Convert Screenshots of Breath of the Wild into Realistic Style using Cycle-Consistent Adversarial Networks

Wei Chen, Quanqing Que, Zhihao Jin
Xiamen University

Abstract

Image-to-image translation aims at generating a new synthetic version of a input image with a specific modification. Recently, since Generative Adversarial Network (GAN) based convolutional neural networks achieved remarkable progress in several computer vision tasks, related technics is also introduced to this field. Especially, Cycle-Consistent Adversarial Networks (CycleGAN) achieves impressive results on this task while trained in a unsupervised manner without using a lot of paired images which is hard to collect. In this course project, we are going explore the potential of CycleGAN. Our goal is to transfer screenshots of a specific video game into a realistic style. By completing this course project, we hope we can understand CycleGAN better and even try to improve it. Screenshots of our project are taken from a popular game, *The Legend of Zelda:Breath of the Wild*, whose visual style is distinctive and representative, so it's suitable for this task.

Introduction

Image-to-image translation aims at generating a new synthetic version of a input image with a specific modification. However, collecting paired images may be expensive for many tasks. Zhu et al. proposed Cycle-Consistent Adversarial Networks (CycleGAN), which is trained in an unsupervised manner, achieves impressive results in this field. By completing this course project, we hope to understand CycleGAN better and even try to improve it. Specifically, we are going to transfer the screenshots from a popular video game, *The Legend of Zelda: Breath of the Wild (BOTW)*, into a realistic style with the help of CycleGAN.

BOTW is a action-adventure game developed and published by Nintendo in 2017. In BOTW, a unique artistic style and brilliant colors are used to outline a beautiful and magnificent fantasy world. The style of BOTW is quite distinctive and fantastic. Therefore, the task of converting the screenshots of BOTW into a realistic style is challenging and interesting.

Figure 1: A screenshot from the popular game Breath of the Wild.

Dataset was prepared in this way. First, we collect the screenshots of BOTW from a video game emulator. Screenshots are then preprocessed and resized to remove special icons and downsampled to a acceptable size. Then, real world images are collected from unsplash.com. Downsampling are also applied on these real world images to reduce their memory footprint. Thanks to the superior design of CycleGAN, CycleGAN can be trained unsupervised in the absence of any paired training examples.

The theory behind CycleGAN is simple. The goal of image-to-image translation is to relate two image domains X and Y. A generator $G : X \rightarrow Y$ is learned to translate images from X to Y under the adversarial loss. So theoretically, let $\hat{y}$ is transfered from $x \in X$ by mapping G, we can expect that the predicted $\hat{y}$ is indistinguishable from Y. However, in practice, this simple pipeline often lead to the well-known problem of mode collapse, where all input images map to the same $\hat{y}$. So CycleGAN introduces a constraint named as "cycle consistency" to ensure that generated images can be transferd to its origin domain under another mapping $F : Y \rightarrow X$. Intuitively, if we translate a sentence from English to French, and then translate it back from French to English, we should arrive back at the original sentence. Mathematically, if we have a thanslator $G : X \rightarrow Y$ and another translator $F : Y \rightarrow X$, then G and F should be inverses of each other, and both mappings should be bijections. CycleGAN introduces *cycle consistency loss* which encourages $F(G(x)) \approx x$ and $G(F(y)) \approx y$. Both adver-

sarial loss and cycle consistency loss are considered in the final objective function of CycleGAN.

Challenges and Contributions

The first challenge we face is the lack of computation resource. We mainly use CycleGAN to achieve the transfer from the style of BOTW into realistic style. And as a comparison, U-GAT-IT, another GAN method, will be used for our experiment. Since these two models are proposed in recent years, they require large computing power. Therefore, only smaller resolution images can be used for training during the experiment.

The second challenge is that BOTW contains many elements that do not exist in reality(such as temples,monsters, towers, etc.). These things that do not exist in reality will result in some problems, e.g. these things are transformed into a strange object. In order to avoid this situation, we filter the screenshots obtained from BOTW and try to choose pictures that do not contain the above elements.

