SEMINAR REPORT

ON

Interpreting Learned Features of Convolutional Neural Networks using Generative Adversarial Networks

Submitted by

Ajay T Shaju (SJC20AD004)

to

the APJ Abdul Kalam Technological University

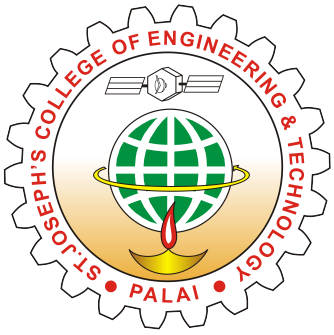
in partial fulfillment of the requirements for the award of the degree

of

Bachelor of Technology

in

Artificial Intelligence and Data Science



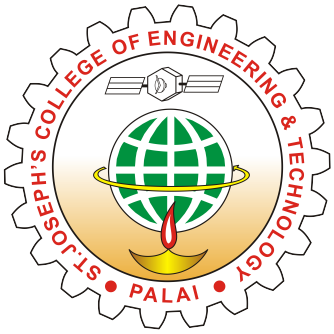
Department of Artificial Intelligence and Data Science

St. Joseph's College of Engineering and Technology, Palai

DECEMBER : 2023

ST. JOSEPH'S COLLEGE OF ENGINEERING AND TECHNOLOGY, PALAI

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE



CERTIFICATE

This is to certify that the seminar report entitled "**Exploring the CNN Learned Features using Generative Adversarial Networks**" submitted by **Ajay T Shaju (SJC20AD004)** to the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Artificial Intelligence and Data Science is a bonafide record of the seminar carried out by him under my guidance and supervision.

Seminar Guide
Dr.Deepa V
Head of the Department
Department of AD

Seminar Coordinator
Mr.Jacob Thomas
Assistant Professor
Department of AD

Head of Department
Dr. Deepa V
Associate Professor
Department of AD

Place: Choondacherry
Date: 01-12-2023

Acknowledgement

I wish to record our indebtedness and thankfulness to all who helped us complete this seminar titled "Exploring the CNN Learned Features using Generative Adversarial Networks". I would like to convey a special gratitude to Dr. V.P. Devassia, Principal, SJCET, Palai, for the facilities. I express my sincere thankfulness to Dr. Deepa. V, Head of the department, Department of Artificial Intelligence & Data Science for her cooperation and valuable suggestions. Also, I express my sincere thanks to the seminar coordinator Mr. Jacob Thomas for his helpful feedback and timely assistance. I am especially thankful to my guide, Dr. Deepa. V, Head of the department, Department of Artificial Intelligence & Data Science for giving me valuable suggestions and critical inputs through guidance and support. I also extend my thanks to college lab technicians, my friends, and others who directly or indirectly helped me during this seminar work.

Ajay T Shaju

Abstract

Generative Adversarial Networks (GANs) are a type of deep learning model that can generate new data, such as images, text, and audio, or enhance the quality of existing data. As a result, they have emerged as a transformative technology in the fields of machine learning and computer vision. GANs are a relatively new technology, but they already had a significant impact on the field of machine learning. GANs have been used to generate realistic images of human faces, translate languages, and create new forms of art and music, serving as a gateway to Generative AI. The introductory GAN has a simple architectural complexity, making it a vibrant research field to explore and create powerful systems. Many research efforts related to GANs are ongoing, and one noteworthy research paper that showcases the versatility of GANs is referred as a case-study for this seminar study. This research work enhances the interpretability of Convolutional Neural Networks (CNNs) in medical imaging, specifically for Pulmonary Edema and Liver Fibrosis. FIGAN utilizes a Conditional GAN (CGAN) to synthesize images that reveal CNN's learned features, making it a functional approach to interpreting medical images. This seminar explores the fascinating world of GANs, delving into their architecture, challenges, and applications, and a recent use case of GANs in-depth, highlighting their significance in generating synthetic data with remarkable realism.

Table of Contents

List of Abbreviations

AI Artificial Intelligence

BNPP B-type Natriuretic Peptide

CAM Class Activation Mapping

CGAN Conditional Generative Adversarial Networks

CIU Contextual Importance And Utility

CNN Convolutional Neural Network

CT Computed Tomography

DCGAN Deep Convolutional Generative Adversarial Network

DCNN Deep Convolutional Neural Network

DL Deep Learning

FID Fréchet Inception Distance

FIGAN Feature Interpretation Using Generative Adversarial Networks

GAN Generative Adversarial Network

GIF Graphics Interchange Format

Grad-CAM Gradient Weighted Class Activation Mapping

HD High Definition

MCMC Markov Chain Monte Carlo

ML Machine Learning

MNIST Modified National Institute of Standards and Technology

MRI Magnetic Resonance Imaging

NN Neural Network

PCA Principal Component Analysis

PGAN Progressive Generative Adversarial Networks

PLS Partial Least Squares

RBM Restricted Boltzmann Machine

SAGAN Self-Attention Generative Adversarial Networks

VAE Variational Autoencoder

VAE Variational Autoencoders

VG Visual Geometry

WGAN Wasserstein Generative Adversarial Networks

XAI Explainable Artificial Intelligence

List of Figures

List of Tables

Chapter 1

Introduction

Generative Adversarial Networks, commonly known as GANs, are a class of artificial neural networks used in unsupervised machine learning. GANs represent a significant breakthrough in the field of Machine Learning(ML) and Artificial Intelligence(AI). These models were first introduced by Ian J. Goodfellow and his colleagues in 2014 [1]. As per the creators, GAN is a framework for estimating generative models via an adversarial process [1]. GANs have gained prominence in recent years due to their unique ability to generate data that mimics real-world data distributions.

