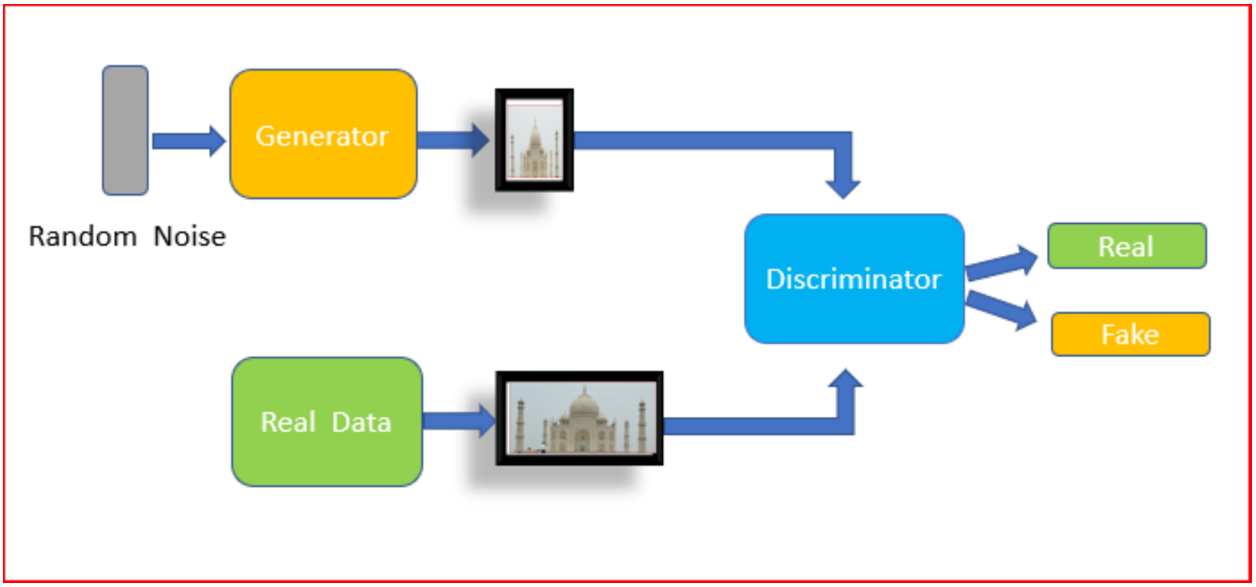
Machine Learning Nanodegree
Capstone Project

By Chingis Oinar

September 4th, 2021

Introduction

Creating human-friendly maps is a very tedious process yet it is one of the most important sources of curated data. Moreover, considering the staggering attention received by smart or even autonomous vehicles, maps have a commercial value to a number of huge companies, including ride-sharing services like Uber, Lyft, vehicle manufacturers like Tesla, or even national security agencies like CIA, NSA and FBI. There is no doubt that the interaction between customers and interfaces play a crucial role in companies' success. Therefore, constructing accurate as well as human-readable maps has been a major concern for companies producing "smart devices". However, an accurate map must also react to all the changes occurring in the ground, which is especially vital for companies producing autonomous vehicles. Thus, the generation of human-readable maps becomes an even more complicated and time-consuming task. One way to tackle the issue is to automate the process of map generation from satellite images. Fortunately, Generative Adversarial Networks (GANs) have recently achieved impressive results in the field. Specifically, by incorporating two separate networks, generator and discriminator, GAN learns a loss function due to which it is able to produce highly realistic images.



