Preliminary exploration of CycleGAN models: Interim Report

G071 (s2056595, s2123157, s2196789)

Abstract

The purpose of image-to-image translation is to learn the mapping between an input picture and an output image using a training set of matched image pairings. In other cases, matched training data is not available. When matched examples are unavailable, we provide a method for learning how an image may be translated from one source domain to another. Using an adversarial loss, we want to discover a mapping G: X → Y such that the distribution of pictures generated by G(X) cannot be distinguished from the distribution generated by Y. We utilise an inverse mapping F: Y → X and a cycle consistency loss to guarantee F(G(X)) ≈ X since this mapping is very under-constrained (and vice versa). For a variety of tasks that lack matched training data, such as collection style transfer, object transformation, seasonality conveyance, or picture improvement, only qualitative findings are provided. The advantage of our methodology has been quantified in comparisons with a number of past methodologies.

1. Introduction

While the image to image translation problems has been widely investigated since the first generation of Generative Adversarial Networks(GAN) was invented in June, 2014 by Ian Goodfellow et al, the most state-of-art method for this field of tasks is still being pursued by people year by year(Ian Goodfellow, 2014). Image-to-image translation has a wide range of applications including changing a carton photo to a realistic photo, or converting real landscape map to satellite map. However, at the early stage of the work, people were required to collect large amount of paired dataset for training neural networks, which is called pix2pix(Phillip Isola, 2016). In reality, it is usually very difficult to collect paired images due to the constraints of different situations, and hence the generated results are not quite satisfactory. Purposed by Jun-Yan Zhu et al, in 2017, CycleGAN had greatly improve the performance in tasks of image-to-image translation, and became very popular as a state-of-art method(Jun-Yan Zhu, 2017).

The most significant improvement in CycleGAN is that it doesn't require paired images, which means we only need to care about one kind of images if we want to translate from A to B.The architecture of CycleGAN consists of two Generative models and two discriminator models. The function of generator G maps from X to Y such that the distribution of X are reproduced in the generated results $G(X)$, and the generator F works exactly in the opposite direction. The reason why it uses two paired models is that it can prevent generators from overfitting. Otherwise the mapping function of the two generators would produce the same results in order to deceive the discriminator despite what their inputs are. There is a cycle consistency loss which guarantees $F(G(X)) \approx X$, which means the input of the generator G is able to be reproduced by the generator F when it takes $G(X)$ as its input, where $G(X)$ is the output of G.

There are some limitations about CycleGAN. Although it has good performance on tasks involving texture or color transformation, such as converting horses to zebras, it does not perform well on geometrical translations. In some examples of geometrical translations it fails to translate from cat to dog. Some distributions of features would also lead to failure. For instance, when translating from horses to zebras, it doesn't consider there could be a man on the horses and as a result the man was also covered with zebra stripes(Saxena, 2021).

In this report we aim to evaluate the performance of the original architecture of CycleGAN on various dataset and investigate some feasible ways to improve its performance especially on tasks involving geometrical translations such as translating from cat to dog as mentioned previously. The main idea is to tune some hyper parameters and explore some other techniques which can be integrated with CycleGAN to improve its performance.

2. Data set and task

In this report, we aim to explore two representative image-to-image translation tasks originally raised by CycleGAN, which are converting horses to zebras and converting cats to dogs. We use specific data set to achieve these tasks. For horses-to-zebras task, we use the original data set employed in CycleGAN, and for cats-to-dogs task, we use the extrinsic data set for geometric changes(Asirra, 2017). For data set in each task, we separate training set and test set for two domains in our task. For each domain, we choose around 1000 images as training set and around 100 images as test set.

We aim to employ the evaluation approaches mentioned in CycleGAN to compare the performance of original CycleGAN and our enhanced models both quantitatively and qualitatively. We will use the approaches in evaluation met-

rics of CycleGAN to evaluate the accuracy of each models in different levels. Besides, we will also use the visualized figures of cycle loss to evaluate the performance of each model on certain task.