GANs are conceptually simple in their architecture, but their simplicity contradicts the complexity of the training process and the results they can achieve. The basic GAN architecture consists of just two components: a generator and a discriminator.

Figure 1.1: Random Images generated by a GAN created by NVIDIA. The GAN Architecture is known as StyleGAN [2], which generates artificial images starting from a low resolution(noise) to a high resolution. It controls the visual features that are expressed in each level, from coarse features(face shape) to finer details(hair color).

1.1 Background

Before going into the world of Generative Adversarial Networks. We need to understand the whole landscape of artificial intelligence and how GANs came into picture. Figure 2.2 gives the relationship between Artificial Intelligence, Machine Learning, Neural Networks, and Deep Learning.

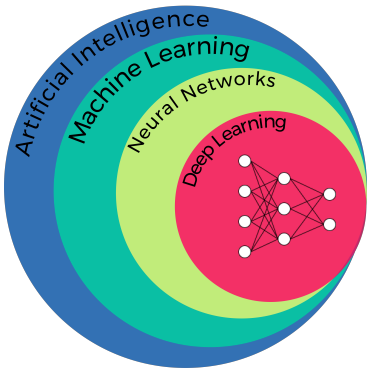


Figure 1.2: Relationship between AI, Machine Learning, Neural Networks and Deep Learning.

Artificial Intelligence is the field of computer science dedicated to creating intelligent agents, that can reason, learn, and act autonomously. Machine Learning is a subfield of AI that combines algorithms that can identify patterns in data without being explicitly programmed. Neural Networks (NNs) are a type of ML algorithm that are made up of interconnected neurons, which are able to learn from data by adjusting the weights between the neuron connections. Deep Learning (DL) is a subfield of ML that uses NNs with multiple layers, and are able to learn from complex data, such as images and languages without defining features explicitly.

Figure 2.3 gives a visual idea of how the learning models are arranged. The machine learning models can be broadly classified into supervised learning, unsupervised learning and reinforcement learning.

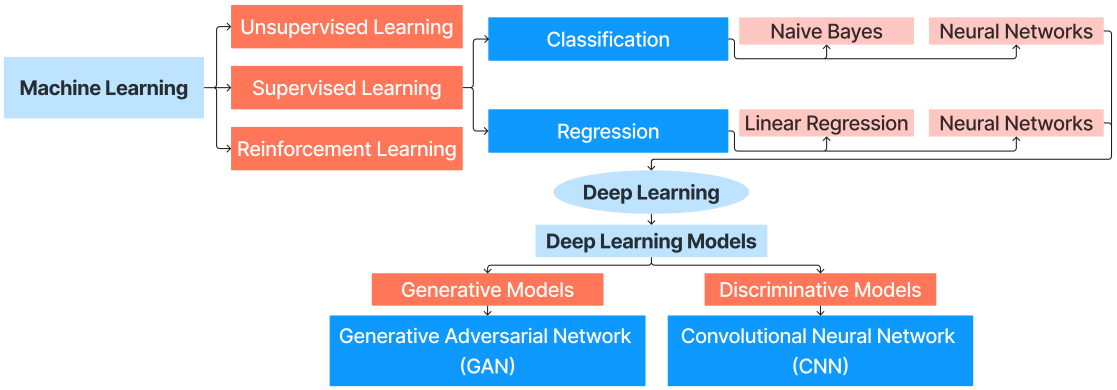


Figure 1.3: A visual guide to the taxonomy of learning models.

Supervised learning is a type of ML in which the data provided for learning has labels(ground truths) in it. It is divided into regression and classification tasks, in which neural networks are commonly used in both tasks, thereby researchers have derived a new branch(or subset) of machine learning which is called Deep Learning.
The Deep Learning models can be broadly categorized into two main types: Generative Models and Discriminative Models.

- **Generative Models:**
  - Generative models are designed to model the underlying probability distribution of a dataset [3]. They generate new data points that are similar to the training data they were exposed to. Generative models are often used for tasks such as image generation, text generation, and data synthesis.
    One of the best examples of a generative deep learning model is GANs - An unsupervised generative machine learning model that can create new and realistic data based on its training data.
- **Discriminative Models:**
  - Discriminative models, on the other hand, aim to learn the boundary or decision surface that separates different classes or categories in the data. They are typically used for classification and regression tasks.
    One of the best examples of a discriminative deep learning model is Convolutional Neural Networks - A model used for classification and other image and audio related tasks in deep learning.

1.2 Motivation

The motivation behind studying GANs lies in their transformative potential in various domains and GANs are considered to be one of the earliest and most effective applications of Generative AI when it comes to real discussions. And GANs also have the power to:

- Generate realistic images, text, and even videos:
  - The work on generating realistic data represents a profound advancement in computer vision and natural language processing.
  - This could even help in industries such as photography, and movies by generating high-quality images and videos that could be used as fillers in their works.
  - Realistic images and videos could be used to make artificial environments, where we could train reinforcement models and create video games.
- Aid in data augmentation and synthesis, benefiting healthcare [4], finance [5], and many other industries.
  - GANs can generate synthetic medical images, such as X-rays, Computed Tomography(CT) scans, or Magnetic Resonance Imaging(MRI) images, which can be used for training and testing machine learning models. This is especially helpful when there is a scarcity of labeled data.
  - GANs can generate normal patient data, allowing healthcare organizations to train models for anomaly detection, such as identifying rare diseases or unusual medical conditions.
  - GANs can generate molecular structures, aiding in the discovery of new drugs by simulating various chemical compounds. This can accelerate the drug development process and reduce costs.

