INFO8010: Project proposal

Antoine Debor[1] and **Pierre Navez**[2]
[1] *antoine.debor@student.uliege.be (s173215)*
[2] *pierre.navez@student.uliege.be (s154062)*

I. PROJECT CONTEXT

Given the sanitary context, more and more students are struggling to cope with the monotony of distance learning. Being in front of the screen all day, isolated from the campus and other students, seems to have a detrimental effect on motivation and, from a general point of view, on mental health. From then on, we thought that using deep learning to brighten up the current routine of students could be an interesting starting point for this project.

Most of the courses are currently given via different platforms like *LifeSize*, *WebEx*, *Zoom*, etc., and most of them are recorded. Therefore, our initial idea would be to perform style transfer on these podcasts in order to make the speech of our teachers funnier, in a way. This style transfer would be done from the style of different great painters, as it has already been done for photos or film excerpts for example. The idea would be so to be able to watch one's favorite course (INFO8010, of course) while choosing to see the teacher evolve in, for instance, a Picasso-ized universe. Our approach considers the general style of a painter rather than focusing on one of its masterpieces.

Aware that such a project may be too ambitious or too laborious to put in place, we may need to simplify it and focus initially on short and simple video clips, or even on still images (which is less funny, but maybe more in our wheelhouse).

II. RELATED WORK

This idea is inspired from the paper "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks" from Zhu et al. [1]

III. TECHNIQUE

The method that will be used to perform such style transfer will rely on what is called *CycleGAN*. In order to understand what is a *CycleGAN*, let's explain what a *GAN* is. A Generative adversarial network (*GAN*) is a neural network architecture that was first described in a 2014 paper written by Ian Goodfellow et al. [2] It consists in a zero-sum game between two neural networks, one being called the generator an the other the discriminator. The basic idea is that from a random input vector the generator will generate a sample in the domain. This sample will be fed into the discriminator model, that will classify this example as a fake or real sample compared to the training set samples. The performance of the discriminator will be used as a feedback for the generator as well as the discriminator in order to adjust their parameters if needed. The generator can be seen as a forger that would create fake bills and would try to put them into its bank account. The discriminator would be the banker which would analyze the money at the counter and accept it if the money seems real. If the banker does not accept the bill, the forger will improve its method to create fake money until he can fool the banker. Similarly, when the banker will eventually see that he has been scammed, he will ameliorate its manner to detect fake bills. This architecture is described graphically on FIG. 1.

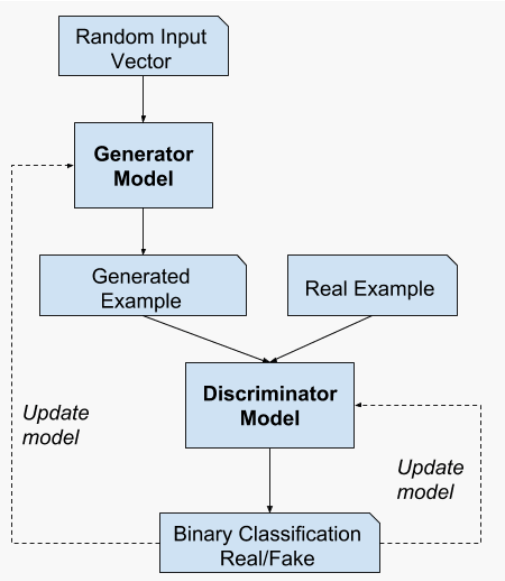

FIG. 1. Generative adversarial network

Now that *GAN* are understood, one can go one step further and explain how this architecture can be adapted to our problem. The goal of this project is to apply a style transfer from paintings to some images. Let's consider that one has a set of Monet paintings at its disposal and, in addition, at set of real pictures. The principle of a *CycleGAN* network is to have one generator-discriminator pair to apply Monet style on the real pictures, along with another generator-discriminator pair which would perform the inverse transformation.

Formally, our goal is to learn a mapping function between two domains X and Y given training samples $\{x_i\}_{i=1}^{N}$ where $x_i \in X$ and $\{y_j\}_{j=1}^{M}$ where $y_j \in Y$. The data distributions are denoted $x\tilde{p}_{data}(x)$ and $y\tilde{p}_{data}(y)$. The model includes two mappings $G : X \to Y$ and $F : Y \to X$. Additionally, there are two discriminators D_X and D_Y where D_X aims to distinguish images x from translated images $F(y)$ and D_Y aims to distinguish images y from translated images $G(x)$. The objective contains two types of error to minimize: the *adversarial losses* and the *cycle consistency losses*

The adversarial loss enforces the matching between the distribution of the generated images and the target domain. For the mapping function $G : X \to Y$ along with the discriminator D_Y the objective is expressed as:

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

In the context of a zero-sum game between G and D_y, the generator tries to minimize the objective that the discriminator tries to maximize. The same loss exists for the mapping function $F : Y \to X$ along with D_X.

The cycle consistency loss is introduced to ensure that the mapping derived from the first objective not only corresponds to a similar data distribution from one domain to another, but is also corresponding visually. Indeed, with networks of large enough capacity, a mapping derived from the adversarial loss can correspond to the data distribution of the target domain but not provide a good qualitative result. The idea here is to say that the mapping should be cycle consistent. If we apply a transformation from image x from domain X to image y in domain Y, then transforming back the image y should provide the image x as well. Formally, $x \to G(x) \to F(G(x)) \approx x$ and the same for y and $G(F(y))$. So the cycle consistency loss is expressed as

$$\mathcal{L}_{cycle}(G, F) = E_{x\sim p_{data}(x)}[||F(G(x)) - x||_1] + E_{y\sim p_{data}(y)}[||G(F(y)) - y||_1] \quad (2)$$

Finally, the overall objective function is expressed as the sum of the objectives with a parameter λ, encoding the importance of the cycle consistency besides the adversarial losses

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) + \lambda\mathcal{L}_{cycle}(G, F) \quad (3)$$

To conclude, dealing with images, it is clear that the network architecture will be constituted of convolutional neural networks.

IV. DATASET

The painting dataset would consist of canvas images labeled with the name of the artist. We think of using the Best Artworks of All Time dataset available on the *Kaggle* platform. This dataset contains three files: *artists.csv*, containing information for each artist (from *Wikipedia*, not relevant for our project), *images.zip*, collection of images (full size), divided in folders and sequentially numbered, and *resized.zip*, same collection (8255 images) but images have been resized and extracted from folder structure (using this file allows to download less data and process faster our model). This dataset contains paintings from the 50 most influential artists of all time, and is 2.16 GB big. A few other similar datasets exist on *Kaggle*, but do not all provide the artist labels along with the paintings.

In order to apply a style transfer to a video podcast of a lecture, we think that this is reasonable to use a face image dataset at the other hand. In particular, the UTKface dataset contains over 20000 face images of various age, gender, pose, facial expression, etc. We would consider only a subset of this whole dataset which is way too big for our problem.

V. COMPUTING RESOURCES

As we will have to deal with large datasets of images, we plan to use Google Colab's free GPU to speed up the computations during the training of our network, as it has been done during the second homework.

VI. NICE-TO-HAVES

We do not know if our main goal is reasonable considering our skills and the time allowed for this project. However, a first additional idea would be to incorporate data from contemporary artists such as street-artists, for instance, and see how this kind of data interacts with videos. A second idea would be to perform real-time style transfer, in order to switch from one painter's style to another during a live lecture.

[1] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

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