3. Methodology

Our objective is to discover mapping functions between two domains X and Y using training data $\{x_i\}_{i=1}^{N}$ where $x_i \in X$ and $\{y_i\}_{j=1}^{N}$ where $y_i \in Y$. The data distribution is denoted by the variables $x \sim p_{data}(x)$ and $y \sim p_{data}(y)$. Our model, as indicated, has two mappings: $G : X \rightarrow Y$ and $F : Y \rightarrow X$. Additionally, we present two adversarial discriminators D_X and D_Y, where D_X seeks to differentiate between pictures $\{x\}$ and translated images $\{F(y)\}$ and D_Y seeks to distinguish between images $\{y\}$ and $\{G(x)\}$. Our aim includes two kinds of terms: *adversarial losses* for matching the distribution of produced pictures to the distribution of data in the target domain; and *cycle consistency losses* for preventing the learned mappings G and F from contradicting one another.

3.1. Adversarial Loss

Both mapping functions are susceptible to adversarial losses(I. Goodfellow, 2014). The aim for the mapping function $G : X \rightarrow Y$ and its discriminator D_Y is as follows:

$$\mathcal{L}_{GAN}(G, D_X, Y, X) = \mathbb{E}_{y \sim p_{data}(y)}[logD_Y(y)] + \mathbb{E}_{x \sim p_{data}(x)}[log(1 - D_Y(G(x)))]$$

in which G attempts to produce pictures $G(x)$ that resemble images from domain Y, and D_Y attempts to differentiate between translated samples $G(x)$ and genuine samples y. G seeks to reduce this goal whereas D seeks to enhance it, i.e., $min_G\ max_{D_Y}\ \mathcal{L}_{GAN}(G, D_Y, X, Y)$. We also include an adversarial loss for the mapping function $F : Y \rightarrow X$ and its discriminator D_X: $min_F\ max_{D_X}\ \mathcal{L}_{GAN}(F, D_X, Y, X)$.

3.2. Cycle Consistency Loss

It is theoretically possible to learn mappings G and F that yield identically distributed outputs as target domain Y and X, respectively, using adversarial training (strictly speaking, this requires G and F to be stochastic functions)(Goodfellow, 2016). Any random permutation of pictures in the target domain, in contrast, may induce an output distribution that matches the target distribution when a network's capacity is high enough. A learnt function's ability to map an individual input x_i to a desired output y_i cannot be guaranteed by adversarial losses alone. We suggest that the learnt mapping functions should be cycle-consistent in order to further minimise the number of viable mapping functions. We use a loss in cycle consistency as a reward for this conduct:

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_{data}(y)}[\|G(F(y)) - y\|_1]$$

Reconstructed pictures $F(G(x))$ end up matching closely to the input images x because of cycle consistency loss.

3.3. Full Objective

Ultimately, our goal is to accomplish the following:

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) + \lambda \mathcal{L}_{cyc}(G, F)$$

In this case, λ is used to regulate the relative relevance of the two goals. We're here to address:

$$G^*, F^* = arg\ min(G, F)\ max(D_X, D_Y)\ \mathcal{L}(G, F, D_X, D_Y)$$

Not that our model may be considered as training two "autoencoders"(G. E. Hinton, 2006): we learn one $F \circ G : X \rightarrow X$ and another $G \circ F : Y \rightarrow Y$ concurrently. Each of these autoencoders, however, has a unique internal structure: they map a picture to itself through an intermediate representation, which is a translation of the image into another domain. This configuration may alternatively be thought of as a subset of "adversarial autoencoders"(A. Makhzani, 2016), which use an adversarial loss to train an autoencoder's bottleneck layer to match an arbitrary target distribution. In our scenario, the $X \rightarrow X$ autoencoder's target distribution is the domain Y.

