GANs - I'm Something of a Painter Myself competition

Covasan Iosif Andrei

January 16, 2024

Abstract

This work presents a solution for the Kaggle competition "I'm Something of a Painter Myself" using a CycleGAN architecture with U-Net generators. Leveraging image augmentation techniques and stabilization efforts, the proposed model achieved a significant improvement, securing a 14th position on the public leaderboard with a MiFID score of 41.45087. Inspired by seminal works and related studies, the approach showcases the effectiveness of the U-Net-based CycleGAN within the competition's constraints. Additionally, preliminary experiments with a ResNet generator hint at its potential superiority with extended training time.

1 Intoduction

1.1 Context

This Kaggle competition aims to generate 7,000 to 10,000 images in the style of Monet using a Generative Adversarial Network (GAN). Evaluation is based on a score called MiFID (Memorization-informed Fréchet Inception Distance), which considers the memorization distance of training samples. Participants are required to create JPG images of size 256x256 pixels, zip them into a file named "images.zip," and submit this file through a Kaggle notebook. Following a recommended Kaggle tutorial is advised, and there are specific time constraints for code execution on different processor types (5-hours run-time on a GPU/CPU Notebook).

1.2 Solution and results

My solution for the aforementioned competition involves a Cycle-Consistent Adversarial Network model inspired by the scientific paper "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks" [6] which I have enhanced and adapted to meet the competition's requirements and limitations. In contrast to the model in the original paper, my model employs a U-net architecture for both generators. Given the relatively small size of the dataset, consisting of only 300 Monet paintings, U-net can capitalize on its ability to learn finer feature relationships, having a higher number of

feedforward and backward connections. Due to time constraints, with the model having only 5 hours for training, U-net's simpler workflow provides better results. Moreover, image augmentation has been employed during training through scaling, horizontal flipping, and color variation transformations, enhancing the learning process by introducing increased diversity into the images. Furthermore, I attempted to stabilize the model training procedure by replacing the negative log likelihood objective with a least-squares loss in the adversarial loss formula, as suggested in the original paper [6]. The best result achieved, based on the MiFID score automatically calculated by Kaggle, is 40.46968, while training on a GPU.

2 State of The Art / Related Works / Literature Review

The primary articles that served as inspiration for my work include "Generative Adversarial Nets" [2], "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks" [6], and "Least Squares Generative Adversarial Networks" [3]. These seminal works have provided foundational concepts and methodologies for my approach in the Kaggle competition. The "I'm Something of a Painter Myself" competition operates indefinitely with a rolling leaderboard. Notable results have been achieved by other participants utilizing CycleGAN with a two-objective dual-head discriminator employing binary cross-entropy (bce) and hinge losses. Additionally, successful implementations have been based on approaches outlined in the paper "CycleGAN with Better Cycles" [4], as well as incorporating concepts from "Differentiable Augmentation for Data-Efficient GAN Training" [5]. The current leading solution incorporates the article "CLIPTraVeLGAN for Semantically Robust Unpaired Image Translation" [1] by Yevgeniy Bodyanskiy, Nataliya Ryabova, and Roman Lavrynenko. My best result secures a 11th position on the competition's public leaderboard with a score of 40.46968, while the best score is 34.43192.

3 Detailed Solution

Data augmentation is done through the next image transformations:

- Scaling: Images are initially resized larger and then randomly cropped back to the original size of 256.
- Horizontal Flipping: Since the content of both photos and Monet paintings is not significantly affected by horizontal orientation, this transformation adds variability to the dataset.
- Color Variation: Slightly changing the colors of the images helps stimulate diverse lighting conditions for photos and introduces variability in the colors of Monet paintings.

In a CycleGAN, two generators (genPM and genMP) and two discriminators (discM and discP) are employed to facilitate photo-to-Monet and Monet-to-photo translations. These

generators and discriminators are optimized using the Adam optimizer during training. To guide the training process, specific loss functions are defined:

- **Discriminator loss**. For real images fed into the discriminator, the output matrix is compared against a matrix of 1s using the mean squared error. For fake images, the output matrix is compared against a matrix of 0s. This suggests that to minimize loss, the perfect discriminator outputs a matrix of 1s for real images and a matrix of 0s for fake images.
- **Generator loss**. This is composed of the three different loss functions below.
  - **Adversarial loss**. Fake images are fed into the discriminator and the output matrix is compared against a matrix of 1s using the mean squared error. To minimize loss, the generator needs to 'fool' the discriminator into thinking that the fake images are real and output a matrix of 1s.
  - **Identity loss**. When a Monet painting is fed into the photo-to-Monet generator, we should get back the same Monet painting because nothing needs to be transformed. The same applies for photos fed into the Monet-to-photo generator. To encourage identity mapping, the difference in pixel values between the input image and generated image is measured using the l1 loss.
  - **Cycle loss**. When a Monet painting is fed into the Monet-to-photo generator, and the generated image is fed back into the photo-to-Monet generator, it should transform back into the original Monet painting. The same applies for photos passed to the two generators to get back the original photos. To preserve information throughout this cycle, the l1 loss is used to measure the difference between the original image and the reconstructed image.