The third challenge is the low resolution pictures from BOTW are already similar to real photos in style, making the discriminator hard to train. Due to the limitation of computation resource, we have to use the low resolution pictures. These low resolution pictures from BOTW are less different from those in reality. In other words, some details of BOTW style become less obvious due to reduction of resolution, which makes it more difficult for our models to learn the difference between BOTW style and realistic style.

Related Work

Image Style Transfer (Gatys, Ecker, and Bethge 2016) involves rendering the semantic content of an image in different styles. They use image representations derived from convolutional neural networks optimised for object recognition, which make image information explicit. Benefited from the deep image representation, their work can combine the content of an arbitrary photograph with the appearance of numerous well-known artworks by matching the Gram matrix statistics of pre-trained deep features. Despite its advantages, it still have many limitation, e.g. their model relies on the user specified "target image", which limited its applications. Besides, a gradient desent based process was adopted to optimize the input image gradually, which harms the efficiency of their algorithm.

Although the method proposed by Gatys, Ecker, and Bethge for *artistic* stylization looks impressive, this method does not aim at transfering images into realistic style. Further research indicates that this method produces noticeable artifacts if a real photo is provided as style image. Fujun et al. introduces a deep-learning approach to style transfer that handles a large variety of image content while faithfully transferring the reference style. This approach builds upon the recent work on painterly transfer that separates style from the content of an image by considering different layers of a neural network. The transformation from the input to the output is constrained to be locally affine in colorspace. Li et al. introduced a closed-form solution for photorealistic image stylization. This method consists of a stylization step and a postprocess step, stylization step performs photorealistic image stylization and the postprocess step improves the stylization effects by reducing the artifacts. Instead of optimize the solution iteratively, this method can produce the output image in a fixed number of operations.

Generative Adversarial Networks (Goodfellow et al. 2014) with convolutional networks have seen huge adoption in computer vision applications. Tricks and experiences introduced by DCGAN described by Radford, Metz, and Chintala significantly makes the training of image generation model powered by GAN easier. WGAN (Arjovsky, Chintala, and Bottou 2017) further improved the original formulation of GAN by mathematical analyzation. GANs have achieved remarkable results in many computer vision tasks, e.g. image generation, image super resolution, and image editing. The key to GANs' success is the idea of *adversarial loss*, which is evaluated by a discriminator which is essentially a deep learning model. Compared to traditional loss functions, adversarial loss is more powerful in many cases, especially some times the hand-crafted objective functions are hard to design. Since its unsupervised training method, GANs are recieving more and more attentions recently.

Cycle-Consistent Adversarial Networks (CycleGAN) Traditional image-to-image translation models either rely on a huge paired image dataset or need a user specified "target image" as a hint. To address these drawbacks, GAN based models were introduced to achieve unsupervised training without paired image data. However, primitive GAN pipeline can easily lead to a well-known problem of "mode collapse". Zhu et al. addressed this problem by introducing cycle-consistency loss. Although CycleGAN achieved impressive results, there are still remaining issues unsolved, e.g. people noticed that CycleGAN is good at performing color or texture transformation, while performs not well on geometrical transformations. Besides, CycleGAN tends to "hide" information about a source image into the images it generates, ensures that the output image can be easily recovered to the origin domain.

Data Collection

The quality of datasets is critical to the training. Our datasets contain two parts: the screenshots of BOTW and realistic pictures. First, we collect the screenshots of BOTW from a video game emulator. BOTW is an action-adventure game, in which the player controls Link to defeat Calamity Ganon and save the kingdom of Hyrule. Therefore BOTW has many elements that do not exist in reality. these elements will bring about some issues, e.g. these elements are transformed into strange objects or disappear from the picture. For this reason, it is necessary to try to avoid these elements in the picture when taking screenshots.