4. Experiments

In this experiment we trained the model on two dataset, one is horses and zebras, and the other one is cats and dogs. For the sake of controlling variables, we use the same hyperparameters for both datasets. The crop size was capped at 256 for each image in order to assure the whole object is included in the image. In the training process we use minibatch with size of 8 at each iteration. The performance affected by size of minibatch has not been investigated yet, but it doesn't make much difference on our results. We also tried to decrease the size of each image to increase the training speed, but the loss of generators was fluctuated and did converge, so we choose 256 by 256 as the input dimension of our model. We use adaptive learning rate with initial value of 0.0002 and decayed after 50 epoch.

Figure1 shows baseline experiment results of outputs for training samples for two datasets, which are horse to zebra and cat to dog. We trained our models on these two datasets for 200 epoch and 131 epoch respectively. There are some notations need to be clarified on this figure. Considering a task translating from A to B, real_A represents the input sample A given to the generator G and then generates the fake image $G(A)$ which is fake_B. Real_B and fake_A is for the other generator F working on the other way around, which is translating from B to A. The rec_A represents $F(G(A))$ which aims to reconstruct A in order to evaluate the cycle loss of the whole system. It is essential to keep the cycle consistency which ensures we can reconstruct A

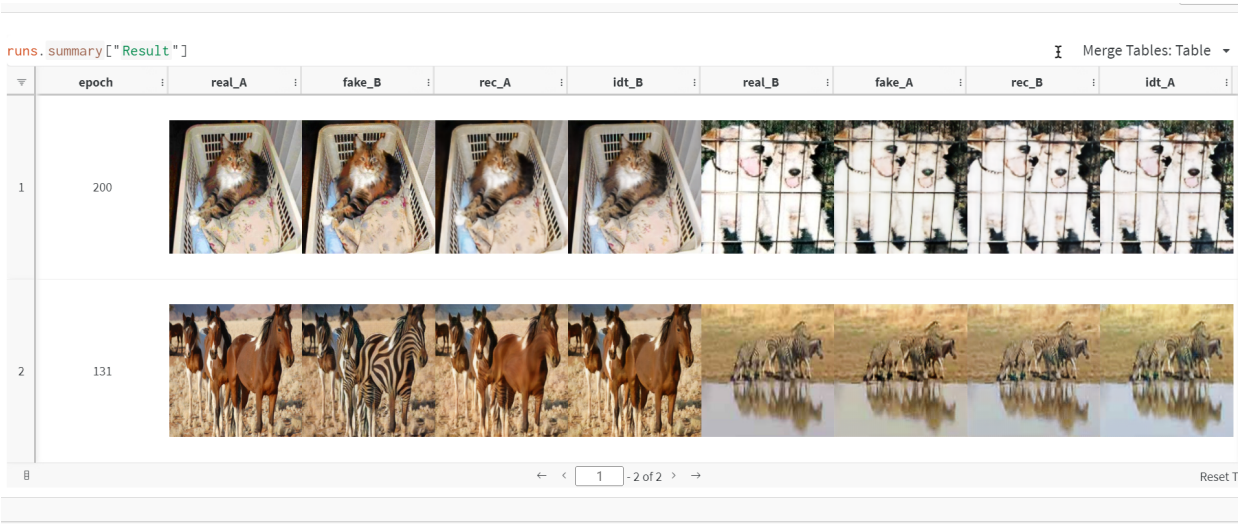

Figure 1. Outputs for training samples for two dataset at training epoch 200 and 131 respectively

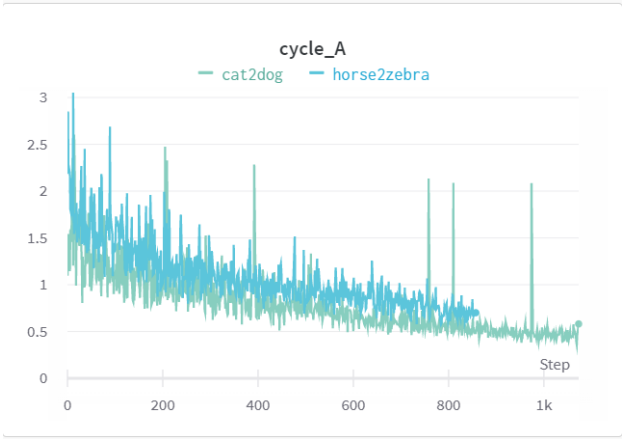

Figure 2. Cycle loss of translating from A to B for cat2dog and horse2zebra models