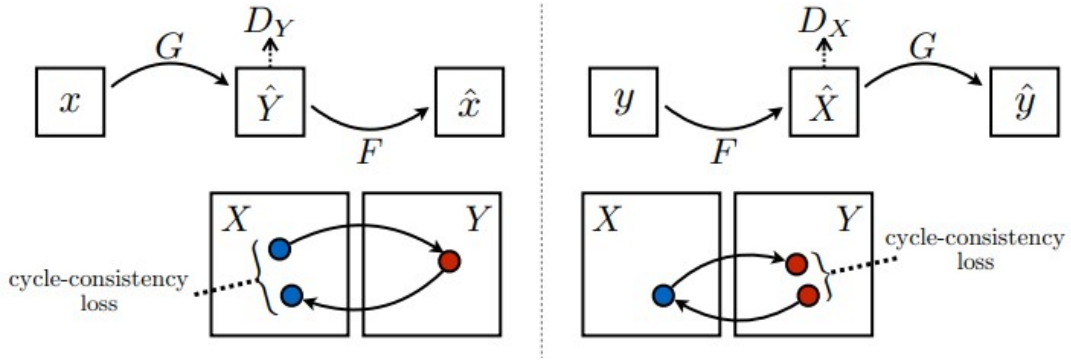

Figure 1: Cycle-consistency loss from [6]

From the above, the **mean squared error** and the **l1** loss are defined as the **adversarial criterion** and the **reconstruction criterion** respectively. The arhitecture for the CycleGAN generator is **U-Net**. U-Net is a network which consists of a sequence of downsampling blocks followed by a sequence of upsampling blocks, giving it the U-shaped architecture. In the upsampling path, we concatenate the outputs of the upsampling blocks and the outputs of the downsampling blocks symmetrically. This can be seen as a

kind of skip connection, facilitating information flow in deep networks and reducing the impact of vanishing gradients. Unlike conventional networks that output a single probability of the input image being real or fake, CycleGAN uses the **PatchGAN** discriminator that outputs a matrix of values. Intuitively, each value of the output matrix checks the corresponding portion of the input image. Values closer to 1 indicate real classification and values closer to 0 indicate fake classification. The idea behind using a PatchGAN discriminator is to provide fine-grained feedback to the generator by evaluating the realism of local patches in the generated image. This allows the generator to focus on capturing and improving details at a local level rather than the entire image. The original CycleGAN implementation updates the discriminator using a history of generated images instead of the latest images generated. This is done by setting up an **image buffer** that stores previously generated images. With probability 0.5, each newly generated image is swapped with a previously generated image stored in the buffer. This stabilizes training by giving the discriminator access to past information.

Figure 2: Photo-to-Monet Translation

3.1 Benchmark Performance

The best result achieved, based on the MiFID score automatically calculated by Kaggle, is 41.45087, while training on a GPU. This marks a substantial improvement compared to the baseline model provided by the competition, which attained a score of 53.76998, while training on a more powerful TPU. My best result secures a 14th position on the competition's public leaderboard, while the best score is 34.43192.

3.2 Ablation Study

Another approach attempted for the competition involved a cycleGAN utilizing a ResNet architecture with 9 residual blocks for both generators, akin to the implementation in the CycleGAN paper. Similar to the U-Net architecture, the ResNet generator consists of the downsampling path and upsampling path. The difference is that the ResNet generator does not have the long skip connections from the concatenations of outputs. Instead, the ResNet generator uses residual blocks between the two paths. These residual blocks have

#	Team	Members	Score	Entries	Last	Join
1	CLIPTraVeLGAN		34.43192	30	5d	
2	StarLabNumberOne		36.20936	3	9d	
3	Nandita Bhattacharya		37.06161	18	19d	
4	Anurag Dixit		37.91659	2	2mo	
5	MeowoeM		38.62360	5	2mo	
6	leeekevin		38.94692	1	2mo	
7	Corina_Code		39.55551	18	4h	
8	yuyuyuyuu		39.94173	2	2mo	
9	hongxiang1117		39.94174	2	2mo	
10	j26129851		40.15675	4	2mo	
11	**Iosif Covasan**		40.46968	5	22m	

