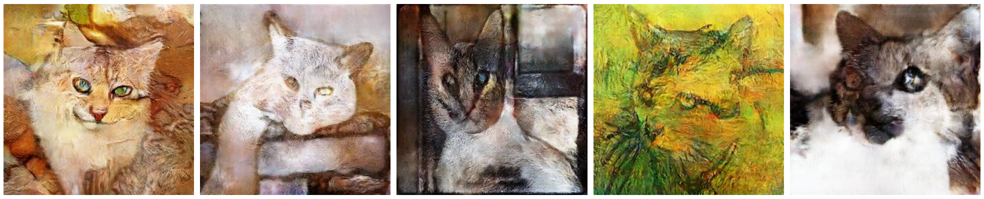
CS230

Everyone could be artist: sketch to style-specified artwork

Fei Jia
CCRMA
Stanford University
jf07@stanford.edu

Qian Dong
Electrical Engineering
Stanford University
qiandong@stanford.edu

Kefan Wang
Electrical Engineering
Stanford University
kefanw@stanford.edu



(a) First (b) Second (c) Third (d) Fourth (e) Fifth

Figure 1: Top five results

Abstract

Sketch-to-image translation is meaningful in real world which provides a powerful tool for people to explore their potential in expressing themselves. Mapping from sketchs to real cat images and Van Gogh paintings separately for content and style is challenging. Therefore, we present a novel approach to learn the translation from a cat sketch to a target image combining cat content and Van Gogh style which is not existed in the real world. Finally, AMT perceptual studies are included to evaluate our generated images.

1 Introduction

Sketch-to-image translation is a promising and powerful tool which has the potential to change people's modes of expression. With the power of deep learning network, artwork and creation won't be limited within a small group of artists; on the contrary, everyone has the possibility to fully explore his art potential and be a talented artist. A simple sketch could be magically converted into a masterpiece only with a bit of guidance provided by the deep learning network. We implemented a modifed Cycle-Consistent Adversarial Network (CycleGAN) in this paper which successfully translate a cat sketch into a cat image with Van Gogh style.

2 Related work

Neural Style Transfer [2, 5] is an optimization technique to combine the content and the style from two images. Instead of learning the relationship between two image categories, the network is only able to capture the mapping between two specific images.

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

StyleGAN is another way to add features from one image to another. In [9], it mixed the features of two images from the same categories and is able to tune the parameters to choose the level of style details from source (e.g. fine, middle, coarse).

Unlike the above work, we want to mix the features from different categories. Specifically, we learn the content from real cat images and the style from van Gogh paintings.

Pix2Pix[4] uses a conditional generative adversarial network (cGAN) to learn the mapping between input and output. cGANs are general-purpose solutions that appear to work well on a wide variety of these problems.

Pix2Pix requires set of aligned image pairs. To use this approach, a generated styled image set needs to be prepared for our task. However, we want to keep our original datasets and let the network extract the content and the style directly.

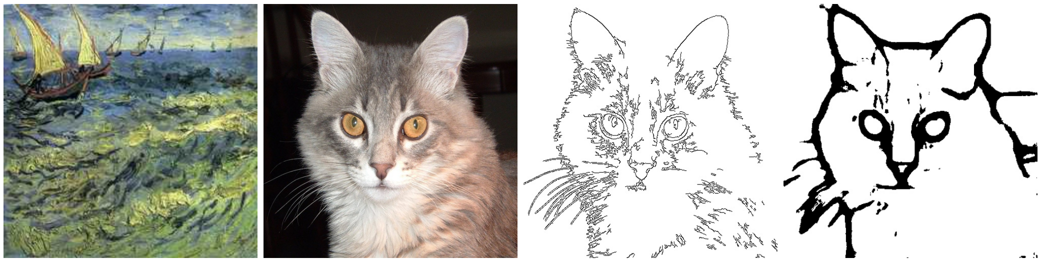
CycleGAN We were inspired by the cycleGAN [3] model because of its unpaired image-to-image translation. By applying cycle consistency loss and identical loss, the network uses transitivity to supervise CNN training. Although CycleGAN is able to find the relationship between unpaired image datasets, it can only be used to map from dataset A to dataset B. Instead of single mapping, our project aims at mapping sketch to both cat and Van Gogh in the respect of content and style.

3 Dataset and Features

Datasets We prepared two datasets: real cat images and Van Gogh paintings. The cat dataset [1] collects many kinds of cats with different poses and locations, having totally over 9000 images. Van Gogh [7] includes 400 images in total.