The figure attached above demonstrates how GAN works. As seen, Generator's objective is to generate data that is indistinguishable from the real data, whereas

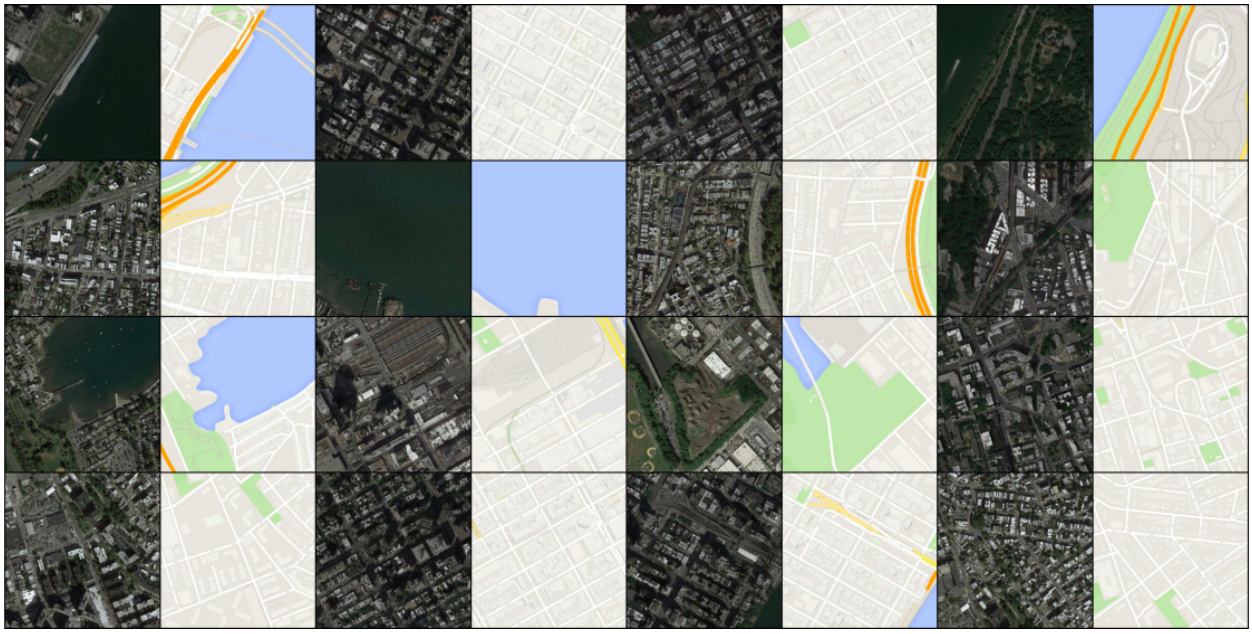
the Discriminator takes both real and generated data and tries to classify them correctly.

Task Definition

In this capstone project, I tackle a satellite to map image translation problem using GAN. For the chosen task, I will be using a Pix2Pix GAN that is proposed by Philip Isola, et al. in the paper titled “Image-to-Image Translation with Conditional Adversarial Networks” and presented at CVPR in 2017. The Pix2Pix model is a type of conditional GAN where the generator takes a source image as a conditional input. The discriminator is provided with both a source and the target/generated as input and predicts whether it’s fake or real. Thus, inputs are aerial photos obtained from Google Maps and the goal is to convert them into user-friendly Google maps format. Previous approaches have observed that it is beneficial to mix the GAN objective with a more conventional loss, such as L2 distance. However, the authors use L1 distance instead as they claim it produces less blurry images. Some of the other related works include Style Loss as well. Therefore, I present a comparison and show how different objective functions affect image quality.

Dataset and Inputs

The dataset is provided on the pix2pix website, hence it is publicly available and can be easily accessed. The train set contains 1,097 images of satellite images of New York and the corresponding Google maps pages, whereas the validation dataset had 1,099 images.



Each image is 1,200 pixels wide and 600 pixels tall. An example of the images in the dataset is attached below, where the target and the source images are shown side by side.

