Understanding GANs with xAI

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

08 June 2023

Abstract

This report explores the use of Generative Adversarial Networks (GANs) for generating random movie posters and employs explainable AI (xAI) techniques to gain insights into the inner workings of the model. The project aims to provide valuable insights into GANs' application and their impact on poster design. Furthermore, this study also highlights the challenges and potential of using GANs for movie poster generation.

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: lehoang@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 (2). 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 (3) 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:

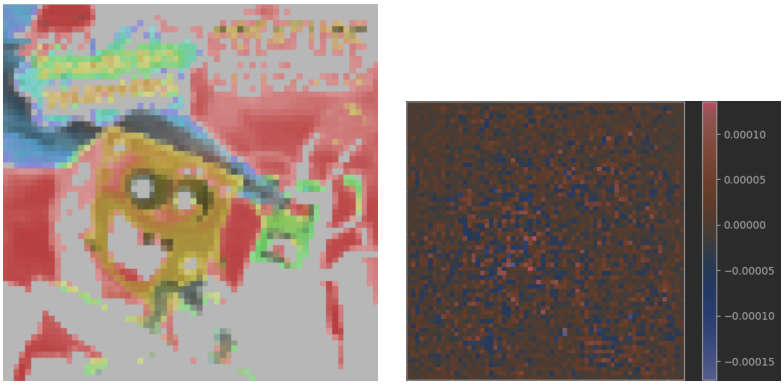
- For the first one, we will try to understand how does it take a decision ; what is the information extracted from every filter and how does the information is encoded into the feature space. In order to achieve this, we will look for patterns or similarities in the neurons' activation map between similar 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.

3 Results

3.1 State of the art tools

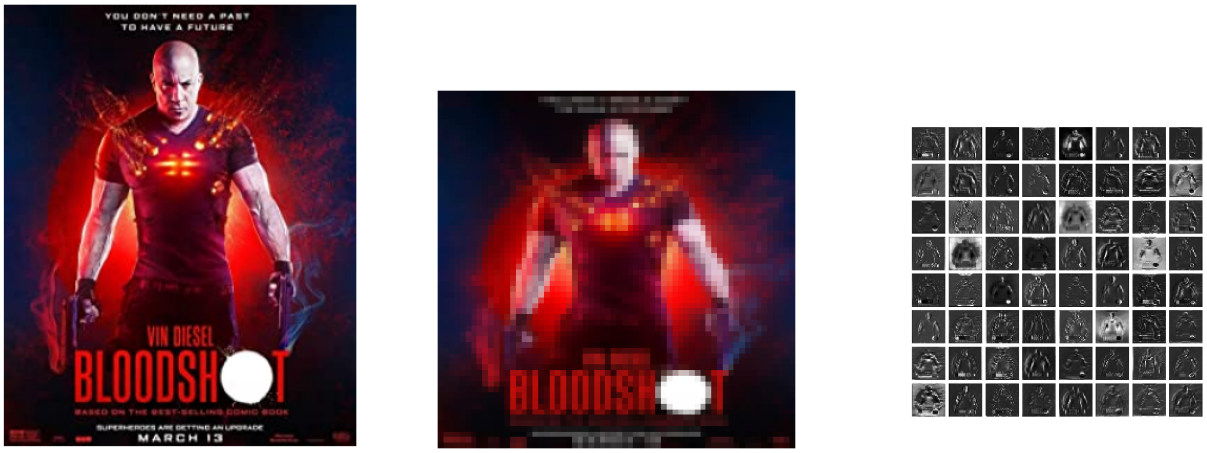
First we tried to analyze the GAN discriminator using state-of-the-art tools:

1. DeepLIFT (1);
2. DeepSHAP (5);



(a) Input image (b) Features heatmap

Figure 1: Feature heatmap using DeepLIFT



(a) Input image (b) Cropped image (c) Activation map 1

Figure 2: Feature maps

3. LIME (4).

We were not able to have significant results neither with DeepSHAP or LIME, as the outputs were not exploitable. However DeepLIFT allowed us to visualize the important features and patterns used by the discriminator, even if the results are still a bit unclear. On Fig 1., we can see for example that the most used features are located around the are of the head of the character that we can see on the input image.

3.2 Discriminator

After analyzing the feature maps generated by the discriminator (Figure 2), it becomes evident that the model has learned to focus on specific aspects of the input images. The feature maps highlight various visual characteristics such as shape, title, and color gradients,...

3.3 Generator

A perturbation analysis has been conducted on the generator. This involves modifying a single value of the input noise to different degrees, and observe the differences in the generated images in order to identify what this value represents. After doing this on every value of the 100-dimensional latent noise, only several values have been possible to understand. In figure 3, the background changes a lot, while the central character remains almost the same. In figure 4, the image takes a vivid red color when the value is increased.

