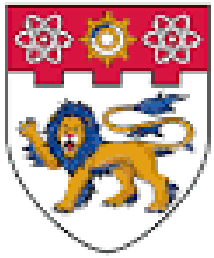

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Master in Science (Artificial Intelligence)

(MSAI)

AI-6121

Direct Reading and Literature Review

Name of Student:
Teo Lim Fong

Matriculation No: G2101964G

Table of content

Contents

Introduction

This work is about image-to-image translation.

Image-to-image translation, also known Generative Adversarial Net (GAN), have been invented since 2004 by Ian Goodfellow **[1]** to generate more fake samples or synthetic datasets. To present my understanding, I will briefly explain the keys factors of GAN. GAN consists of two models, generative and discriminative shown in Figure 1. For generative model, the input noise(z) will be mapped into $G(z;\theta_g)$ data space to learn the generative distribution while for discriminative model, it is represented as $D(x;\theta_d)$ where x is the original samples. In other words, discriminative model is to evaluate the authenticity produced from the generative model. The model will keep training until the discriminative model approved the generative distribution p(z) to be similar as the original sample x.

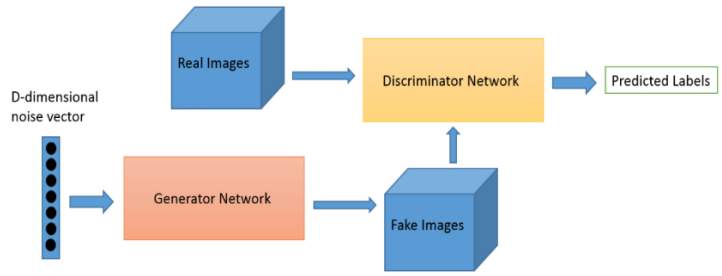

Figure 1: Example of Generative Adversarial Network (GAN) architecture

Problems

One of the problems that computer vision is facing is lack of datasets. In traditional computer vision, it uses different type of data augmentation such as flipping vertically and horizontally, rotating different angle, adjust the color saturation, random cropping and many more. However, with such augmentation, it is more or less the same distribution. Neural network needs a new set of datasets that never seen before. For example, we have a lot of front view car images but we are lacking the side view of the car. Segmentation in this case does not work. No matter how you segmented the data, it is still the front view of the car. For possible solution, we can use pix2pix GAN **[2]** paired training datasets to generate different domain. We will discuss this in the next section.

Also in fashion industry, it is expensive to hire a model just to take picture of different color of the dress. To minimize the cost, is it possible to change the color of the dress from one color to another color that we want? CycleGAN is possible to generate two separate domains. We will discuss this in the next section.

They are many different types of GAN for different purposes. Mainly, I will be concentrating on pix2pix GAN and CycleGAN.

Type of GAN

1) Pix2pix GAN

Pix2pix GAN [2] are capable of mapping the input image with output image. Discriminator model is trained to distinguish between fake sample G(x), produced by generative model, and the real image (y) while the generative model is trained to trick the discriminator shown in Figure 2. Unlike GAN, sample x will be used in both generative and discriminator.

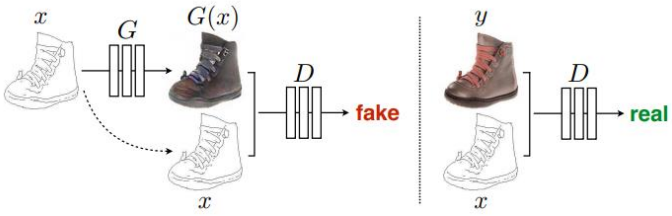

Figure 2: Example of Pix2pix GAN architecture

Previously in GAN, input noise (z) is used to learn a mapping with $G(z;\theta_g)$. According to the article, the author tried training a net without input noise (z). It turned out that sample x is still able to learn a mapping with sample y. However, the result is deterministic. Next, the author tried adding Gaussian noise (z) into generator but the generator seems to ignore the noise. The final solution the author tried is by applying dropout noise to several layers in the generator which only resulted minor stochasticity in the output.

Pix2pix GAN required a pair image that consist of domain x and y, where x and y must be in the same geometric but different domain/style. It is useful when you have insufficient datasets of another domain. Let's say we have a few of sketches of a car in side view but is not in RGB shown in Figure 3. How can we convert this sketch into our datasets?