Technical Approach

Objective

There were a number of objective functions tested yet most of them contain a traditional conditional GAN's objective, which is expressed as follows:

$$\mathcal{L}_{cGAN}(G, D) = \mathbf{E}_{x,y}[(x, y)] + \mathbf{E}_{x,z}[\log(1 - D(x, G(x, z)))]$$

Where G and D are for generator and discriminator respectively, whereas x, y and z are for an image, target and noise vector respectively. As some previous works in

the field observed, adding an additional objective for a generator such as L2 or L1 distance boosts the performance and results in less blurry images. Therefore, the experiments involved such objectives as L2, L1 and even Style Loss, which was first introduced by Leon A. Gatys, et al. in the paper titled "A Neural Algorithm of Artistic Style", on top of conditional GAN. To summarize, the combinations of objectives used for the experiments are as follows:

Baseline:

$$G = argmin_G max_D \mathcal{L}_{cGAN}(G, D)$$

Experiment #1:

$$G = argmin_G max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L_1}(G)$$

This framework is introduced by Pix2Pix GAN and the authors demonstrate that L1 distance helps achieve high quality, especially less blurry, images. It was tested within a set of different image-to-image translation tasks showing a significant boost in the image quality.

Experiment #2:

$$G = argmin_G max_D \mathcal{L}_{L_2}(G, D) + \lambda \mathcal{L}_{L_1}(G)$$

This framework was introduced by Xudong Mao, et al. in the paper titled "Least Squares Generative Adversarial Networks" (LSGAN). They argue and demonstrate that it is able to generate higher quality images than regular GANs and it results in a more stable learning process.

Experiment #3:

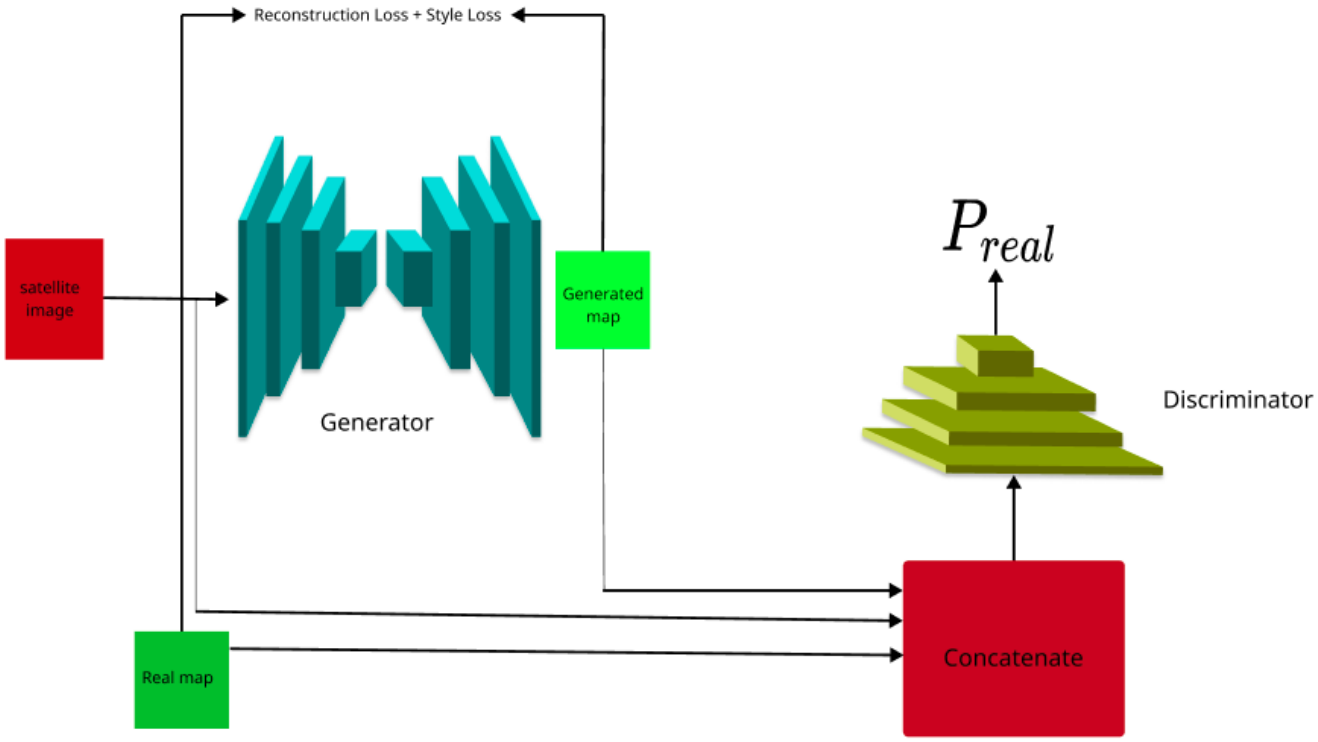
$$G = argmin_G max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L_1}(G) + \mathcal{L}_{Style}(G)$$