Besides, we collect real world images from unsplash.com. When collecting real world pictures, it is important to ensure that the datasets contain the following types of landscape photos: grasslands, forests, lakes, deserts, mountains and snowcapped mountains. The reason for this is that screenshots from BOTW contain these types of scenery. If the real world dataset lacks snow mountains picture, for example, the color of snow mountains in BOTW may be converted into a peculiar color.

Method Selection

There have already many proposed image-to-image translation methods, from which some representative models were choosed and experimented by us. These models are described briefly and the corresponding experiment results are represented as below.

Deep Photo Style Transfer Deep Photo Style Transfer mainly puts forward two ways of improvement to realize better style tranfer tasks. Before this work, the output of style transfer method tend to looks like a painting. Inspired by the Matting Laplacian, in order to the remove painting-like effects the algorithm builds a transformation model that is locally affine in colorspace to prevent spatial distortion. On the other hand, problems caused by content differences between the input image and the reference image often occur, which may result in undesired transmission between irrelevant content. To address this issue, Deep Photo Style Transfer chooses to use semantic segmentation of the input and reference images.

To implement Deep Photo Style Transfer into our work of converting screenshots of Breath of the Wild into realistic style, we would have to first find a pair of pictures that are sufficiently similar in composition. But due to the limited size of the data set and the diversity of the screenshot content itself, it's very difficult to pick out a pair images that contains the same elements and shares similar struct. In our work to test this algorithm, We select a set of pictures that all included the three main components of the house, the sky and the lawn, and the composition of the two were also roughly similar. We do the semantic segmentation of images by using the Photoshop quick selection tool manually. There are two main problems with the results we get through this algorithm, the output image looks blurry especially where different parts meet and some minor element such as branches near the wall are distorted. Our experiment results are even more unsatisfactory on other image pairs with lower matching levels, which means the quality of output image produced by Deep Photo Style Transfer is highly relied on the input images and the semantic segmentation done before. In conclusion, to use this method to achieve our goal, we need to collect a large number of realistic pictures that match the screenshots of the game, which is not realistic and the quality of the generated pictures is not satisfactory either.

(a) Original Image (b) Style Image (c) Transfered Image

Figure 2: Some results of Deep Photo Style Transfer. The overall style of the picture has been changed, but it is still blurry, and the background element has been added by mistake.

(a) (b)

Figure 3: Some representative results of the original CycleGAN. Although translated image seems realistic to some extent, the image looks blurry. (a) is the original input, (b) is the translated image.

CycleGAN CycleGAN is a kind of image-to-image translation algorithm with no need of paired input-output examples. CycleGAN assume that there is some underlying relationship between the domains and seek to learn that relationship. Since standard GAN based methods often lead to mode collapse, "cycle consistent" is then introduced to facilitate the training process.

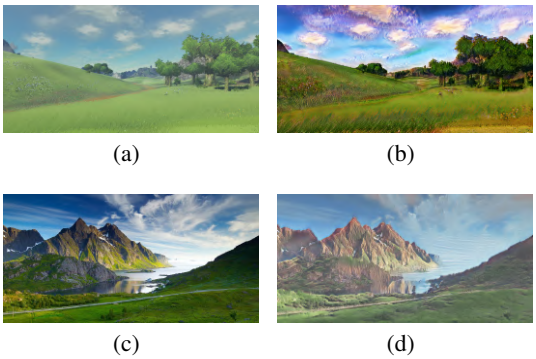
CycleGAN is really a ground breaking work which is useful in many image translation tasks, for example, transfer the time of day, weather, season, and artistic edits. It's natural to try to utilize CycleGAN to translated the screenshots of BOTW into realistic style. However, although we found CycleGAN actually effective to some extent, it's performance is not very perfect as we expected. As far as our task is concerned, the drawbacks of CycleGAN is listed as below:

1. The output image looks blurry;
2. It tends to adjust the color of the images, while the texture is generally the same before and after translation.

Some example results of the original CycleGAN are illustrated by figure 3.