Data processing We used 800 cat images and 400 Van Gogh paintings for training and 50 cat images for testing. Since images from the cat dataset have different resolution, we resized all the images to 256×256. To accelerate the training speed, we normalized the images as well

Edge extraction Due to the lack of adequate cat sketch dataset with good quality, we used edge extractor to generate sketch data by ourselves. We conducted a popular approach, Canny Edge Detector at first, whose result is shown in Figure 2c. However, the extracted result contained too many details which is not the simple sketch we wanted. Then we explored Holistically-Nested Edge Detection [10] which performed image-to-image prediction by using deep learning model to learn rich hierarchical representations. We used a pre-trained model to predict result as shown in Figure 2d, which ignored most undesired edges and still preserved the contour of the cat making it more like a simple sketch and could be created by anyone.



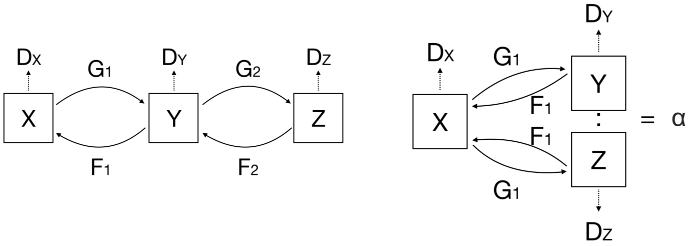
(a) Van Gogh painting (b) Real cat image (c) cat sketch generated by canny (d) cat sketch generated by HED

Figure 2: Dataset samples

4 Methods

4.1 cycleGAN

The architecture of cycleGAN's generator is adopted from [8], which achieves great results in neural style transfer and super-resolution. Discriminator network is a 70×70 PatchGAN, which aims to classify whether the generated images are real or fake. Such a patch-level discriminator architecture has fewer parameters than a full-image discriminator and can work on arbitrarily-sized images in a fully convolutional fashion [4]. Moreover, unlike recent work on “neural style transfer” [6, 8], this method learns to mimic the style of an entire collection of artworks, rather than a single selected piece of art.



(a) Serial

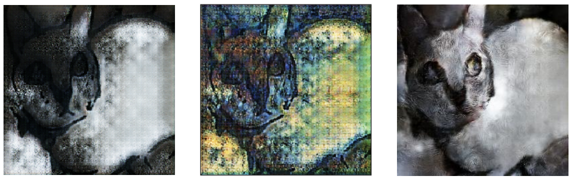
(b) Parallel

Figure 3: Two cycleGAN structure

4.2 Serial cycleGAN structure

We first proposed serial cycleGAN structure, which contains two cycleGANs. The first cycleGAN generates real cat images from edge, and the second cycleGAN translates the real cat images into images with specified style. These two cycleGANs are trained together serially with the same number of epochs. This model didn’t work very well for two main reasons. First, the learning tasks for two cycleGAN have different levels of difficulty. It is much harder to learn to generate real cat images from edge than to generate a image with Van Gogh style. Real cat image generation focuses more on higher-level features including localization of various cat components including eyes, nose, etc., while Van Gogh style generator learns more about lower-level features such as edges, simple textures and color.

Second, generating styled cat image based on the intermediate generated cat image is also a factor that leads to the unsatisfying result. As shown in Figure 4, there is repeated pattern on the cat body in the intermediate cat image generated from edge using the first cycleGAN, and this pattern is preserved in the styled image generated by the second cycleGAN.



(a) Serial mid output (b) Serial output (c) Parallel output

Figure 4: Output comparison between serial structure and parallel structure

4.3 Parallel cycleGAN structure

Since our serial cycleGAN structure wasn’t satisfying enough, we designed a parallel cycleGAN structure which fixed issues in our serial model and greatly improved performance. There are two main ideas behind this design. First, because our goal is generating images that look like both cat and Van Gogh painting, our task can be divided into two sub-tasks, generating real cat and generating Van Gogh styled images. These two sub-tasks can be done with one cycleGAN by sharing the same generators and discriminators in both directions. Considering the different levels of difficulty of two sub-tasks, we proposed loss function to be the weighted sum of the loss of two sub-tasks. The weight is adjustable. For a harder task, which is generating real cat images in our case, we assign higher weight to push the model to learn faster.

Second, with a single cycleGAN, the result will not be affected by any intermediate generated result which might have unwanted patterns that could otherwise be preserved or even amplified in the latter generating stage.