This framework is proposed by Swetava Ganguli, et al. in the paper titled "GeoGAN: A Conditional GAN with Reconstruction and Style Loss to Generate Standard Layer of Maps from Satellite Images". The authors report that it helped them achieve a better image quality compared to the framework presented by Pix2Pix GAN. I would like to note that, following GeoGAN, the Style loss is calculated on the map images directly rather than taking a sum over the activation of all the layers.

Network Architecture

Following Philip Isola, et al., the same generator and discriminator are used for all the experiments. Thus, the generator is a modified U-net architecture and the discriminator is known as a PatchGAN. This discriminator tries to classify if each N ×N patch in an image is real or fake. Thus, it judges a generated image patch-wise instead of giving a single value for the whole image. Moreover, the authors state that the PatchGAN acts as a kind of texture/style loss due to which they obtain highly realistic images.

Finally, the whole framework can be formulated as follows:



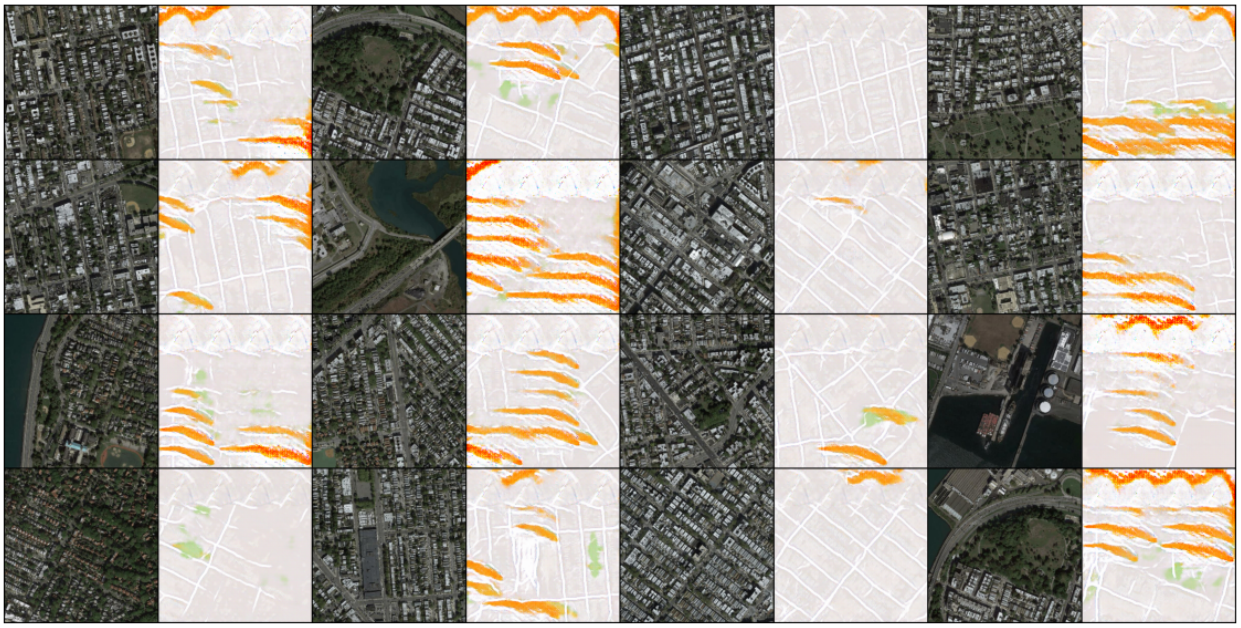
Training Details

Following Philip Isola, et al., I use Adam optimizer with a learning rate of 0.0002, and momentum parameters $\beta 1 = 0.5$, $\beta 2 = 0.999$. The batch size is set to 1. Random jitter and horizontal flips are used as augmentations. Weights were initialized from a Gaussian distribution with a mean of 0 and a standard deviation of 0.02. Finally, the networks are trained for 200 epochs and the lambda parameter is set to 100 for all the experiments.

Evaluation metrics

Evaluating generated images is an open and difficult issue. There have been numerous methods proposed. Although the difference in images might be lucid, I will provide numerical results by averaging discriminator losses (BCE), which are obtained after each experiment. Therefore, a generator should be able to fool all 4 discriminators well resulting in high discriminator losses.