U-GAT-IT

U-GAT-IT is another method for unsupervised image-to-image translation, which incorporates a new attention module and a new learnable normalization function in an end-to-end manner. The attention module guides the model to focus on more important regions distinguishing between source and target domains based on the attention map obtained by the auxiliary classifier. Unlike previous attention-based method which cannot handle the geometric changes between



(a) (b) (c) (d)

Figure 4: Comparison of source image and converted image: (a) Source image from BOTW, (b) Realistic style transformed by U-GAT-IT, (c) Source image from real world, (d) BOTW style transformed by U-GAT-IT

domains, the model can translate both images requiring holistic changes and images requiring large shape changes. Moreover, the new AdaLIN(Adaptive Layer-Instance Normalization) function helps the attention-guided model to flexibly control the amount of change in shape and texture by learned parameters depending on datasets.

Some example results of UGATIT are illustrated by figure 4. It can be seen that the BOTW image is still quite different from the real style after the conversion, and the color of the clouds appears strange. On the other hand, the color and texture of the real picture after conversion are closer to the BOTW style, which shows that the conversion effect is better in this case. Comparing two pictures after conversion, it can be seen that the styles of the two pictures both show a special style like the oil painting. We guess that this result is mainly due to insufficient training of the model.

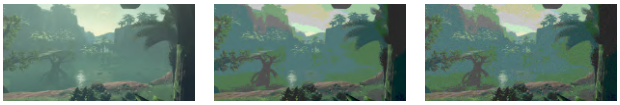
Finally, we choosed CycleGAN as our base model to achieve our goal. This is because that image-style-transfer based methods need extra style image as input, this is not convenient, and the style image should be carefully selected in order to produce a satisfactory output. And U-GAT-IT is expensive and hard to train. Only small image can be used to train U-GAT-IT while our screenshots are generally high resolution.

Compared with image-style-transfer based methods and U-GAT-IT, the original CycleGAN model just fits our requirement better. We will choose CycleGAN as our base model and try to improve it to achieve a better performance.

Model Improvement and Experiment Details

Preprocess

We found a better proprocess can improve the subjective quality of the output images. As is shown in our early experiment, the color tune of the output image looks different from the real scene. We suppose that the it may be because in the original style data set, some realistic pictures have brighter tones, but in the world of Zelda, the range of tones is not so large, which led to this situation. So we did some preprocessing on the real image data set, mainly to make the tones of the image uniform. The experiment results of this method looks relatively natural in color, but new problem occurs as well that some color tunes are changed incorrectly such as a sunny beach scene is tranfered into a blue style. The key point of this issue is that it is difficult to determine which pictures need to be preprocessed, and it is not easy to determine to what extent it should be processed. We decide whether it needs to be processed according to the average size of the rgb channels of the picture. After several experiments, we finally determined a relatively satisfactory threshold.



(a) (b) (c)

Figure 5: (a) Input image before preprocess; (b) Image after quantization; (c) Gaussian noise is added to (b).

As we mentioned earlier, CycleGAN tends to vary the color of the input image while remain the texture unchanged. Instead of change the inner structure of the CycleGAN model, we apply quantization to input images. By quantization, we expect some detail can be "erased" from the original images. Let's denote the input image as I, the quantization operation is formulated as:

$$I' = ceil(I/\alpha) * \alpha$$

Where the α is the "quantization level", the lager α is, the more information will be erased from I.

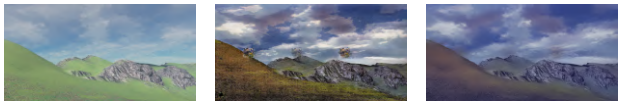
After quantization, gaussian noise is added into I'. By quantization and gaussian noise, we wish we can reduce the prior knowledge about the input image, make CycleGAN free to do more creative work.

Hyper Parameters and Training Details

All input image are resized and cropped into 1024x512. When training, images are loaded and cropped into 512x512. Unet is adapted as our generator backbone, and instance normalization is choosed as the normalization layer. λ_A is set to 16 while λ_B is set to 10, which encourages the model weights more on improving the performance of translating A (game screenshots) to B (real sceneries). The weight of identity loss, $\lambda_{identity}$, is set to 0.2 which forces generators do not modify the input if the input is already translated.

