From Human face sketches to realistic photographs
January 22, 2019

Hind Dadoun
hind.dadoun@gmail.com

Raphael Montaud
raphael.montaud@gmail.com

Abstract

Generative adversarial network (GAN) has shown great results in many generative tasks. However, it is still rather challenging to train a GAN model due to training instability or failure to converge. In this report we focus on the particular case of image-to-image translation, more specifically, translating human-drawn sketches to realistic photo portraits. This problem is rather interesting because of an additional challenge: the lack of clean databases. To tackle this problem we apply the techniques of Conditionnal GAN ([3]) and Cycle GAN ([7]) to different datasets. To evaluate the quality of resulting images we start by a visual analysis and then we use a proxy evaluation metric named Inception score.

1. Introduction

In this report we set A to be the distribution of face sketches and B the distribution of face photographs.
In [3], the authors exposed a new method named 'Conditionnal GANs' for image translation, it is a variant of the original GAN that adds additional information to inputs of the generator and discriminator.Indeed, it learns a mapping from observed image X and random noise vector Z to Y, $G : X, Z \rightarrow Y$. Therefore it enables to train generative networks that can create an image in B from an image in A, but it must be fed with paired examples.
In [7], the authors propose an alternative method named 'Cycle GANs', that can learn from unpaired images. The goal is to learn a mapping $G : X \rightarrow Y$ such that the distribution of images from $G(X)$ is indistinguishable from the distribution Y using an adversarial loss. They do so by coupling the first mapping with an inverse mapping $F : Y \rightarrow X$. We propose to experiment both methods with different datasets.

2. Datasets

In this section, we describe three different datasets that we used in our project.The first data set is the Chinese University of Hong Kong database which contains 200 paired images of faces and human-drwan portraits [5].The drawings are realistic and we have the real picture to compare our results with.
The celebA dataset is composed of images of celebrity faces. It is cropped to be centered around the face of the person. In addition it is very clean and has varied background in comparison with the CUHK dataset. The size of the database is a big advantage as well as it allows us to study varied face attributes. If training goes well it can be applied to a wide range of human sketches. We also propose an algorithm to transform those pictures with a "pencil-like" filter. Hence we can obtain paired examples of sketches and photographs. This filter consists on converting the RGB image to a gray one then applying a Gaussian Blur and finally dividing the gray image by the blured image. This way, only contours are kept in a pencil sketch fashion.
Finally, we scrapped images of drawings and photographs on Flickr. The images are not always well tagged on Flickr so images dit not always match our queries.

3. Experiments

You'll find here [1] all the code used for the experiments. We start by creating the 'benchmark' set which includes images from the three datasets. The goal is to see how well the methods perform when tested on a different set than the one used in training. Moreover to be able to compare both conditional GANs and the cycle GAN methods we test them on the CUHK and CelebA datasets. The reason we test them on two different datasets is to see if performance varies with the size of the database(200 vs 200 000), it's diversity (there are only chinese people in the CUHK dataset) and the quality of sketches (human-drawn vs pencil-like). With the Flickr data set, we can only train a cycle GAN since the data in unpaired. We try to train a model from scratch, but we also try a tranfert learning approach where the generators and discriminators are initialized with the ones obtained after the Cycle Gan training on CelebA.

[1] https://github.com/raph-m/pytorch-CycleGAN-and-pix2pix/blob/master/README.md

4. Results

We present the results in Appendix A.

- In Figure 1 we test the models on a drawing from the CUHK dataset. The best performance is obtained for the CUHK Cycle and celebA Cycle models. It should be noted that there is a bias in the results. In the CUHK model, the background is blue (since this is the case in all the training set). Also in the celebA cycle model, the eyes have been colored in blue, which probably comes from the training set too. Other models perform poorly.
- In Figure 2 we test our models with an image from the Flickr drawings distribution. Visually, the best result comes from the pretrained Flickr model. Since it performs better (if not better, they are "not worse") than the celeba Cycle and the Flcikr Cycle, it suggests that the transfer learning is a success. The models that were not trained on the Flickr distribution perform quite poorly.
- In Figure 3, once again the best results come from the models that were trained on this type of data.
- In Figure 4, we test our models against an "automatic sketch" generated from an image in the celebA dataset. We get outstanding results from the celebA conditonal GAN model. It could almost fool a human. Just like other examples, the models that were not trained on this data perform poorly. The cycle GAN - celebA model performs very poorly too, we will discuss the difficulties encountered during the training in 5.
- Finally, in Figure 5, we wanted to show an example of the celebA conditional model performing poorly (this is unusual).

5. Notes on training

As we said, we encountered some issues during the training of the Cycle GANs. Notably, during the training on the celebA dataset, we found that the cycle reconstruction is quite well respected, as we can see in Figure 6, but the generated photographs are not realistic. When noticing this, we decided to change the parameters in the loss in order to make the cycle loss approximately 10 times smaller than the GAN loss, in the hope that it would allow the generator to be more "creative" and generate more realistic images. This did not work and we observed that the GAN loss oscillates a lot. Also, we notice that strange artefacts appear on the generated drawings and pictures. In Figure 7, we show another example where the generated drawing does not correspond to what we expect it to be. To conclude, we found it hard to train the Cycle model and the heavy computational time of the cycle method makes it harder to experiment different settings.

