University of Dhaka
Department of Computer Science and Engineering

Artificial Intelligence Lab

CSE - 4111

Final Lab Project

Submitted by:
Tanjid Hasan Tonmoy
Roll: 09

April 21, 2019

Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

1 Introduction

The task of learning mapping between images from two different domains without aligned pairs of training data is referred to as unpaired image to image translation. The proposed technique, referred to as CycleGAN [1], combines Generative Adversarial Networks (GAN) [2] in conjunction with minimization of cycle consistency loss to produce noteworthy results for domain transfer of images. In this context, minimization of cycle consistency loss refers to constraint that the generated image should be identical to the the input image when converted back to the source domain. Additionally, the CycleGAN is composed of generators that take images from one domain as input and create images in the corresponding domain and a discriminators that distinguish which domain the images are from. The aim is to learn such a mapping so that the discriminators are no longer able to distinguish the generated images from the genuine target domain images.

The major contribution of the paper is the idea of using cycle consistency loss, eliminating the need for paired training data which is unavailable in most cases. Experiments conducted to reproduce the results reported by the authors and analysis of the obtained results follow in subsequent sections.

2 Experiments and Results

The experiments designed to replicate the results in [1] are described in subsection 2.1, the evaluaion criteria and the obtained results are discussed in subsections 2.2, 2.3 and 2.4. The limitations of the proposed model is discussed in subsection 2.5.

2.1 Implementation Details

The complete model consists of two generators and two discriminators. The generators (G_{AB} and G_{BA}) translate image from one domain to the corresponding domain. Here, A and B indicate the two domains and their order indicate the direction of the image translation. The discriminators (D_A and D_B) distinguish between images from their own or different domains.

The generator and discriminator architectures were implemented according to the suggested methods in the paper. For training of the model, the hyper-parameters and the loss function were set as proposed in the paper. The optimization objective for training of the model is to minimize the generator losses and maximize the discriminator losses. Here, combined loss function includes cyclic consistency or identity loss in addition to general generator loss minimized by producing close to target domain images. Cycle consistency loss is explained in Figure 1 with an example. For the combined model, the generators were trained jointly. However, the discriminators were trained separately. During the combined training of the model, training of the discriminators was turned off. However, The implementation differs from the original paper in the following ways-

- Image resolution of 128x128 has been used instead of 256x256 due to limited GPU memory and to reduce the training time.
- Batch size of 64 has been used instead of 1 to reduce model training time.
- All of the models have been trained for 100 epochs due to resource constraints, whereas the models described in original paper were trained for significantly more epochs.

2.2 Evaluation Criteria

Different metrics have been used to evaluate the results and compare them against the existing base line approaches. One of the metrics is human evaluation, where people hired on Amazon Mechanical Turk were asked to distinguish between the generated images and the real images. This metric is referred to as AMT perceptual realism in [3]. The result for this metric has been included without any modification from the paper.

The generated images have been analyzed for various application specific use cases in subsection 2.4. Such analysis addresses the conundrum that is faced in the evaluation of images generated by adversarial networks for their usability in practical tasks in various application domains.