Postprocess

As illustrated by figure 6, image translated by CycleGAN come up with noticeable artifacts. We tried to remove these artifacts by guided image filter (He, Sun, and Tang 2010)



(a) (b) (c)

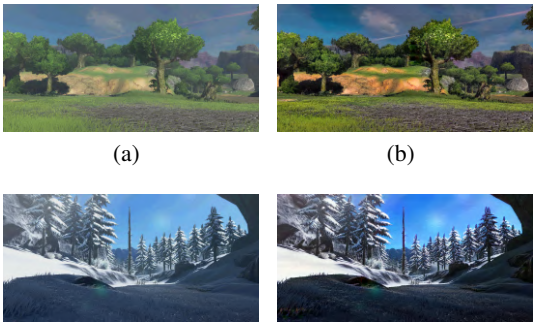
Figure 6: (a) The input image; (b) CycleGAN sometimes produces visual noticeable artifacts; (c) Artifacts are removed by guided image filter.

with the original input provided as the guide image. Guided image filter generates the filtering ouptut by considering the content of a guidance image, performs as an edge-preserving smoothing operator like the popular bilateral filter but with better behavior near the edges and more efficient. Guided image filter also mitigate the blurry of the output image since it utilizes the structure of the original input image.

However, is clear that although the guided image filter improves the image quality, it also discards too much information from the translated images. So we only use the guided images filter for those image with serious artifacts.

Results

Both success and fail cases are listed in figure 7 and figure 8.



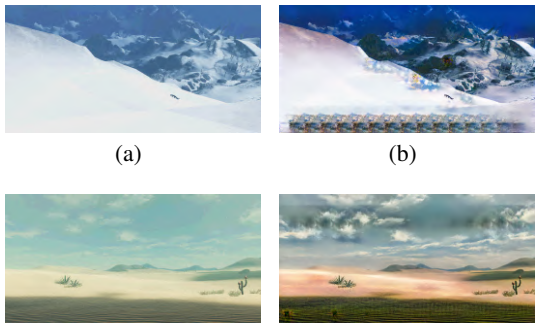
(a) (b)

(c) (d)

Figure 7: Success cases of CycleGAN. Images on the left side are original inputs, the others are translated images.

Conclusion

Compared with many other existing approaches, CycleGAN is the most suitable model for our task because it's relatively light weight than U-GAT-IT, and CycleGAN does not require additional style images as inputs which is necessary for image style transfer based methods. So finnaly CycleGAN model was adopted and improved by us to handle the problem of transfering the screenshots of BOTW into a realisitc style. Quantization followed by Gaussian noise were introduced for preprocessing in order to reduce the prior knowledge taken by the input image, make CycleGAN free to do more creative work. Guided image filter is used



(a) (b)

(c) (d)

Figure 8: Fail cases of CycleGAN. Images on the left side are original inputs, the others are translated images.

for postprocess those images with significant artifacts and makes the translated image more clear.

However, there are still many issues remain unsolved. First, we found that models with normalization layer are easy to suffer from artifacts, however without normalization, the capability of the models will be seriously limited. Secondly, although CycleGAN actually makes the translated image different in color and texture, we have to admit that the translated looks not realistic enough, since the CycleGAN model are not good at change the shape of objects. There is still a gap between our work and the perfect.

References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875* .

Fujun, L.; Sylvain, P.; Eli, S.; and Kavita, B. 2017. Deep Photo Style Transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4990–4998.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2414–2423.

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

He, K.; Sun, J.; and Tang, X. 2010. Guided image filtering. In *European conference on computer vision*, 1–14. Springer.

Li, Y.; Liu, M.-Y.; Li, X.; Yang, M.-H.; and Kautz, J. 2018. A closed-form solution to photorealistic image stylization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 453–468.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* .

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.