GANs stand as a pivotal innovation with profound implications across diverse sectors like visual media, data synthesis, and problem-solving. The transformative potential of GANs in augmenting creativity and addressing complex problems underscores GAN being the focal point of this seminar.

1.3 Outline

The outlined structure delves deeply into Generative Adversarial Networks and their diverse applications. It starts with an introduction that sets the stage for understanding GANs and proceeds with a comprehensive exploration of their historical journey, tracing their evolution with insights gained from extensive research writings on the subject. This study helps in understanding how GANs have developed and transformed over time.

Moving forward, the report takes a closer look at the fundamental components of GANs, breaking down complex concepts into more digestible bits. It details their architecture, explaining how they're built, the methods used to train them, and the various types that exist. By unraveling these aspects, the aim is to simplify the understanding of how GANs function and their different variations.

Furthermore, the report dives into the methodologies used by researchers to study GANs, shedding light on the processes involved in gathering important data. It explores the theories behind GANs, aiming to identify areas where further research could be fruitful. This section also focuses on making AI systems easier to comprehend, highlighting the limitations of current methods and proposing a new approach called FIGAN that aims to enhance the clarity of AI systems.

In addition to theoretical discussions, the report describes frameworks employed in addressing a critical aspect, AI explainability. Which scrutinizes widespread limitations or doubt about how AI systems make decisions, and introduces innovative methodologies, the FIGAN architecture, to enhance explainability within AI systems.

Lastly, the conclusion sums up the accumulated insights, acknowledging the existing limitations in understanding GANs while pointing out potential avenues for future research and innovation. It acts as a guide for further exploration and improvement in the field of Generative Adversarial Networks.

Chapter 2

Literature Review

The literature review conducted on 25 research papers on GANs has provided a comprehensive understanding of the evolution, applications, and challenges inherent in this field of study. Spanning from the influential work by Ian J. Goodfellow in 2014 to the development of diverse GAN variants such as Deep Convolutional GAN(DCGAN) and Variational Autoencoders GAN(VAEGAN), the review contains the thorough study of applications of GAN variants across disciplines like art, medicine, and data generation. Therefore the literature survey acts as a roadmap for comprehending the intricacies, practical implementations, and prospective research trajectories within the GAN landscape.

2.1 Evolution of GANs

2.1.1 Early Generative Models: Foundations Preceding GANs

There were many models, which tried to explore the Generative AI landscape. But three of the most important works (Probabilistic Models) were Variational Autoencoders (VAEs) [6], Restricted Boltzmann Machines (RBMs) [7], and Markov Chain Monte Carlo (MCMC) [8] Methods.

VAEs were used in generating realistic images by learning a latent space representation of images. By sampling points from this learned latent space, VAEs could produce new, yet similar, images. For instance, in generating images of human faces, digits, or scenes, VAEs provided a method to create novel images with controlled variations.

RBMs were employed in collaborative filtering tasks within recommendation systems. By learning the latent features of users and items, RBMs could make recommendations based on a user's preferences and historical data, suggesting movies, products, or content similar to those a user had interacted with previously.

MCMC methods were extensively used for Bayesian parameter estimation in complex models. They provided a means to sample from the posterior distribution over model parameters, enabling probabilistic reasoning and uncertainty estimation in various domains such as finance, epidemiology, and ecological modeling.

2.1.2 Emergence of GANs: Transformative Innovations

Ian J. Goodfellow et al. [1] proposed *"Generative Adversarial Nets"* in June 2014. Goodfellow introduced the concept of GANs, which was based on the field of Generative AI. Thereby after two years, Goodfellow alone presented a tutorial at Neural Information Processing Systems on GANs, where the tutorial delved more into the mathematical and practical sides of GANs which is explained in the paper *"NIPS 2016 Tutorial: Generative Adversarial Networks"* [9] in December 2016.

Mehdi et al. has proposed a new architecture for GANs *"Conditional Generative Adversarial Nets (CGAN)"* [10] in November 2014. Where they added an extra term to the general equation of GANs to gain more control over the output of GAN.

After the very introduction of GAN Architecture, Radford et al. proposed *"Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks"* [11] in November 2015, Where they bridged the gap of using CNN for unsupervised machine learning tasks by introducing a new GAN architecture named Deep Convolutional Generative Adversarial Networks.

During the years 2016-2020, many GAN architectures were introduced (about 500+), and most of them solved some or many problems with the traditional GAN, and others created new systems that were powered by GANs, for example, Head Models [12].

2.1.3 Post-GAN Era: Evolving Generative Models

Zhang et al. wrote *"Self-Attention Generative Adversarial Networks"* [13] in June 2019. In this paper, the authors integrated self-attention mechanisms within the framework of GAN, aiming to enhance the model's ability to capture long-range dependencies in images. By enabling GANs to focus on relevant image regions at different scales simultaneously, which facilitated more globally consistent image synthesis.

Jiang et al. *"TransGAN: Two Pure Transformers Can Make One Strong GAN, and That Can Scale Up"* [14] in December 2021. In this paper, the authors tried to get rid of CNN from GAN Architecture and replace it with Transformers, a neural network architecture that powers most Large Language Models nowadays.

Zhaoqing et al. wrote a survey paper *"Recent Progress on Generative Adversarial Networks (GANs): A Survey"* [15] in March 2019, where they described various GAN architectures with their performance analyses.

Der-Lor et al. have proposed *TwinGAN: Twin Generative Adversarial Network for Chinese Landscape Painting Style Transfer* [16] which is trained to imitate multiple styles of Chinese landscape paintings. The TwinGAN network has successfully imitated five styles of Chinese landscape paintings.