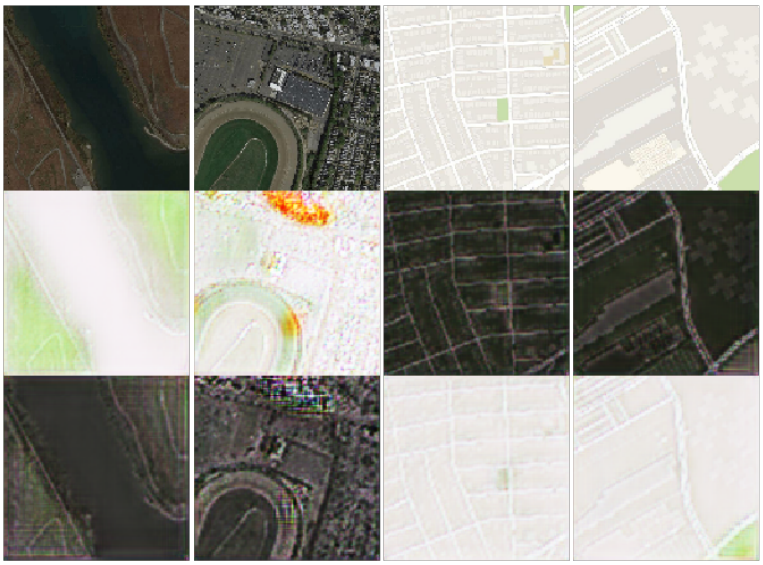

Figure 1: Illustration of cyclic consistency enforcement: Input images, translated images and reconstructed images are shown on the top row, middle and bottom row respectively

Figure 2: Translation of urban images to semantically labeled image and vice-versa using CycleGAN, top row indicates input image and bottom row indicates translated image

Moreover, FCN score, defined for the cityscapes benchmark [4] is used to compare the proposed model with several baseline approaches described in the next section.

2.3 Comparison against baselines

The baseline methods for comparison with the proposed model include CoGAN [5], Pixel loss + GAN [6], Feature loss + GAN [6], BiGAN/ALI [7] [8] and pix2pix [3]. Comparative results for AMT perceptual realism metric are given in Table 1 and results for FCN-score are given in Table 2. Example of translation of urban images to semantically labeled images and vice-versa are illustrated in Figure 2. A comparison in terms of images are shown in Figure 3.

Table 1: AMT "real vs fake" test on maps → aerial photos.

Loss function	**Map → Photo % Humans labeled real**	**Photo → Map % Humans labeled real**
CoGAN [5]	0.6% ± 0.5%	0.9% ± 0.5%
BiGAN/ALI [7] [8]	2.1% ± 1.0%	1.9% ± 0.9%
Pixel loss + GAN [6]	0.7% ± 0.5%	2.6% ± 1.1%
Feature loss + GAN [6]	1.2% ± 0.6%	0.3% ± 0.2%
CycleGAN [1] (Original Result)	26.8% ± 2.8%	23.2% ± 3.4%

Table 2: Comparison of FCN-scores for different methods, evaluated on Cityscapes dataset

Loss Function	**Per-pixel acc.**	**Per-class acc.**	**Class IOU**
CoGAN [5]	0.4	0.1	0.06
BiGAN/ALI [7] [8]	0.19	0.06	0.02
Pixel loss + GAN [6]	0.2	0.1	0.04
Feature loss + GAN [6]	0.06	0.04	0.01
pix2pix [3]	0.71	0.25	0.18
CycleGAN [1] (Original result)	***0.52***	***0.17***	***0.11***
CycleGAN [1] (Obtained result)	***0.39***	***0.08***	***0.06***

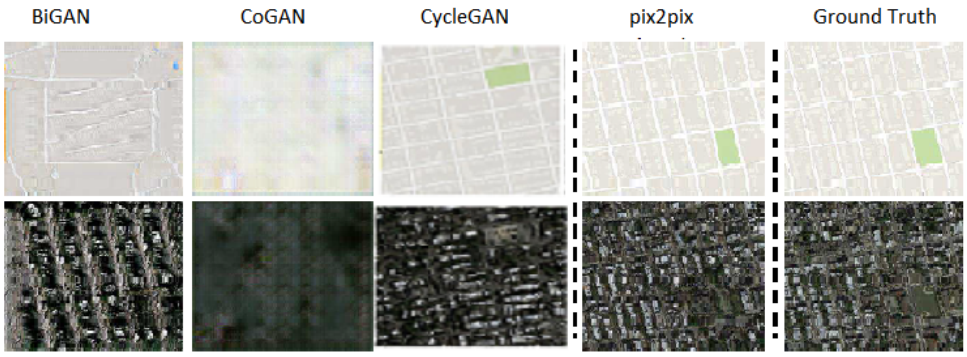

Figure 3: Comparison of baseline approaches with cycleGAN for the task of converting photos to map images.