6. Quantitative analysis of results

To evaluate the quality of resulting images we used the Inception score metric as introduced in [4]. The idea is to train a classifier on the distribution B and then process the generated images through this classifier. According to this note [2] the inception score uses two criteria in measuring the performance of generative models, the class predictions should be:

- Diversified. That is to says that the generated images must belong to various classes according to the classifier.
- The classifier should be sure of his results. This means that the generated image is of good quality and that the classifier is able to make predictions on them. This translates into a great variance in the predictions for a class (better to have only 0s and 1s than only 0.5 values)

Thus, we want the conditional probability $\mathbb{P}(y|x)$ to be highly predictable (low entropy). So we use an Inception network to classify the generated images and predict the label of x. This gives us an idea about the quality of the images. To measure the diversity of images we check if the data distribution for y is uniform (high entropy). $\mathbb{P}(y)$, the marginal probability is computed this way:

$$\int_z \mathbb{P}(y|x = G(z))dz$$

Finally to combine those two criteria we compute their KL-divergence and use the equation below for the inception score(IS):

$$IS(G) = exp\left(\mathbb{E}_{x\sim p}[D_{KL}(\ \mathbb{P}(y|x)\ ||\ \mathbb{P}(y))]\right)$$

If we denote $p = \mathbb{P}(y|x)$ and $q = \mathbb{P}(y)$ then :

$$D_{KL} = \int_X p \times \log \frac{p}{q} d\mu$$

In our opinion this metric should not be used as an absolute metric but rather as a relative metric to compare several generators (for example to find the best parameter during a grid search).

In order to adapt the inception score to our problem, we trained a classifier to predict the gender on the celebA dataset (cf [1]). We then process 10,000 images of the celebA dataset through 3 generators and then process the images through the classifier. The results are presented below:

Inception Score	Mean	Standard deviation
Cycle Gan(5 epochs)	1.33	0.01
Conditional Gan(5 epochs)	1.69	0.015
Conditional Gan(10 epochs)	1.65	0.014

The results suggest that the conditional GAN is better than the Cycle GAN and this correlates with our visual appreciation. Also it looks like training 5 more epochs on the conditional GAN did not improve the results. Visually results are quite close and it is delicate to prefer one or another.

Following the logic of the inception score, best images should be the ones with high probability of being a male or a female. And the worst images should be the ones where the classifier is unsure about the gender. We present in Figure 9 the best and worst images for different generators. It appears that the "best" images are actually the ones where the gender features are obvious (make-up and long hair for women) and the worst are the ones where the gender features are more tricky (men with long hair, women with short hair, etc). To conclude, we are unsure about the relevance of the inception score. We might have more convincing results by training a classifier on all attributes available on the celebA dataset. We can also imagine that since we have the real attributes available we could create a score based on the accuracy of the classifier on the generated images.

7. Conclusion and further work

In this report we investigated the conditional and cycle GAN techniques on various datasets. We found that it can perform well to generate realistic photographs on the CUHK dataset, but the generation performs poorly when the input sketches come from antoher distribution (eg: Flickr). Therefore, we tried to implement a cycle GAN model on the Flickr data, but it fails to produce satisfying results. This comes perhaps from the lack of data. To overcome the lack of data, we proposed to pretrain the model on the celebA dataset and to finetune it on Flickr data. Results seem to improve with this transfer learning technique but are still not realistic. Also the models fail to add features to make the resulting images more realistic (they mostly add color to the drawing). Finally, the results of the conditional GAN on the celebA dataset are very realistic, but they do not perform well on real drawings. This suggests that a paired dataset like CUHK with more data and various conditions (background, lighting) would perform well on real drawings.

In further work, we should explore the possibilities of transfer learning, maybe by rethinking the "pencil drawing" filter. Also we should try to implement more state of the art techniques to improve the cycle GANs performance, like historical averaging and virtual batch normalization (cf [6]).

References

[1] Kaggle kernel: gender detection on celeba dataset.

[2] S. Barratt and R. Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

[3] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976. IEEE, 2017.

[4] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

[5] X. Wang and X. Tang. Face photo-sketch synthesis and recognition, 2009.

[6] L. Weng. From gan to wgan. 2017.

[7] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017.