Anders et al. have proposed *Autoencoding beyond pixels using a learned similarity metric* [17] in June 2023. The paper pioneers a novel amalgamation of Variational Autoencoders and GAN, amplifying image generation and reconstruction by emphasizing feature-wise errors using the GAN discriminator. It surpasses prior approaches that rely on handcrafted similarity metrics or conditional networks by introducing a learned similarity measure and a unified training framework.

2.2 Generative Technologies in Medical Imaging

S Band et al. wrote a systematic review paper of interpretability methods titled *Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods* [18].

In that paper [18] the main DL methods discussed were:

- Gradient-Weighted Class Activation Mapping (Grad-CAM): Class Activation Mapping (CAM) utilizes gradient data from the last convolutional layer in a CNN to generate a basic map of significant areas in an image, correlating with the classification. Grad-CAM extends CAM's scope, applicable to various CNN architectures without the need for retraining. [19]

- NeuroXAI: A CNN is employed to analyze brain images, generating convolutional feature maps and specific output computations. These calculations include visualizations depicting clusters associated with tumor classification and segmentation.

- Contextual Importance And Utility (CIU): It uses the concept of attribute importance and its reliance on other attributes' value. It considers the correlation of the importance of an attribute. When a combination of features is appropriate or causes interaction on prediction, it is interpreted as feature interaction, which leads to higher-level explanations.

In the main reference of this study, GAN was tested for use as Explainable AI(XAI) method in medical imaging tasks by exploring the learned features(feature maps) of CNN by Hasenstab et al.in their paper *"Feature Interpretation Using Generative Adversarial Networks (FIGAN): A Framework for Visualizing a CNN's Learned Features"* [20] in January 2023.

2.3 Summary

The evolution of Generative Adversarial Networks commenced with foundational models VAEs, RBMs, and MCMC methods, which laid the groundwork for generative AI. The introduction of GANs in 2014, lead to diverse GAN architectures like CGAN and DCGAN, addressing many limitations.

In medical imaging, GANs play a crucial role in interpretability methods. Techniques like Grad-CAM and FIGAN are employed to understand the learned features of CNNs. This exploration aids in contributing to XAI methodologies.

Chapter 3

Theoretical Aspects

The chapter serves as a comprehensive guide to the processes of conducting a study or development of new Generative Adversarial Networks. This is divided into two sections:

- Research Methods and Data Collection, and
- Theoretical Explanation and Math behind GANs

3.1 Research Methods and Data Collection

Researchers employ a variety of methods in GAN-related studies, such as:

- Collecting and Preprocessing Data
  - Data Sources: Researchers start by collecting the data required for their specific GAN application. This data can be in various formats, including images, text, audio, or any other type of data relevant to the problem at hand.
  - Data Preprocessing: Before feeding data into a GAN, it often requires preprocessing. This may involve tasks like data cleaning, normalization, resizing, and data augmentation to ensure that the data is suitable for training.
    * Image Generation: In a study focused on generating realistic human faces, researchers collect a dataset of thousands of portrait images. These images are preprocessed to standardize the size, and crop faces, and adjust brightness and contrast to ensure uniform quality.

    - Text Generation: For generating text using GANs, a text corpus of novels is collected. Preprocessing involves tokenization, removing punctuation, and lowercasing the text to create a clean and consistent dataset.

- Training GAN models using popular frameworks
  - Researchers typically choose popular deep learning frameworks like TensorFlow, PyTorch, or Keras for developing and training GAN models. These frameworks offer a wide range of tools and resources for building and training neural networks, which are essential components of GANs.
  - Architecture Selection: Depending on the specific GAN variant or the application, researchers choose the appropriate architecture for the generator and discriminator networks. For instance, they may opt for a Convolutional GAN for image generation tasks.

- Fine-tuning models and optimizing hyperparameters.
  - After initial training, researchers may perform fine-tuning, which is a crucial step to enhance the quality and diversity of generated data. This can involve adjusting parameters like learning rates, and batch sizes, adding regularization techniques and network architecture specifics to achieve the best performance. It is essential to ensure that the GAN meets the specific requirements of the task, whether it's generating high-resolution images, text, or any other content.
    - Image Super-Resolution: After training a GAN to enhance image resolution, fine-tuning might involve adjusting the network architecture by adding skip connections or residual blocks to improve the visual quality of the generated high-resolution images.
    - Conditional Text Generation: In text generation tasks, if researchers want to condition the GAN on specific attributes (e.g., genre for generating stories), they fine-tune the GAN by incorporating these conditions into the generator and discriminator networks.

3.2 Theoretical Explanation for GANs

This section embarks on a journey to delve deep into the core theoretical foundations and mathematical intricacies that underpin the remarkable success of Generative Adversarial Networks. Within this exploration, two main components of GANs, architecture, loss functions, training procedures, types of GANs, and practical applications are explained.

Figure 3.1: Smooth transition of an image using GAN

3.2.1 Components of GAN

At the core, GANs contain two main components:

- The Generator, and
- The Discriminator

The Generator: It is the creative element of the GAN. It takes random noise as input and transforms it into data samples. These samples can be anything from images to text. *The primary objective of the generator is to produce data that is indistinguishable from real data.* In other words, the generator is responsible for creating new data(a.k.a fake data). It is implemented as a neural network, often a Deep Convolutional Neural Network(DCNN), which is well-suited for image generation tasks [21].