4.4 Objective

The adversarial losses were implemented to all mapping functions between different datasets. In mapping function G(from X to Y) with discriminator D_Y, the equation is expressed as below.

$$L_{GAN}(G, D_Y, X, Y) = \mathbb{E}_y[\log D_Y(y)] + \mathbb{E}_x[\log(1 - D_Y(G(x)))] \tag{1}$$

The cycle consistency loss including mappings from X to Y and from Y to X is

$$L_{cyc}(G, F, X, Y) = \mathbb{E}_x[||F(G(x)) - x||_1] + \mathbb{E}_y[||G(F(x)) - y||_1] \tag{2}$$

In our project, the full serial and parallel objectives are equation 3, 4 respectively.

$$\begin{aligned} L(G, F, X, Y, Z) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + L_{cyc}(G, F, X, Y) \\ +\alpha(L_{GAN}(G1, D_Z, Y, Z) + L_{GAN}(F1, D_Y, Z, Y) + L_{cyc}(G1, F1, Y, Z)) \end{aligned} \tag{3}$$

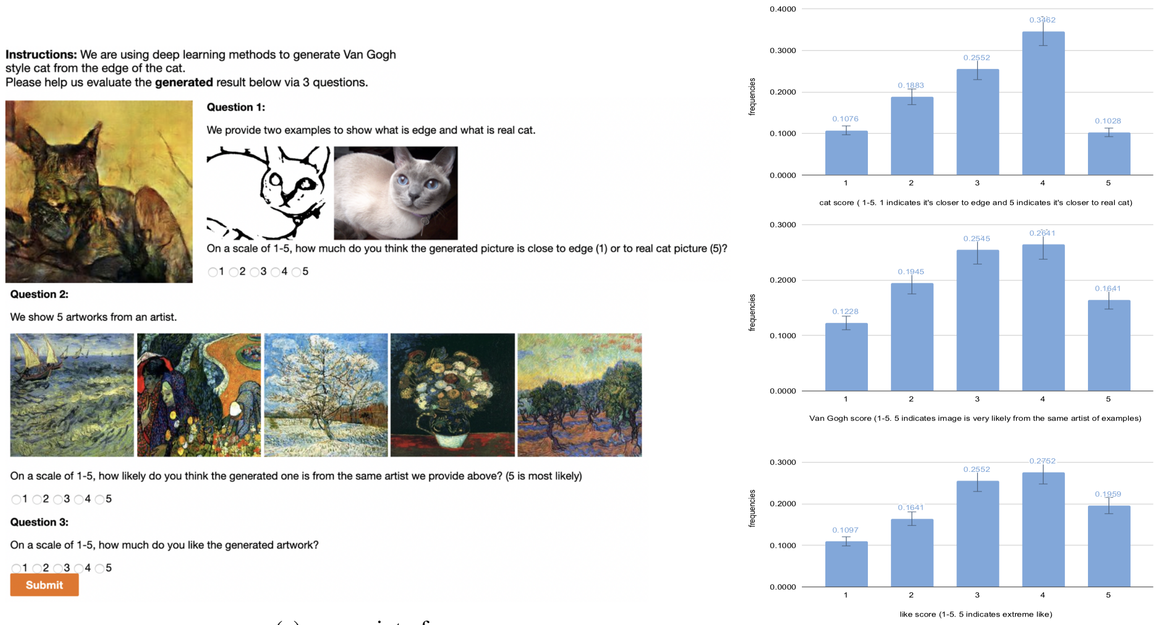
$$\begin{aligned} L(G, F, X, Y, Z) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + L_{cyc}(G, F, X, Y) \\ +\alpha(L_{GAN}(G, D_Z, X, Z) + L_{GAN}(F, D_X, Z, X) + L_{cyc}(G, F, X, Z)) \end{aligned} \tag{4}$$

5 Experiments

5.1 Parallel cycleGAN structure weight

We trained our parallel cycleGAN structure model with 1:1, 2:1 and 4:1 as weighted loss of real cat image generator to loss of styled image generator. Output of development set shows that ratio 2:1 leads to optimal results, which makes sense since generating real cat image is harder than learning style, and higher weight boosts it to learn faster.

6 Results & Discuss



Instructions: We are using deep learning methods to generate Van Gogh style cat from the edge of the cat.
Please help us evaluate the **generated** result below via 3 questions.

Question 1:

We provide two examples to show what is edge and what is real cat.

On a scale of 1-5, how much do you think the generated picture is close to edge (1) or to real cat picture (5)?

○1 ○2 ○3 ○4 ○5

Question 2:

We show 5 artworks from an artist.

On a scale of 1-5, how likely do you think the generated one is from the same artist we provide above? (5 is most likely)

○1 ○2 ○3 ○4 ○5

Question 3:

On a scale of 1-5, how much do you like the generated artwork?