Appendices

A. Testing results

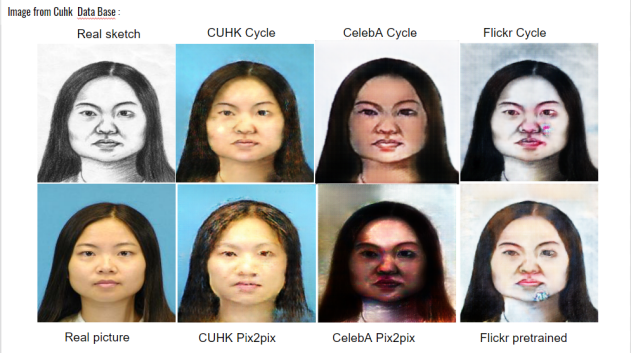

Figure 1: Test of the models on a drawing from the CUHK dataset

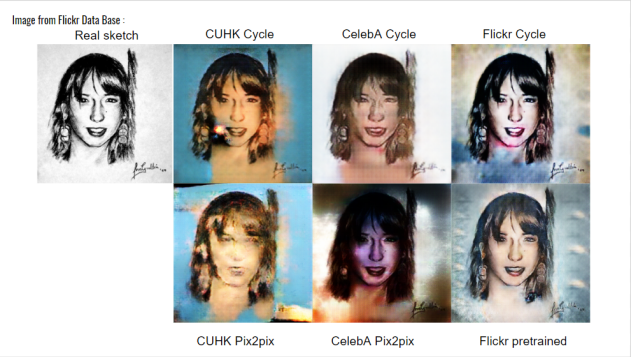

Figure 2: Test of the models on a drawing from the Flickr dataset

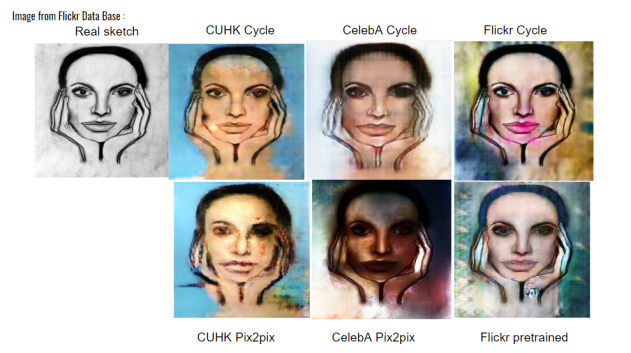

Figure 3: Test of the models on a drawing from the Flickr dataset

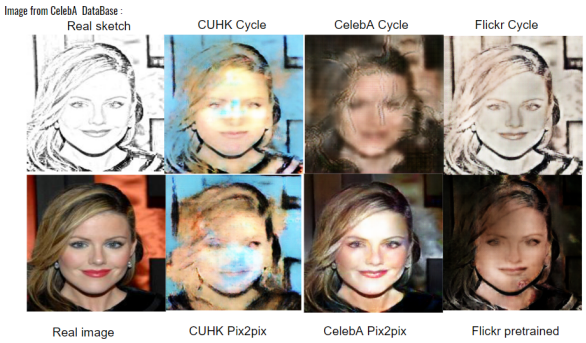

Figure 4: Test of the models on a drawing from the CelebA dataset

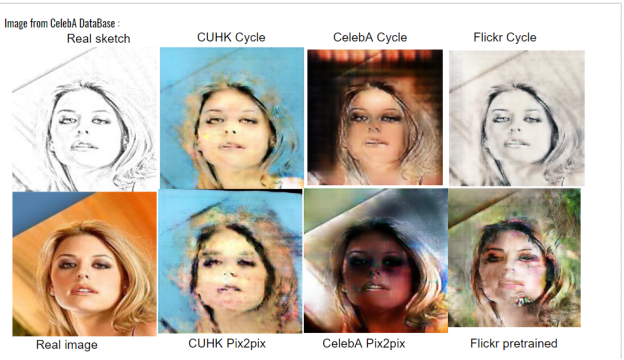

Figure 5: Test of the models on a drawing from the CelebA dataset

B. Training results

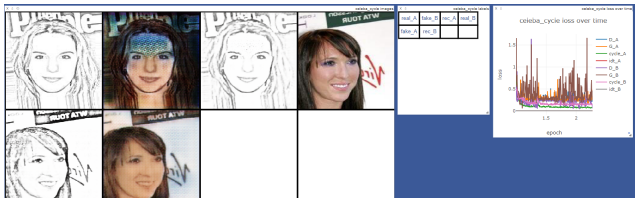

Figure 6: Visdom outputs for Cycle Gan training at epoch3

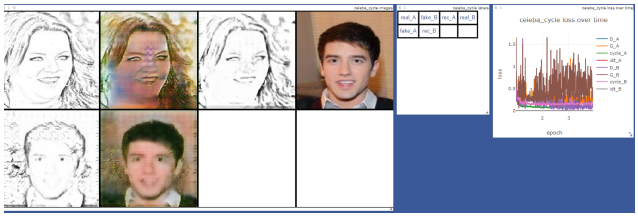

Figure 7: Visdom outputs for Cycle Gan training at epoch4

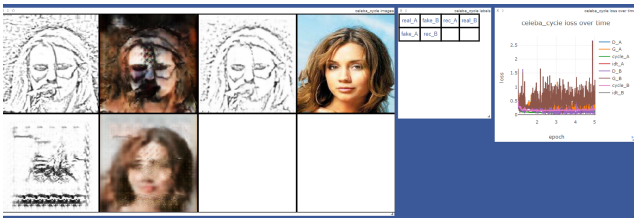

Figure 8: Visdom outputs for Cycle Gan training at epoch5

C. Inception score results

Figure 9: Inception score: best images on the right and worst images on the left