The Discriminator: The discriminator plays the role of a binary classifier. *It evaluates whether a given data sample is real (from the training set) or fake (generated by the generator).* In other words, the discriminator is responsible for distinguishing between real and fake data. Like the generator, the discriminator is also implemented as a neural network. Its goal is to become increasingly accurate at distinguishing real data from fake data over time.

GANs were inspired by the game theory, more precisely, a zero-sum game, in which one's gain(+1) is the other's loss(-1), and the net outcome will be zero(+1-1 = 0). The generator and discriminator will compete with each other to achieve the Nash equilibrium(i.e. achieving the desired outcome not by deviating from the initial strategy) during its training process. [1]

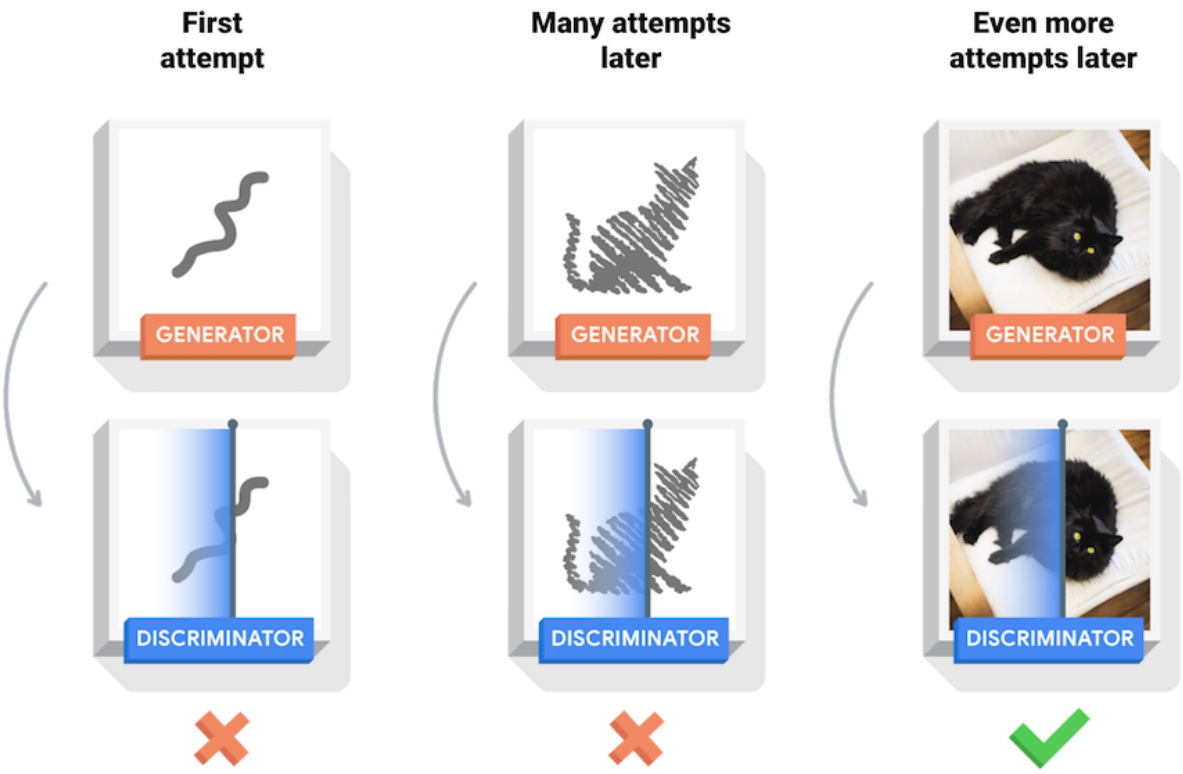


Figure 3.2: Working of generator and discriminator

Figure 3.2 depicts the working of the generator and discriminator in different attempts. At the beginning, the generator had no idea of what to create, so it made a doodle-like output. After many attempts, the generator has created a visually appealing and realistic image that is indistinguishable from the real image.

3.2.2 Generator

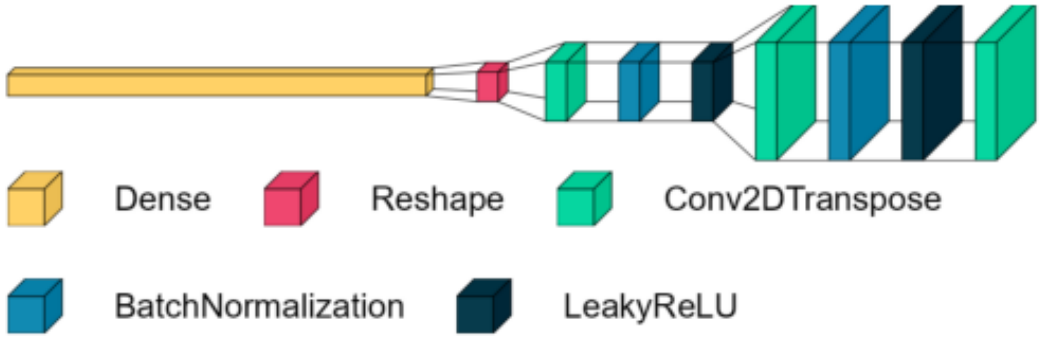


Figure 3.3: Generator Architecture

With the introduction of Deep Convolutional GAN by Radford, Chintala, et al. Convolutional Neural Networks have become the primary architecture for image generation tasks. Figure 3.3 shows a sample image of such image generation networks.