2.4 Application based results

2.4.1 Object Transfiguration

The process of converting image of one object to another one is referred as object transfiguration. Examples of such image translation is given in Figure 4

2.4.2 Collection Style Transfer

Style transfer algorithms typically translate one image in the style of another image. However, using cycleGAN images can be translated to style of a collection of images e.g. Monet paintings. One such example is shown in Figure 5.

2.4.3 Generation of Photo from Paintings

Similar to style transfer, cycleGAN may be used to transform paintings to photographs preserving the salient features of the paintings. It should be noted that paired images for such tasks are nearly impossible to collect making this the only viable approach. Figure 6 shows one such example.

2.4.4 Season Transfer in Images

This is illustrated in Figure 7.

2.4.5 Enhancement of Images

Similar to the other approaches cycleGAN may be used for enhancement of images.

Figure 4: Example of converting images of apples and oranges, top row indicates input and bottom row indicates output image

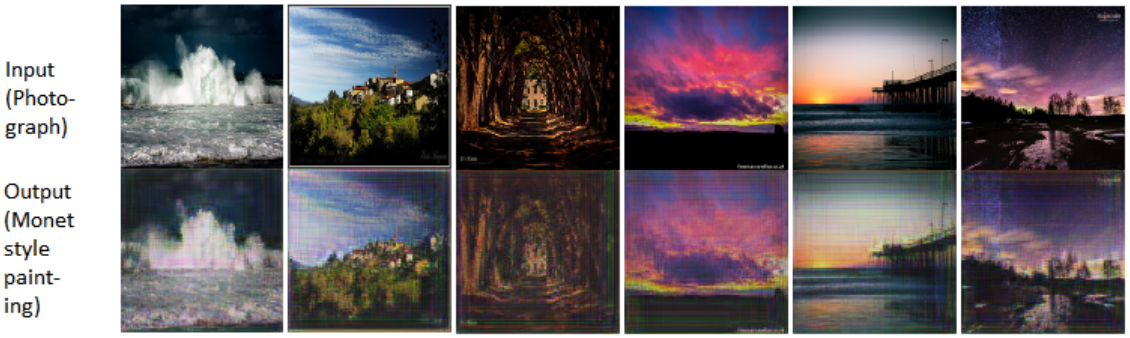

Figure 5: Example of collection style transfer- conversion of photo to Monet style

Figure 6: Example of generation of photo from painting

Figure 7: Example of season transfer, trained on photos of Yosemite

Figure 8: Poor results obtained during conversion of orange image to apple image, only color transformation performed

2.5 Limitations of CycleGAN

CycleGAN produces the best outputs when domain transfer could be accomplished by transformation of color and texture. However, when geometric transformations are required, results often degrade. Moreover, poor results have also been observed in cases where the distribution of training and test data are significantly different. Example of such poor result is illustrated in Figure 8.

3 Discussion and Conclusion

Since human evaluation has not been possible on the reproduced outputs, the original evaluation has been relied on. It should be noted that the models have been able to produce some convincing results as shown in section in spite of severe limitations in training.

The FCN scores were low compared to the original paper. However it may be explained by of the limitations in training described in section. The reduction of image resolution and subsequent resizing impacts the performance of pre-trained semantic segmentation algorithm used to calculate FCN score. The other limitations due to resource constraints also impacts the score. The superior results of pix2pix [3] may be explained by the fact that it requires and was trained on paired training data.

The application based results indicate that the model is able to learn mappings that can produce reasonably realistic image domain transfer results.

This work established the plausibility of unpaired image translation and indicate significant future scope of work to improve the method and overcome the limitations as well as the incorporation of such idea in different fields.

References

[1] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 2242–2251, 2017.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[3] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

[5] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Advances in Neural Information Processing Systems 29* (D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, eds.), pp. 469–477, Curran Associates, Inc., 2016.

[6] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2107–2116, 2017.

[7] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville, "Adversarially learned inference," *arXiv preprint arXiv:1606.00704*, 2016.

[8] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," *arXiv preprint arXiv:1605.09782*, 2016.