Figure 3: Competition Leaderboard

gen arhitecture	ResNet9	ResNet6	U-Net
MiFID score	50.20616	46.75595	40.46968

Table 1: arhitecture scores in competition

short skip connections where the original input is added to the output. However, due to limitations related to the small size of the dataset and time constraints, with the model having only 5 hours for training, this model does not achieve better results than the one utilizing a U-Net architecture for generators. This model attains a score of 50.20616, training for only 16 epochs due to the competition's time constraints. However, allowing this model to run for an extended period, approximately 50 epochs, results in superior values for the adversarial criterion of the generators and discriminators compared to the model using the U-Net architecture if both models are trained for an equivalent duration. Therefore, the ResNet model should yield better results on this competition if we alleviate the time constraints. None of the proposed models in the paper is suitable for running on CPU, considering the constraints related to the small domain space and time limit of the competition.

Future research directions could focus on addressing significant issues faced by current CycleGAN models, including:

1. **Cycle Consistency Loss:**
   (a) The fundamental problem lies in data loss during image translation. For instance, when converting zebras to horses, stripes are lost, and models struggle to accurately restore these details.
   (b) Cycle Consistency Loss fails to provide motivation for the precise preservation of semantics in image translation. A future direction might involve seeking a more suitable objective function, possibly starting with a regular GAN.
2. **Random Initialization Problem:**

(a) Learning outcomes significantly depend on the type of initialization. This is an essential hyperparameter, and exploring more efficient methods than the current brute-force approach could be beneficial.

(b) A research hypothesis could involve pre-training the generator and discriminator separately on supervised learning problems, such as classification.

3. **Small Dataset Issue:**

   (a) Another significant problem arises when dealing with a relatively small dataset, such as the 300 Monet paintings. The Monet discriminator tends to specialize in these specific images, leading to a decrease in performance when presented with other real Monet paintings.

4 Conclusion

Through the proposed approach in the Kaggle competition "I'm Something of a Painter Myself" I have demonstrated the effectiveness of the CycleGAN model, employing U-Net generators, in generating images in the style of Monet. The achieved results underscore the potential of this architecture under limited dataset and training time conditions. Furthermore, experiments with a ResNet generator suggest a promising direction for future improvements under extended training time. The approach draws inspiration from seminal works and related studies, aligning itself with established methodologies in the field. In summary, this work contributes a competitive solution to the Kaggle competition, showcasing the potential of CycleGAN architectures with U-Net generators while highlighting avenues for future improvements and research.

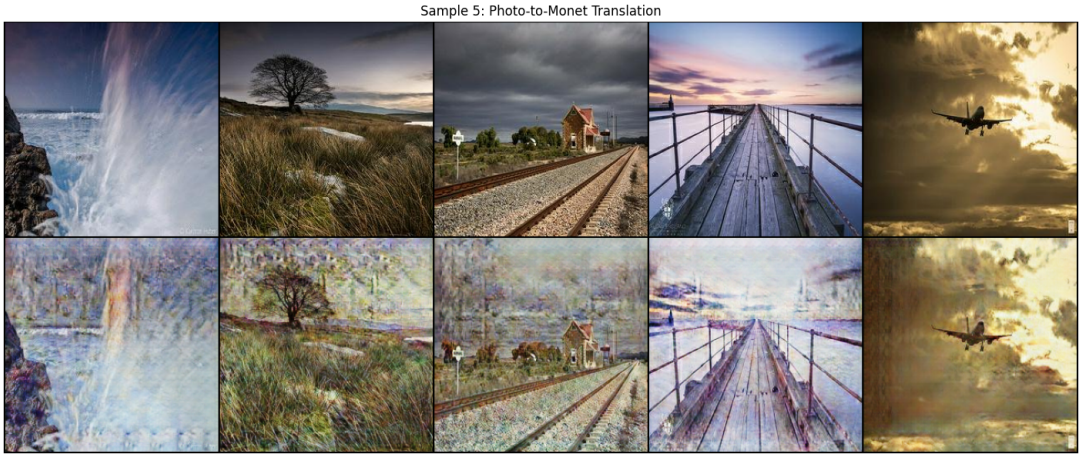

Figure 4: Photo-to-Monet Translation

References

[1] Yevgeniy Bodyanskiy, Nataliya Ryabova, and Roman Lavrynenko. Cliptravelgan for semantically robust unpaired image translation. 2023.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[3] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[4] Tongzhou Wang and Yihan Lin. Cyclegan with better cycles, 2018.

[5] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *Advances in neural information processing systems*, 33:7559–7570, 2020.

[6] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.