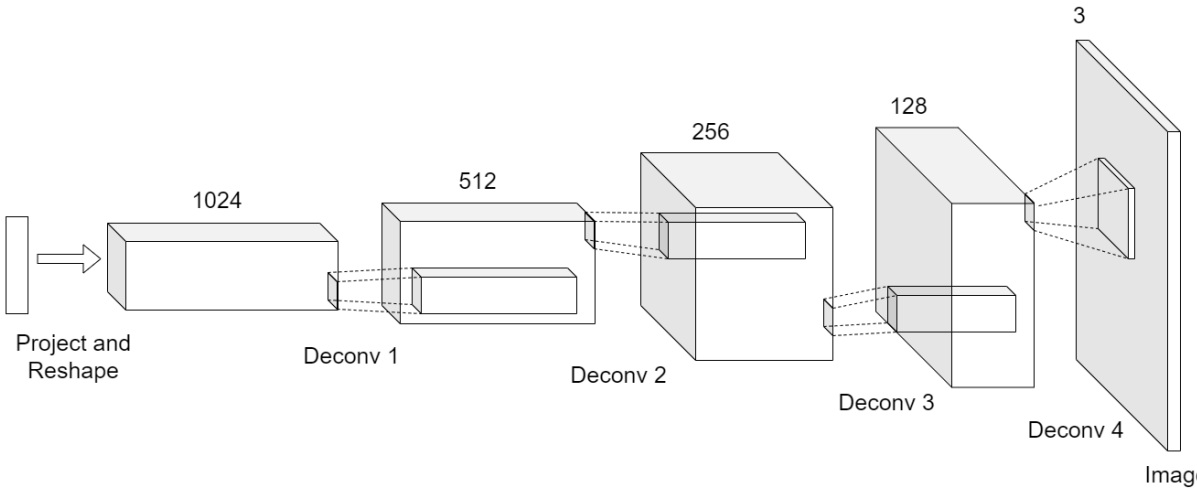


Figure 3.4: Deconvolution Network used for upsampling in Generator

Figure 3.4 depicts an Upsampling/Deconvolution Network, where it starts from a lower dimension, and becomes a clear and Full HD image at the end.

3.2.3 Discriminator

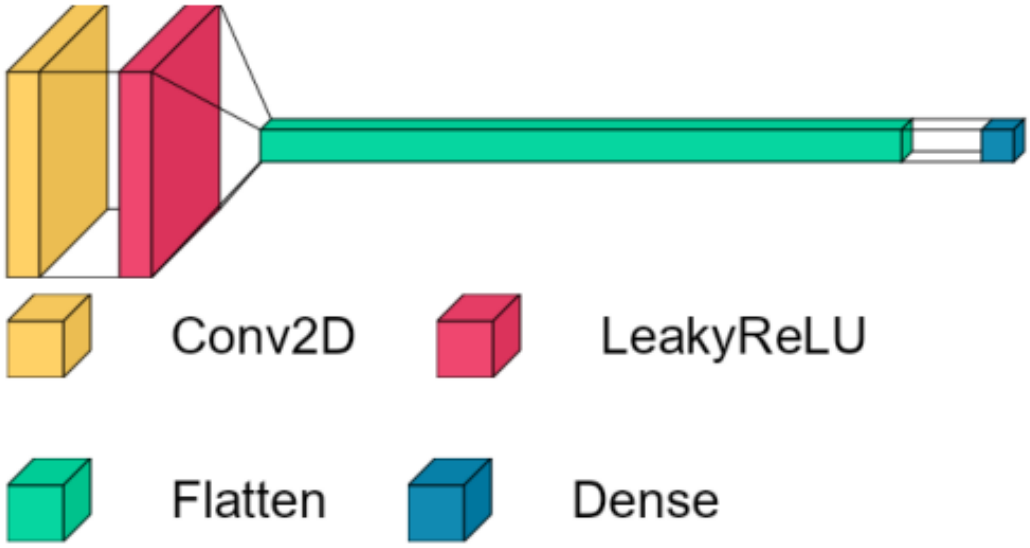


Figure 3.5: Discriminator Architecture

The architecture of a discriminator is very similar to a normal binary classifier. It could use any network architecture appropriate to the type of data it's classifying.

The discriminator will analyze the image using CNN layers and predict whether the image is real or fake (indicated by the integer value 1 or 0). The real example comes from the training dataset, and the generated examples are output by the generator model.

In some advanced GAN variants, the discriminator may be designed to adapt to different levels of image quality, and it may provide more refined feedback to guide the generator's training.

3.2.4 Architecture

The general architecture of GAN is a single diagram representing the training and testing process. Explanation for the architecture diagram (Figure 3.6):

