Face2Anime

Valeriy Girkin, Anton Streltsov
Faculty of Computer Science
National Research University Higher School of Economics
Moscow, Russia

***Abstract*—Style transfer is a task of applying the style of a particular image to another image while preserving its content. In this work, we try to solve a more complex problem: given two unpaired sets of images, we want to learn a mapping which could shift an image from one domain into another. For that purpose, we use latest developments in generative adversarial networks such as CycleGAN and XGAN. We start with the review of related work and methods then describe our datasets. Finally, we define our neural network.**

Keywords—generative adversarial networks; style transfer; unsupervised; machine learning;

Introduction

Background of the study. One of the main long-term goals of style transfer is to shift the style of one picture, called reference image, onto another picture, called content image. For example, one can achieve getting pictures of different illumination, weather or time of the day or completely another style while preserving the object depicted in the original image by choosing the right style picture. The introduction of generative adversarial networks [4] gave way to many improvements of results of computer vision problems, which include changing emotions [1], colourising sketches [2] and photo styling [3]. However, the architectures vary greatly from work to work, heavily depending on the task at hand. Therefore, we have ample opportunities for experiments.

The problem statement. Here we tackle the task of image-to-image translation, converting an image from one representation of a given scene to another. In particular, we plan to transfer styles of anime characters' pictures and of real people's photos between each other. The main challenge is the absence of any correspondence between these two image collections. Also, we aim to keep different facial attributes such as hair, eye colour, emotions and head orientation. To achieve this, we will need to devise a way to preserve semantic content rather than pixel-level details. To generalise, let X be the domain of real-world photos and Y be the domain of anime pictures. Then we want to train a function $G_{xy} : X \to Y$ which generates pictures $G(X)$ that are indistinguishable from Y both in appearance and adversarial loss. To constrain this function further we also introduce an inverse function $G_{yx} : Y \to X$ and enforce $G_{yx}(G_{xy}(X)) \approx X$ for consistency.

Professional significance. Unsupervised semantic style transfer has recently gained a lot of popularity and plenty of new solutions have been proposed. The resulting model is anticipated to be suitable for applying to a diverse set of style transfer tasks. Although this objective might seem to be used mostly for entertainment purposes, it can also be used in communication technologies [5] and as a baseline in art projects.

Literature Review

Style transfer is a subject of many recent papers in machine learning community. In global style transfer algorithms an image is processed by applying transfer function. Although these methods are effective, they can handle only simple styles like global colour shifts and tone curves. First approaches involving deep learning used convolutional neural nets [6]. They had a fixed pair of a content c and a style s images, an input image x initialised with white noise and a VGG network [7], trained to perform object recognition. One iteration consisted of feeding the images to the net reducing the difference of c and x representations from shallow layers of the net and reducing the difference of s and x representations from deeper layers. In the end, they got an image x_r which depicted the same scene as c while having the same style as s. However, this approach uses one style images at a time and performs poorly if given a collection of style pictures, mostly due to the necessity of manually averaging results.

Later, synthesising natural images has become one of the most intriguing and challenging tasks in graphics, vision and machine learning research. Nowadays, application of generative frameworks is a more common approach to unsupervised learning. GAN methods train a network that synthesises samples from a target distribution from white noise images called generator network G. This network is trained simultaneously with another network known as a discriminator D. Its goal is to distinguish generated by G images from samples from the training that are drawn from the target distribution. Thus the task of G is to generate images which force D to make mistakes and classify them as original samples.

Another approach [8] uses a new GAN-based model for unpaired image-to-image translation — CycleGAN. The formulation of this paper's problem is quite similar to ours. The authors achieved good results in transferring colour and texture details, but the method does not work well if the contents of images are too different. The reason for that might be that they use only pixel-wise semantic consistency and reconstruction losses.

[9] tries to mend these limitations and introduces semantic consistency and domain-adversarial losses in their XGAN

model. A shared representation of the two input sets is learned in this unsupervised dual adversarial auto-encoder, which then helps to restore most of the content specific features in both domains. The main drawback of this method is the huge number of task-specific optimisations and parameter tunings, which makes it less reproducible.

Methods

Generative adversarial networks. GAN is a network consisting of a generator network G and discriminator network D. The task of G is to generate a picture from the target distribution given a noise image as an input. At the same time, D is supposed to distinct generated images from the original ones. This lets us introduce adversarial loss which is used when the quality metric cannot be explicitly defined.

