Understanding GANs with xAI

NGUYEN Le Hoang* AGUILA–MULTNER Adrien† CHAUVIN Etienne‡

21 April 2023

1 Introduction

1.1 Background

As part of the CSED/AIGS538 - Deep Learning course at Postech (Pohang University of Science and Technologies), we will be carrying out a term project in which we will try to demystify a puzzling phenomenon. This proposal report is an introduction to it and lays its groundwork and objectives.

Our term project aims to better understand the internal functioning of Generative Adversarial Networks (GANs). A GAN is a deep learning model able to produce new and artificial data that is comparable to a given dataset on which the model has been trained. A generator and a discriminator make up the model. The generator generates new data from random inputs while the discriminator is simply a classifier that tries to distinguish real data from the data created by the generator. In the end a GAN is supposed to be able to generate data that is indistinguishable from the original dataset, so that the disciminator fails in its task (assuming it is efficient).

We will be working on an off-the-shelf model generating movie posters, trained on an IMDb public and large dataset.

1.2 Motivation/Problem

Movie posters play a significant role in generating interest and promoting films. However, creating posters that accurately capture the mood and genre of the movie can be a complex task. By using GANs, we can create visually appealing posters from existing one that match the statistical distribution of the original posters. Ultimately, this project can provide valuable insights into the use of GANs in the movie industry and their impact on poster design.

*Department of CSED, POSTECH, South Korea; e-mail: member@postech.ac.kr
†Department of CSED, POSTECH, South Korea; e-mail: aaguilamultn@postech.ac.kr
‡Department of CSED, POSTECH, South Korea; e-mail: echauvin@postech.ac.kr

1.3 Related Work

1.3.1 GANs

Generative Adversarial Networks (GANs) have been shown to be effective in improving image classification accuracy. A study compared the performance of GAN models with different architectures on image generation and image classification tasks, with the GAN model using convolution layers outperforming other models in both tasks (1). It demonstrates the potential of GANs in improving image classification accuracy and suggests that GANs can be a valuable tool for image processing and analysis.

1.3.2 Explainable AI(XAI)

Explainable AI is an important area of research in machine learning, which seeks to provide insights into how models make decisions. In deep learning, explainability is particularly challenging, as models can be highly complex and difficult to interpret. Therefore, the high accuracy of these complex models has also raised concerns about how the machine making decision. A research (2) conducted by Vinay Jogani team on the application of XAI in medical domain serves as a valuable example of how explainability can be achieved in complex deep learning models. Their work highlights the importance of interpretable AI systems in the medical field, and the need for further research in this area.

2 Main Idea

Our goal is to analyze the inner knowledge and decisions of a convolutional Generative Adversarial Network. The task of this model will be to generate new images from random noise, after being trained on a dataset of IMDb movie posters. Once satisfactory results are obtained, we will analyze both the discriminator and the generator:

- For the first one, we will try to understand how does it take a decision ; what is the information extracted from every filter and how does the information is encoded into the feature space. In order to achieve this, we will look for patterns or similarities in the neurons' activation map between similar input images ; semantically or visually.
- For the generator, we will study how the initial noise is interpreted, and what transformations are applied by each layer. We will look at how the generated image varies while changing values of the input noise, one by one. We will also try to visualise different channels from different hidden layers' feature map to see the complete evolution.

This analysis can be done at different epochs during training, to evaluate how both networks evolve.

3 Expected Results

First of all, our model should be able to generate plausible new movie posters, even though they may not be of a very high quality. In the early stages, the discriminator will probably decide based on the smoothness of the image, maybe with Sobel-like filters to detect edges which won't

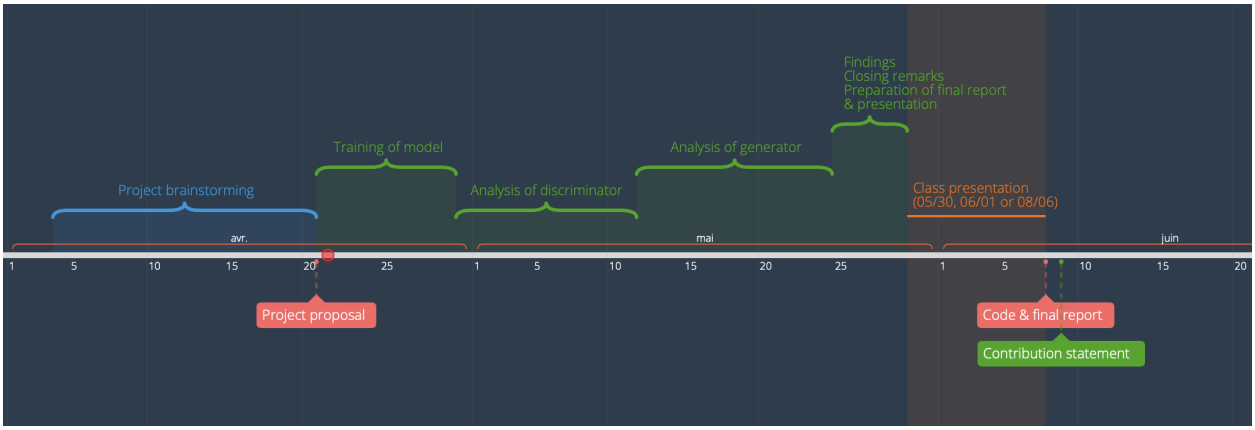
appear in the noise created by the generator. Meanwhile the generator will just apply random transformations. After the training process, we expect the discriminator to learn more complicated filters to extract informations about specific patterns that differentiate the real images and the generated ones. We also suppose that the generator will learn to map the latent noise to features of the image ; for instance one can represent the overall brightness level, another one the average color of the background.

4 Plan

Our plan is organized as follows:

1. First, we will find an off-the-shelf efficient GAN model, understand it and train it;
2. Then we will analyze the GAN discriminator in order to understand how decisions are made;
3. Next we will analyze the GAN generator to see how inputs are interpreted and how is the evolution through the layers;
4. Afterwards, we will pool all the results and findings to conclude our project, and start preparing our presentation and final report;
5. Finally we will do the presentation, write our final report and our statements of contribution.

For a better visualization of our schedule, you can have a look at our timeline:



References

[1] Meng, H. and Guo, F. (2021). Image classification and generation based on gan model. In *2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*. 180–183.

[2] Purohit, J., Shivhare, I., Jogani, V., Attari, S., and Surtkar, S. (2023). Adversarial attacks and defences for skin cancer classification. In *2023 International Conference for Advancement in Technology (ICONAT)*. IEEE. https://doi.org/10.1109%2Ficonat57137.2023.10080537.