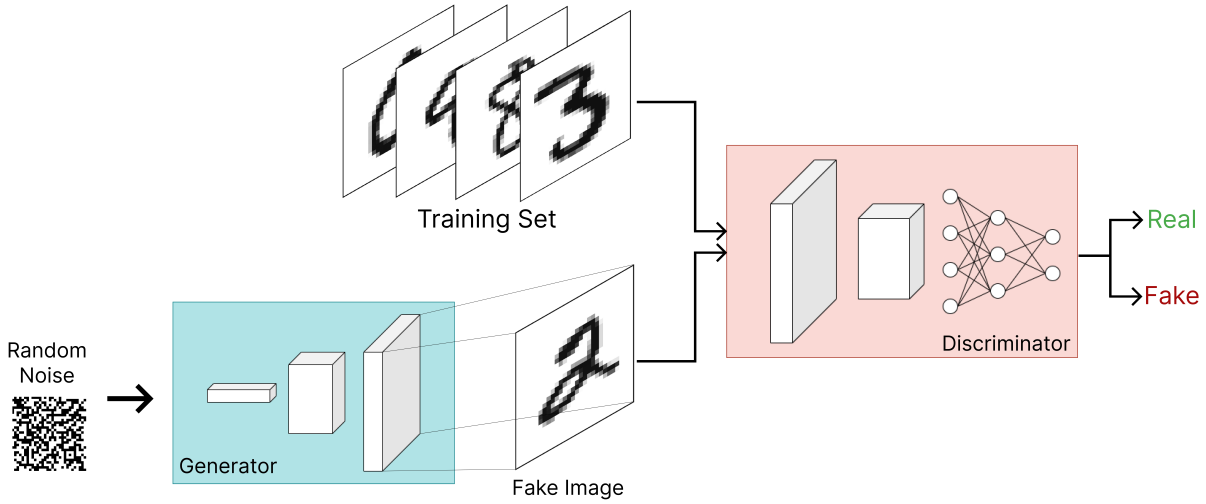


Figure 3.6: General Architecture of GAN

First, the initialization is done with some random noise, it can be considered as a matrix of values that is taken from the training/testing data distribution. This random noise is given as the input of the generator, where the generator uses this random noise to create fake data.

In the next step training sets and fake images are inputted to the discriminator during the training step. However, during the testing step, we only input the fake data created by the generator. During both phases, the discriminator has only one role, which is to identify whether the data given to it is real or fake. This output is given as an integer value of 1 or 0.

The block diagram of the same architecture is given in Figure 3.7, it also depicts the same idea explained above. This diagram is the first point of reference for diagrams that will be shown in the GAN training section.

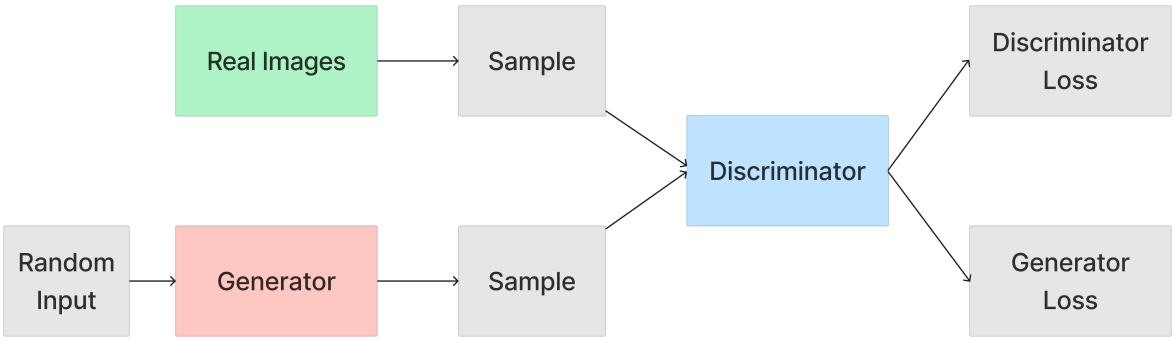


Figure 3.7: Block Diagram of GAN

The Generator will get its loss after the discriminator checks whether the created data is real or fake. The discriminator's loss depends on its output for the corresponding input of the generator.

3.2.5 Loss Function

The Standard GAN Loss Function is also known as the Min-Max Loss [9]. This loss function is divided into Discriminator loss and Generator loss. The equation is given below, equation (1).

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim P_z(z)}[\log(1 - D(G(z)))] \quad - \quad (1)$$

Explanation for the terms in equation (1):

- **E** represents the expected value or the average.
- **x** represents real data samples drawn from the real data distribution
- **D(x)** is the output of Discriminator when it distinguishes real data from fake data.
- **z** represents random noise.
- **G(z)** is the output of the Generator when given random noise as input.
- **1-D(G(z))** represents the probability that the Discriminator assigns to the generated data being fake(1 minus the probability it assigns to it being real).

The general GAN loss function is the sum of derived forms of discriminator and generator loss functions, which could be derived using the general equation of Binary Cross-Entropy, which is given by

$$L(y, \hat{y}) = -\left(y \cdot \log \hat{y} + (1-y) \cdot \log(1-\hat{y})\right) \; - \; (2)$$

Where $\mathbf{L(y, \hat{y})}$ is the loss.
We substitute y = 1 for real images and y = 0 for fake images in the Binary Cross-Entropy equation, to get:

$$L(\text{Discriminator}) = \min\left[-\left(\log(D(x)) + \log(1 - D(G(z)))\right)\right] \; - \; (3)$$

Removing the negative sign will change min to max:

$$L(\text{Discriminator}) = \max\left[\log(D(x)) + \log(1 - D(G(z)))\right] \; - \; (4)$$

Maximizing the loss function means to put D(x) = 1, and D(G(z)) = 0. So, the loss function will result in 0 (Fake image).
Note: $\mathbf{\log(1) = 0}$
Generator loss is calculated from the discriminator loss. The generator will try to fool the discriminator into classifying the fake data as real data. This implies that the generator tries to minimize the second term in the discriminator loss equation.

$$L(\text{Generator}) = \min\left[\log(1 - D(G(z)))\right] \; - \; (5)$$

Minimizing the loss function means to put D(G(z)) = close to 1. So, the loss function will result in 1 (Real image).

By taking the sum of both the equations (4) & (5), we get the general equation for the GAN loss function back, equation (6):

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim P_{\text{data}}(x)}\left[\log D(x)\right] + \mathbb{E}_{z \sim P_z(z)}\left[\log(1 - D(G(z)))\right] \; - \; (1)$$

3.2.6 Training

The training of GANs has three main phases:

- The starting of training
- The Optimal training time, and
- The stopping of training

The GAN starts training in alternating periods. The discriminator trains for one or more epochs then the generator's turn and this step repeats. Careful training is required, as most challenges of GANs are in the training phase.