Figure 3: Example of car testing datasets for Pix2pix GAN

First, we can consider Figure 3 as our test set. Earlier on we mention that we are lacking datasets of a car that is side view. This mean that we can use python programing to convert existing RGB side view into sketch then pair it with the RGB datasets to fulfill the pix2pix GAN requirement. To make my explanation clearly, I demonstrate a diagram shown in Figure 4.

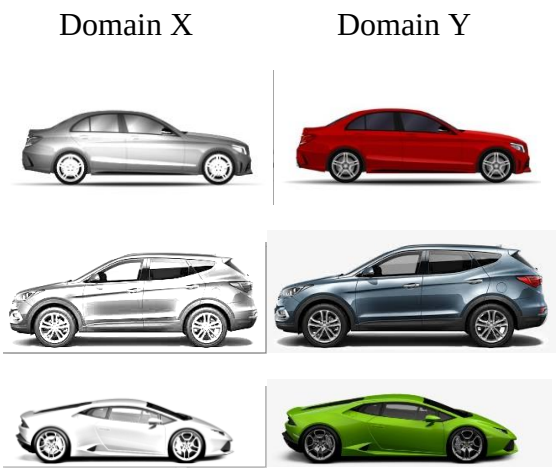

Figure 4: Example of car training datasets for Pix2pix GAN

Domain Y is the existing datasets that we have. Domain X is the sketch produced by using python code. We need to combine these two images together. The final output of the new image will be (256 x 512), we assumed the original image is 256 x 256. As such, this will be our training datasets. After training, we will apply the trained model to the test datasets in Figure 3 to convert it into RGB. Thus, we will have more datasets.

However, for both images to combine together, both images required same amount of channel. Grayscale contains 1 channel while RGB contains 3. We can convert grayscale into RGB using the open cv library. This function is not converting gray to RGB but instead stacking the channel value into 3.

However, there is a limitation for this method. Both domains must be in the same geometric. In reality, it is very hard to find same geometric for both domains. Hence this may lead to lack of training datasets for pix2pix GAN. However, the author has further implemented unpaired datasets which means we can train with two separated domains without any restriction. This implementation is called CycleGAN.

2) CycleGAN

As mention earlier, is it possible to generate different color clothing from GAN? For example, in Figure 5, given such training datasets, is it possible for Pix2pix GAN to generate black to red dress? The answer is no. These training datasets is not paired. However, CycleGAN are capable of generating it.

Figure 5: Example of dress training datasets for CycleGAN

CycleGAN as shown in Figure 6 consists of two Generators and two Discriminators. In CycleGAN, unpaired training dataset is used for training, From the article **[3]**, unpaired training dataset is proven to be better than paired training dataset as it does not need any information from Generated X. On the other hand, paired dataset is expensive and has limited resources such as semantic segmentation

The purpose of having two domains in the network is to successfully learn a mapping using the given training datasets. For example, Generated X (black dress) will learn a mapping function from G :X → Y (Generator X to Y) to become indistinguishable from Generated Y (red dress). Adversarial loss is applied to both mode G :X → Y and its Discriminator (D_Y) to calculate the minimum and maximum value.

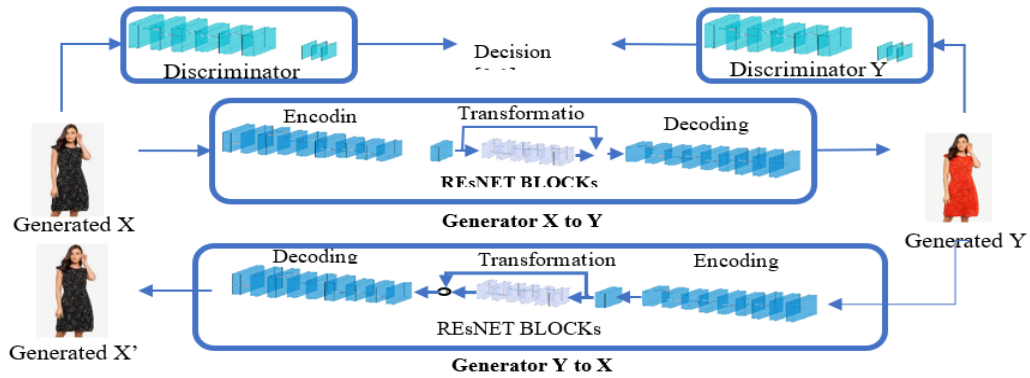

Figure 6: Example of CycleGAN architecture*

CycleGAN are not only are capable of generating two different domains but also able to remove noise in the image such as dehazing and de-rain. Since the generated output does not have any ground truth, how are we going to evaluate the quality of the image? In the next section, I will be explaining some of the evaluators for CycleGAN.

