGANs (T2F-MSGGANs)

3.1. Multi-Scale Gradients GANs (MSGGANs)

Training instability a fundamental issue with GANs has been widely reported by previous works [6, 7, 8]. MSGGANs [4] address this issue by investigating how gradients at multiple scales can be utilized to produce high-resolution images (typically more challenging due to the data dimensionality) without relying on previous greedy approaches. MSGGANs allow the discriminator to look at not only the final output (highest resolution) of the generator but also at the outputs of the intermediate layers (Figure. 1). As a result, the discriminator becomes a function of multiple scale outputs of the generator and importantly, passes gradients to all the scales simultaneously.

3.2. Text Embedding

In this work, we utilize a Facebook's text embedding model named Infersent2. Infersent2 is a method of sentence embedding that provides semantic representations of English sentences. It is trained on natural language sets and generalizes well for many different tasks. Therefore, it is widely used in many fields of embedding sentences.

Depending on the intended use, Infersent2 can be configured to produce a vector with different dimensions. The output of Infersent2 is a vector representing the level of importance of words containing information in that sentence. In this work, we configure Infersent2 output dimension is 4096.

3.3. Algorithm

Given a pair input including a text description input and an image. Firstly, we utilize the Infersent2 model to generate a text embedding vector. After that, we combine a text embedding vector with a Gaussian noise vector to produce input for our T2F-MSGGANs architecture.

Based on Conditional MSGGANs, we can configure resolution for output images based on hardware. In this work, we produce the highest resolution at 128x128 pixels (see Figure 1).

4. Experimental Results

4.1. Dataset and implementation details

In this work, we use a public dataset CelebFaces Attributes Dataset (CelebA) [9], CelebA is a large-scale face attributes dataset

Figure 2. Experimental results

The chubby man has a double chin. He has a receding hairline. He has big nose, narrow eyes and a slightly open mouth. The man is smiling. He is wearing a necktie.

The woman has oval face and high cheekbones. Her hair is blond in colour with bangs. She has pointy nose and a slightly open mouth. The smiling, young attractive woman has heavy makeup. She is wearing lipstick.

The man has oval face and high cheekbones. He has wavy hair which is brown in colour. He has a slightly open mouth. The young attractive man is smiling.

Figure 3. Testing trained model with unknown text description

with more than 200K celebrity images, each with 40 attribute annotations. The images in this dataset cover large pose variations and background clutter.

We develop the proposed model based on the Pytorch framework. We train our model from scratch without using any pre-trained model weights. For optimization, we use Adam [5] optimizer (beta 1 : 0, beta 2 : 0.99).

4.2. Results

In training progress, we print intermediate outputs and arrange them in Figure 2 for better visualization. In the horizontal axis, images are arranged following to the epoch. In the vertical axis, images are arranged from low-resolution to high-resolution (top to bottom).

For testing with unknown input text descriptions, we use a trained model to test and arrange generated images in Figure 3.

Acknowledgments

This research was supported by Korea government(MSIT,IITP), under the ICT Creative Consilience program(IITP-2021-2020-0-01821) and National Research Foundation of Korea (NRF-2020R1A2C2008447, Deep Adversarial Learning Driven Virtual Edge: Self-supervised virtual edge mobility, resource placement and allocation)“ This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program(IITP-2021-2015-0-00742) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation)”.

References

[1] Goodfellow, Ian, et al. "Generative adversarial networks." Communications of the ACM 63.11 (2020): 139-144.

[2] Mirza, Mehdi, and Simon Osindero. "Conditional generative adversarial nets." arXiv preprint arXiv:1411.1784 (2014).

[3] Conneau, Alexis, et al. "Supervised learning of universal sentence representations from natural language inference data." arXiv preprint arXiv:1705.02364 (2017).

[4] Karnewar, Animesh, and Oliver Wang. "Msg-gan: Multi-scale gradients for generative

adversarial networks." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[5] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).
[6] Salimans, Tim, et al. "Improved techniques for training gans." Advances in neural information processing systems 29 (2016): 2234-2242.
[7] Mao, Xudong, et al. "Multi-class generative adversarial networks with the L2 loss function." arXiv preprint arXiv:1611.04076 5 (2016): 1057-7149.
[8] Arjovsky, Martin, Soumith Chintala, and Léon Bottou. "Wasserstein generative adversarial networks." International conference on machine learning. PMLR, 2017.
[9] Liu, Ziwei, et al. "Large-scale celebfaces attributes (celeba) dataset." Retrieved August 15.2018 (2018): 11.