○1 ○2 ○3 ○4 ○5

Submit

(a) survey interface

0.4000
0.3000
0.2000
0.1000
0.0000
frequencies
0.1076
0.1883
0.2552
0.3462
0.1028
1
2
3
4
5
cat score (1-5. 1 indicates it's closer to edge and 5 indicates it's closer to real cat)
0.3000
0.2000
0.1000
0.0000
frequencies
0.1228
0.1945
0.2545
0.2641
0.1641
1
2
3
4
5
Van Gogh score (1-5. 5 indicates image is very likely from the same artist of examples)
0.3000
0.2000
0.1000
0.0000
frequencies
0.1097
0.1641
0.2552
0.2752
0.1959
1
2
3
4
5
like score (1-5. 5 indicates extreme like)

(b) histogram of responses

Figure 5: AMT survey and result of collected responses

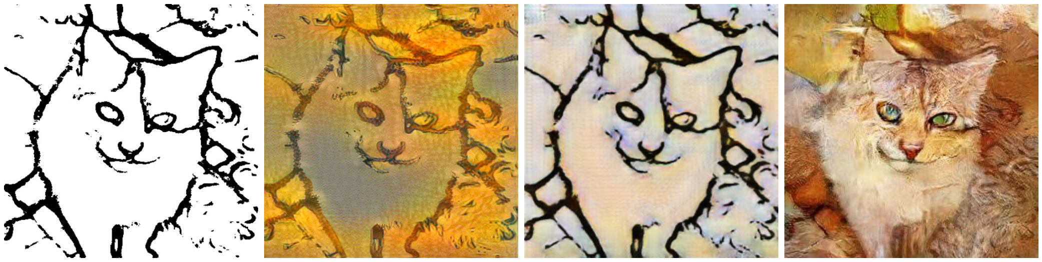
6.1 Evaluation

The goal of our model is to generate an image from the sketch that preserves some characteristic of a real cat but also has style of Van Gogh. It's not a typical Neural Style Transfer problem cause the *style* here is not a single artwork but a broad style which we hope to learn from a bunch of artworks of Van Gogh. Therefore, the evaluation metrics for Neural Style Transfer are not suitable for our task.

AMT perceptual studies We designed to evaluate our models from three dimensions: degree of cat characteristics, degree of style and satisfaction. We generated 50 results and run perceptual studies on Amazon Mechanical Turk (AMT) to qualitatively assess the generated style-specified artwork as shown in Figure 5a.

6.2 Comparison with baseline

We implemented two baseline models: neural style transfer and pretrained cycleGAN. As shown in Figure 6, the baseline models simply fills the edge with color from Van Gogh painting, while our parallel cycleGAN outputs detailed and artistic image.



(a) Input edge (b) Neural style output (c) Pretrained cycleGAN output (d) Parallel cycleGAN output

Figure 6: Output comparison between baseline models and parallel cycleGAN structure

6.3 Our model

We collected 2791 responses from 263 participant for 50 generated images and excluded abnormal responses. Each image has been assigned to 48 to 60 different raters. As we can see from the Figure 5b, considering the entire image set, over 70% people believed the images were closer to real cat (cat score ≥ 3) and 68.3% people considered the images are Van Gogh's work by comparing to several piece of artworks of him (Van Gogh score ≥ 3). Surprisingly, 72.6% people indicated that they like our generated artworks (like score ≥ 3).

7 Conclusion / Future Work

We implemented two baseline models including neural style transfer and pretrained cycleGAN, and designed two creative models: serial cycleGAN structure and parallel cycleGAN structure. Since parallel cycleGAN structure takes level of difficulty of sub-tasks into account and thus pushes different generators to learn at a different rate with adjustable weights, it performs the best to generate images with both real cat details and Van Gogh artistic style. Based on human evaluation results, the images generated by our model are greatly acknowledged and liked by the majority.

In our future work, we would like to try different architectures for cycleGAN generator. In the project, we used U-net for the generator. Since some images in our result still lack of cat details(e.g. cat eyes, mouth), we would like to use pretrained VGG16 network and transfer learning in the next step. We also want to improve our evaluation strategies especially including some quantitative evaluations and extend the participant number in our AMT perceptual studies.

8 Contributions

Fei Jia and Qian Dong and Kefan Wang brainstormed ideas and proposed models together. Fei Jia is responsible for data prepossessing, evaluation designing and deployment, as well as result analysis. Qian Dong is responsible for model comparison, training and optimization. Kefan Wang is responsible for collecting data, training model as well as model comparison. We together complete the report and poster.

9 Acknowledgement

We would like to give our special thanks to professor Andrew Ng and Kian Katanforoosh for inspiring lectures, our mentor Sarah Ciresi for valuable suggestions on project, and all the course staff for this wonderful course. We also appreciate the good words and useful feedback from MTurkers.

References

[1] Chris Crawford. Cat dataset, Feb 2018.

[2] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[5] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 2019.

[6] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[7] Junyanz. junyanz/pytorch-cyclegan-and-pix2pix, Nov 2019.

[8] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):119, 2016.

[9] Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. In *European Conference on Computer Vision*, pages 318–335. Springer, 2016.

[10] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. *CoRR*, abs/1504.06375, 2015.