CycleGAN. It is composed of two pairs of G and D related to two domains A and B accordingly. At each training step the model is trained in two directions: a picture X_A from A is fed into the model so that we get $D_B(G_{AB}(X_A))$ and $G_{BA}(G_{AB}(X_A))$ using which we can count adversarial and reconstruction losses for the generator; and discriminator loss from $D_B(G_{AB}(X_A))$ and make optimisation step for $G_{AB}(X_A)$ and $D_B(G_{AB}(X_A))$. The process of training in other direction is similar. Also, the authors used labels when they were available, that gave a considerable increase in resulting quality. Unfortunately, we do not have the labels due to the fact that we have to gather one of the datasets manually.

Distances between distributions. It is usually not possible to match the input with the output sample in unsupervised problem setup, so most of the developed methods rely on the measure of distance between generated distributions.

Semantic consistency loss. The results of generative models trained by minimising mean squared errors are often blurred. The possible reason for this is that such model does not take into account the patterns of high-level features but only consider pixel differences. This loss is introduced in [9] and it encourages the embeddings learned by encoders parts of generators to lie in the same subspace. That is achieved by minimising the distance between the encoding of the original image and the encoding of the picture generated from this image.

$$L_{semAB} = \frac{1}{batch\_size} \sum \|e_{AB}(X_A) - e_{BA}(G_{AB}(X_A))\|$$

Domain-adversarial loss. [10] contribution is a domain-adversarial loss, which also tries to force embeddings to lie in the same subspace. For that purpose an extra discriminator D_{dom} is used. Its purpose is to maximise the following loss:

$$L_{dom} = mean((D_{dom}(e_{AB}(X_A)))^2 + (D_{dom}(e_{BA}(X_B)) - 1)^2)$$

At the same time, the generators' encoders are supposed to minimise this loss.

Perceptual loss. An additional trained recognition convolutional neural network is fed with generated and original images to extract features after different layers to calculate the difference between them. This method brings substantial improvements to generated images. However, it is quite important that the net is trained on data similar to yours. And we were not able to find a pretrained CNN for anime pictures, thus we might be forced to train such network ourselves.

Wasserstein critic. Wasserstein critic [11] is a powerful and mathematically justified substitution for discriminators. It is usually hard to tell from the loss of discriminator if the training process goes well. Wasserstein critic's loss is more interpretable, however, it is not suitable for all tasks and requires a lot of tweaking.

Results Anticipated

We will use CelebsA dataset for photos of real people because it has aligned and preprocessed faces. It also has several labels, which can be used to improve the generation of images from anime into real ones. As for the anime dataset, we found only one diverse collection — Danbooru Faces, but its problem is that the vast majority of it is faces of females. Despite this fact, we consider the size of both datasets adequate, even after removing most of male photos from CelebsA collection. Still, crawling of anime pictures represents a non-trivial task, as systemised sources of them are scarce and combining several sets might lend a problem of repeated images.

At first, we plan to implement a basic CycleGAN architecture, but it can be predicted that the results will not be satisfactory. That is highly probable because this architecture is more fit to the task of transferring only the style without keeping the semantics. Afterwards, we will experiment with different mentioned losses, starting with domain adversarial, as it seems to be the most promising addition to the model. Moreover, we would like to make the final model as simple as possible, so not all of the mentioned methods might be used. It is expected that the model will generate both high-quality anime and real pictures that will be indistinguishable from their originals.

Conclusion

In this paper we showed the motivation and ways of solving our stated problem. Then we presented a lot of existing methods and network architectures from related works. Furthermore, we described many ways and solutions for semantic style transfer as well as proved that this problem can be solved. Also, we showed that there are plenty of opportunities for experiments with the proposed model. Our final model is also expected to solve other general tasks including all those that were mentioned in previous parts of this work.

References

[1] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, 2017.

[2] Wenqi Xian, Patsorn Sangkloy, Varun Agrawal, Amit Raj, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Texturegan: Controlling deep image synthesis with texture patches, 2017.

[3] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer, 2017.

[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[5] Xiaohan Jin, Ye Qi, and Shangxuan Wu. Cyclegan face-off, 2017.

[6] L. Gatys Image Style Transfer Using Convolutional Neural Networks

[7] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs], Sept. 2014. arXiv: 1409.1556

[8] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2017.

[9] Am´elie Royer, Konstantinos Bousmalis, Stephan Gouws, Fred Bertsch, Inbar Mosseri, Forrester Cole, and Kevin Murphy. Xgan: Unsupervised image-to-image translation for many-to-many mappings, 2017.

[10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. Journal of Machine Learning Research, 2016.

[11] Martin Arjovsky, Soumith Chintala, and L´eon Bottou. Wasserstein gan, 2017

Word count 1533