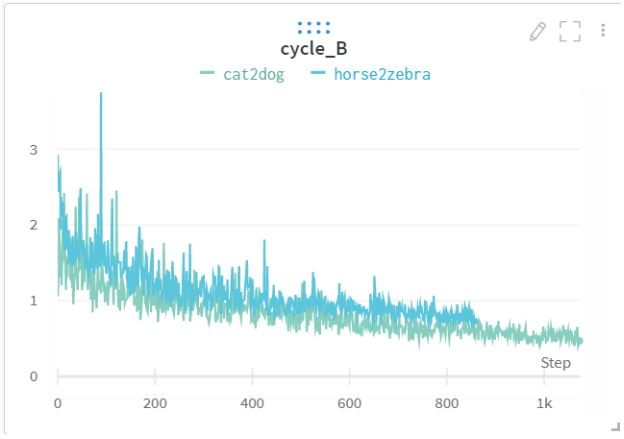

Figure 3. Cycle loss of translating from B to A for cat2dog and horse2zebra models

using generator F. The idt_A and idt_B is identity mapping which represents $G(B)$ and $F(A)$ to ensure $G(B) = B$ and $F(A) = A$.

It's not hard to notice that there is hardly no difference between real_A and fake_B, and hence our model failed to translate from cat to dog even after 200 epoch. On the contrary, the model did well on translating from horse to zebra. By comparing the cycle loss for both of these models, we could find that their loss have the same trend. The cycle loss is indeed decreasing after a certain amount of iterations. In the model of translating from cat to dog, it learned some patterns to reduce the overall loss, yet it didn't understand its task to translating the graphical features of the images.

5. Interim conclusions

5.1. Acquired knowledge

In picture creation, image manipulation, and representation learning, Generative Adversarial Networks (GANs) have achieved remarkable achievements.

Image-to-image translation dates all the way back to Hertzmann's Image Analogies(A. Hertzmann, 2001), which apply a non-parametric texturing model on a single input-output training image pair.

Numerous more strategies are equally applicable in the unpaired context, when the objective is to connect two distinct data domains.

The concept of employing transitivity to regularise structured data is not new. Enforcing basic forward-backward consistency has been a typical method in visual tracking

for decades(Z. Kalal, 2010).

Neural Style Transfer(L. A. Gatys, 2016) is another technique for doing image-to-image translation, in which the content of one picture is combined with the style of another image based on the Gram matrix statistics of pre-trained deep features.

5.2. Feasibility

Johnson's architecture(J. Johnson, 2016) for generative networks is used in our experiments, and he has shown outstanding results in the areas of neural style transfer and superresolution. Among the network's components are three convolutions, six residual blocks, two fractionally-strided convolutions with stride $\frac{1}{2}$, and one convolution that converts features to RGB.

To stabilise our model training approach, we use two strategies from recent research. To begin, we substitute a least-squares loss(X. Mao, 2017) for the negative log likelihood goal in $\mathcal{L}_{GAN}$. This loss is more consistent during training and resulting in higher-quality outcomes. Second, in order to minimise model oscillation, we use Shrivastava's method(A. Shrivastava & Webb, 2017) of updating the discriminators using a history of generated pictures rather than the most recent generators' images. We maintain an image buffer in which the 50 previously produced pictures are stored. Each network was trained from start using a 0.0002 learning rate. For the first 100 epochs, we maintain the same learning rate and then linearly decline the rate to zero over the following 100 epochs.

5.3. Flaws and defects

While our strategy often produces persuasive findings, the outcomes are far from universally good. Additionally, we investigated challenges that involve geometric modifications, although with little success. For instance, on the dog → cat transfiguration job, this failure might be a result of our generator designs being optimised for appearance modifications. Managing increasingly complex and intense transformations, particularly geometric transformations, is a critical issue for future work.