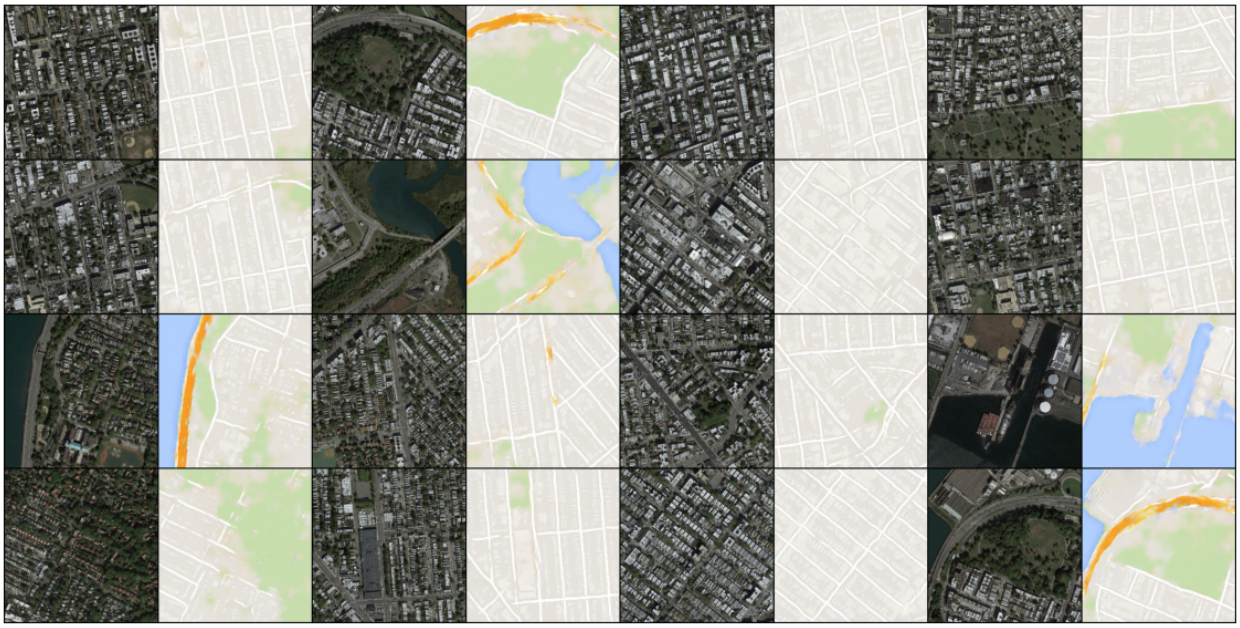
BCE



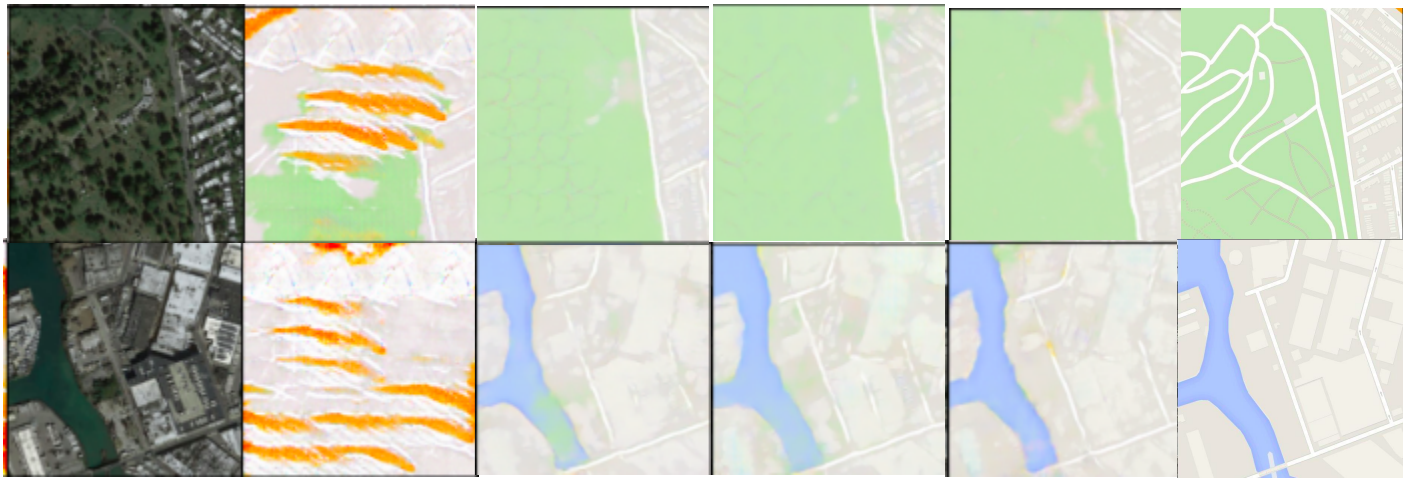
BCE + L1

BCE + L1 + Style Loss

L2 + L1



Satellite image	BCE	BCE + L1	BCE + L1 + Style	L2 + L1	Ground truth



As it is seen the baseline model (BCE) turned out to be biased towards orange lines since it always produces them. Furthermore, I can clearly see that adding L1 distance significantly improves the image quality. Likewise, adding Style Loss results in slightly better images. Comparing these two models, we can observe that Style Loss improves on some small but crucial details. For example, it produces better images for satellite photos containing both water and ground, whereas the

former model struggles to make an accurate transition between two surfaces. Finally, LSGAN turned out to be as good as the one with the Style Loss. However, it is interesting to note that LSGAN produces better maps for satellite images containing a lot of trees. As it can be noticed trees make other GANs generate circular shapes, whereas LSGAN is able to ignore that and produces maps that are closer to the ground truths. Next, I used all the discriminators to evaluate GANs. The following table shows average BCE losses, the higher the loss the better GAN is. As it is observed LSGAN is also the best in terms of average discriminator loss; however, the one with the Style loss demonstrates a comparable performance.

Method	**Average BCE loss**
BCE	**0.23**
BCE + L1	**0.48**
BCE + L1 + Style	**0.57**
L2 + L1	**0.59**

Conclusion

Although there is room for improvements, the results in this capstone project suggest that conditional GANs are a promising approach for human-readable map generation tasks. Additionally, a series of experiments demonstrate that objective functions significantly affect the final image quality.

References

Ganguli, Swetava, Pedro Garzon, and Noa Glaser. "GeoGAN: A conditional GAN with reconstruction and style loss to generate standard layer of maps from satellite images." *arXiv preprint arXiv:1902.05611* (2019).

Mao, Xudong, et al. "Least squares generative adversarial networks." *Proceedings of the IEEE international conference on computer vision*. 2017.

Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.