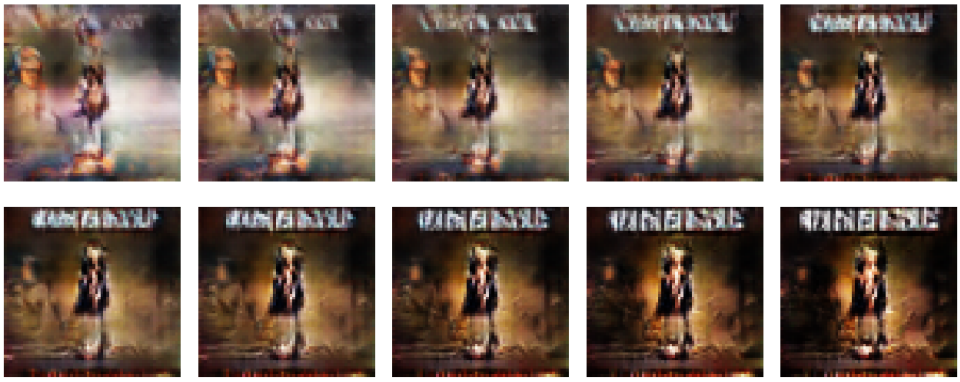


Figure 3: Images generated by the generator with slight modification in 1st value of input noise

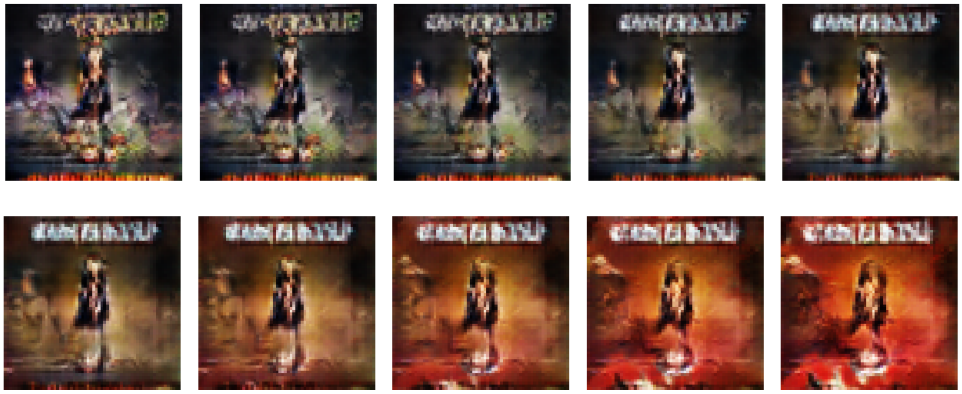


Figure 4: Images generated by the generator with slight modification in 6th value of input noise

The activation maps of the generator, in particular the last layers, have also been analyzed for the same image as in figure 3 and 4. It is interesting to notice that some of them have important activations in the regions of interest (figure 5), while some others add nothing to the output (figure 6). This could indicate that some filters are not useful and therefore could be pruned from the architecture to accelerate computations and reduce memory usage.

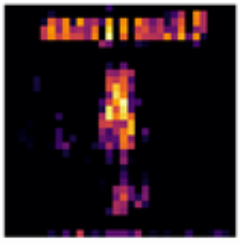


Figure 5: Activation map of filter 31 from last hidden layer

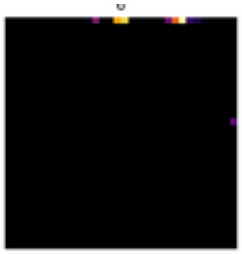


Figure 6: Activation map of filter 6 from last hidden layer

4 Discussion

Upon analyzing the results obtained from the movie poster generation process, two main challenges can be observed. Firstly, the simplicity of the employed GAN model limits its ability to learn the complex patterns and diverse elements found in movie posters. Unlike face images or objects that have relatively simple patterns to recognize, movie posters encompass various styles, layouts, color tones, and titles, which can vary based on the genres they represent. This complexity makes it a more complex task for the GAN model to grasp, resulting in generated images that lack meaningful content and suffer from low resolution.

Secondly, the low resolution of the input images further hampers the model's ability to capture detailed features. Due to the need for uniformity in input size, the images are cropped, leading to a deterioration in resolution. This reduction in resolution prevents the model from effectively learning and representing intricate details, resulting in a loss of visual quality and coherency in the generated images.

To address these challenges, two potential solutions can be considered. Firstly, employing Conditional GANs (CGANs) can enhance the model's performance by conditioning it on the movie genre. By providing the model with additional information about the genre, it may gain a better understanding of the diverse elements associated with each genre, ultimately improving the quality and relevance of the generated posters. Additionally, including higher-resolution movie posters into the dataset or exploring the use of Progressive GANs (PGANs) can help enhance the resolution and overall quality of the generated images.

A Code Submission

We verify that we have included all the resources relevant to our project, including demo files written in Jupyter Notebook snippets and experimental results. Link: https://drive.google.com/drive/folders/1lV9TrQQqQxF8GzDfHABPGsAJnx3kppAI?usp=drive_link

References

[1] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[2] 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.

[3] 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.

[4] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier.

[5] Shrikumar, A., Greenside, P., and Kundaje, A. (2019). Learning important features through propagating activation differences.