Certain types of failures are triggered by the distribution properties of the training datasets. Because our model was trained on ImageNet's wild horse and zebra synsets, which do not include photos of humans riding horses or zebras.

Additionally, we notice a persistent discrepancy between the outcomes obtained using paired training data and those obtained using our unpaired strategy. In other situations, closing this gap may be very difficult. To resolve this uncertainty, some type of weak semantic monitoring may be required.

6. Plan

The plan's overall research goal and design concepts will remain unchanged, and its essence will remain the use of image-to-image translation to tackle vision and graphics transformation problems(A. A. Efros, 1999). This plan's overarching objective is to further optimise the CycleGAN models based on our preliminary exploration. The technique of presenting is unique in that it processes data sets with specialised representations, namely picture transformations between cats and dogs, and fully exploits the optimization concept to considerably improve the effect over the original CycleGAN models.

By conducting a series of experiments, we found that it is difficult to improvement the performance of CycleGAN on tasks involving graphical changes by just tuning the hyperparameters. This limitations of CycleGAN could be solved by changing its basic architecture. In future experiment we plan to add an attention or self-attention mechanism which can improve its robustness. The attention mechanism(AM) has been widely used in deep learning areas such as natural language processing(NLP), computer vision, and speech processing. I

References

A. A. Efros, T. K. Leung. *Proceedings of the seventh IEEE international conference on computer vision*, 1999. URL https://ieeexplore.ieee.org/abstract/document/790383.

A. Hertzmann, C. E. Jacobs, N. Oliver B. Curless D. H. Salesin. Image analogies. *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001. URL https://dl.acm.org/doi/abs/10.1145/383259.383295.

A. Makhzani, J. Shlens, N. Jaitly I. Goodfellow B. Frey. Adversarial autoencoders. *ICLR*, 2016. URL https://arxiv.org/abs/1511.05644.

A. Shrivastava, T. Pfister, O. Tuzel J. Susskind W. Wang and Webb, R. Learning from simulated and unsupervised images through adversarial training. *Computer Vision and Pattern Recognition (CVPR)*, 2017. URL https://arxiv.org/abs/1612.07828.

Asirra, Petfinder. *Microsoft Research*, 2017. URL https://www.microsoft.com/en-us/download/details.aspx?id=54765.

G. E. Hinton, R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 2006. URL https://www.science.org/doi/10.1126/science.1127647.

Goodfellow, Ian. Generative adversarial networks. *NIPS 2016 Tutorial*, 2016. URL https://arxiv.org/abs/1701.00160.

I. Goodfellow, J. Pouget-Abadie, M. Mirza B. Xu D. Warde-Farley S. Ozair A. Courville Y. Bengio. Generative adversarial nets. *NIPS*, 2014. URL https://arxiv.org/abs/1406.2661.

Ian Goodfellow, Jean Pouget-Abadie. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. URL https://arxiv.org/abs/1406.2661.

J. Johnson, A. Alahi, L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *ECCV*, 2016. URL https://arxiv.org/abs/1603.08155.

Jun-Yan Zhu, Taesung Park, Phillip Isola. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017. URL https://arxiv.org/abs/1703.10593.

L. A. Gatys, A. S. Ecker, M. Bethge. Image style transfer using convolutional neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Gatys_Image_Style_Transfer_CVPR_2016_paper.html.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016. URL https://arxiv.org/abs/1611.07004.

Saxena, Pawan. Cycle generative adversarial network (cyclegan), 2021. URL https://www.geeksforgeeks.org/cycle-generative-adversarial-network-cyclegan-2/.

X. Mao, Q. Li, H. Xie R. Y. Lau Z. Wang-S. P. Smolley. Least squares generative adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. URL https://ieeexplore.ieee.org/document/8237566.

Z. Kalal, K. Mikolajczyk, J. Matas. Forward-backward error: Automatic detection of tracking failures. *2010 20th international conference on pattern recognition*, 2010. URL https://ieeexplore.ieee.org/abstract/document/5596017.