Image Quality Comparison

Image quality comparison is important in image enhancement as it determines the quality of the image. For CycleGAN, we used quantitative comparison and qualitative analysis. In machine learning, quantitative comparison refers to a comparison with a ground truth as shown in Figure 7 while qualitative analysis does not require a ground truth. Quantitative comparison is used in training data set while qualitative analysis can be used in both training and testing data set. Example of quantitative technique that can be used are the following: SSIM (Structural Similarity Index) and qualitative analysis is NIQE (Naturalness Image Quality Evaluator) **[4]**

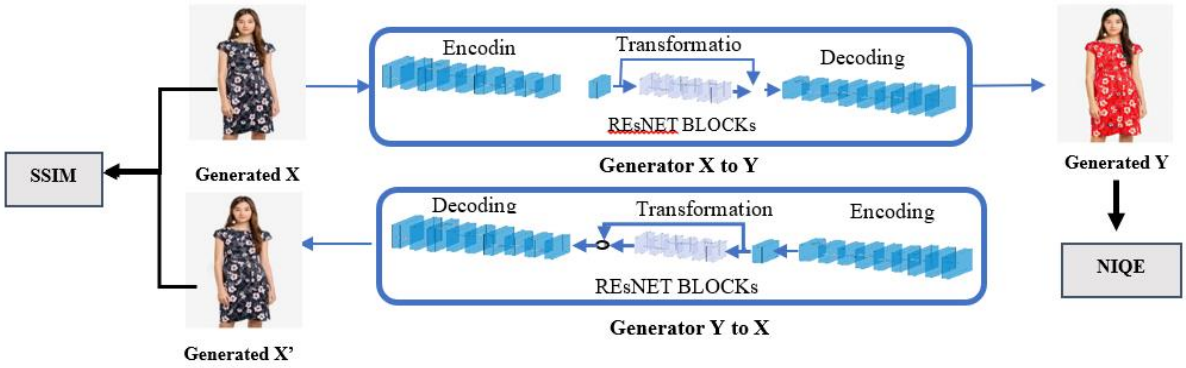

Figure 7: Using SSIM and NIQE metrics to evaluate the result

1) **SSIM**

SSIM is a measurement of a ground truth with a processed image (Generated X). It measured the similarity between the ground truth and the processed image with a range from 0 to 1, where 1 is the ideal quality measurement. SSIM is used in training dataset (ground truth) and the Generated X' (Reconstructed X) to compare how well the network perform.

2) **NIQE**

NIQE is the qualitative analysis which also known as the 'completely blind' image quality assessment. It does not need any dependency to predict the quality of the image. It is mainly used on the testing images as it does not have any ground truth to compared with.

Discussion and Limitation

Every algorithm has its limitation. For CycleGAN, the training distribution between domain X and Y cannot have any extreme changes, especially huge geometric changes. For example, domain X contains dog and domain Y contain crocodile. Other than that, each domain should clearly identify the distribution rather than mixing with multiple distribution. If not, it will fail. A good example of such failure cases is the famous 'zebra human' that translate human riding on a horse into a zebra. Their training datasets does not include any human image in the distribution. Thus, causing the model to translate incorrectly.

All the limitation that I just mentioned can be considered as a future improvement.

Conclusion

In computer vision, generating more synthetic dataset is critical. It is not only helping to enhance the prediction but also increase different variant of the object in different environment. Nevertheless, image-to-image translation is just a method to generate more datasets for real life application such as detection of an object. Training a robust detection model is not easy. There are many factors that could interfere the detection such as lighting, occlusion and many more. Therefore, experts keep implementing novel algorithms to encounter insufficient datasets issue.

Perhaps, 3D GAN maybe be more impactful. 3D contains much more information than 2D such as the front view, slide view, top view, etc.

Reference

[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza. Generative Adversarial Nets. 2014

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou. Image-to-Image Translation with Conditional Adversarial Networks. 2017

[3] Jun-Yan Zhu, Taesung Park, Phillip Isola. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. 2017

[4] Anish Mittal, Rajiv Soundararajan and Alan C. Bovik. Making a 'Completely Blind' Image Quality Analyzer. IEEE 2012