The optimal training duration will depend on:

- Complexity of the data being generated
- Size of the training dataset, and
- The desired quality of the generated data.

The training can be stopped when:

- The generator is able to produce high-quality samples that are identical to real data.
- The discriminator is no longer able to classify real and fake data with high accuracy.

3.2.7 Generator Training

In a typical Generative Adversarial Network training process, random noise is first fed into the generator, which produces a synthetic data sample. This generated sample is then passed to the discriminator, whose role is to classify whether the sample is real or fake. The discriminator is kept constant during the generator training phase. Otherwise, the generator would try to hit a target (output of discriminator) and never converge. The discriminator's classification forms the basis for calculating the loss, quantifying the difference between the discriminator's predictions and the actual labels (real or fake).

This loss is backpropagated through both the discriminator and the generator to obtain gradients, which describe the direction and magnitude of the necessary weight adjustments. However, it's crucial to note that during this training phase, these gradients are used exclusively to modify the generator's weights. This adversarial interplay between the generator and the discriminator continues iteratively, with the generator aiming to produce increasingly realistic data while the discriminator seeks to become more selective, ultimately leading to improved GAN performance.

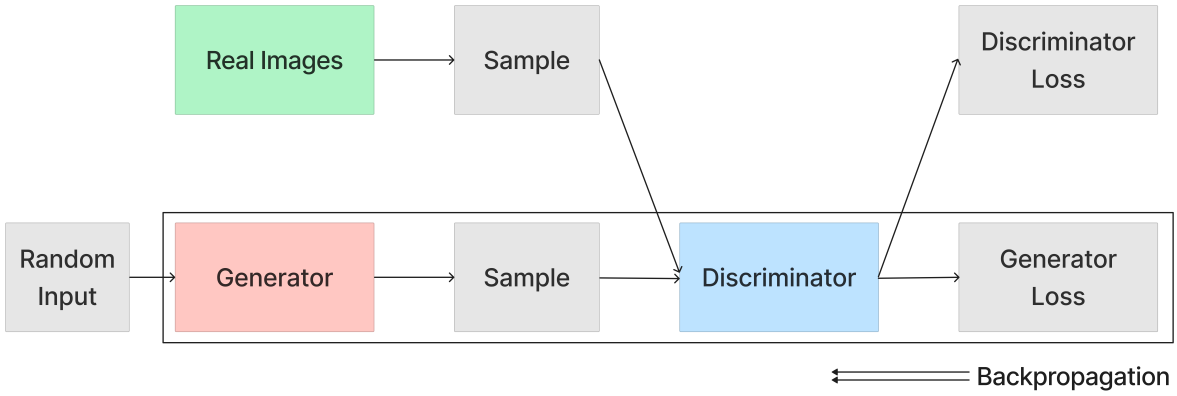


Figure 3.8: Generator Training

3.2.8 Discriminator Training

In the discriminator training process, a fake image generated by the generator is passed to the discriminator. The discriminator's role is to classify whether the image is real (from the true data distribution) or fake (generated by the generator). During discriminator training, the generator is paused. Its weights remain constant while it produces examples for the discriminator to train on. Based on this classification, a loss is calculated, quantifying how well the discriminator is distinguishing between real and fake images. The gradients of this loss are then obtained through backpropagation and used to adjust the discriminator's weights. This iterative process helps the discriminator become better at distinguishing real from fake images.

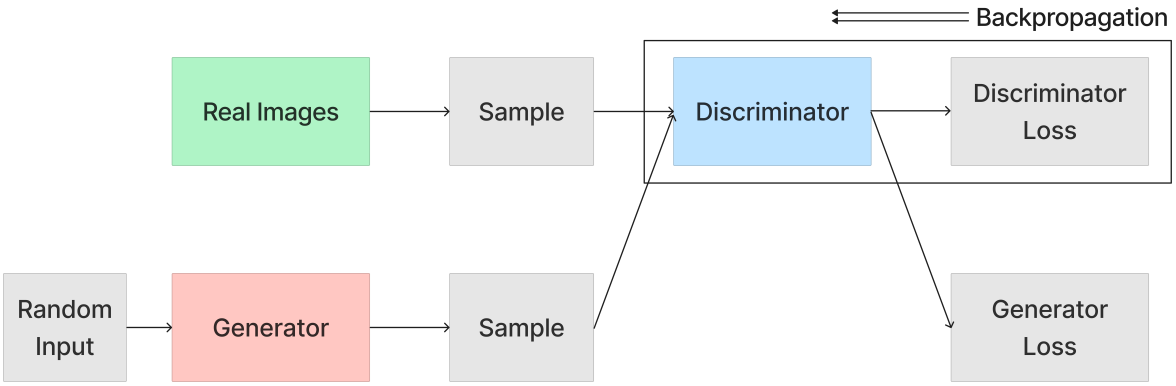


Figure 3.9: Discriminator Training

3.2.9 Types of GANs

Table 3.1: Types of GANs with Specializations

GAN Type	Specialization
Normal GAN	Basic GAN architecture for image generation
CGAN	Controlled image generation(conditional inputs)
DCGAN	Deep Convolutional GAN for image generation
StyleGAN	High-quality image synthesis with style control
BigGAN [22]	High-quality images at high resolutions
MuseGAN [23]	Music Generation using GAN
DiscoGAN [24]	Learn cross-domain relations in unsupervised data.
CycleGAN [25]	Unpaired image-to-image translation
SocialGAN [26]	Socially Acceptable Trajectories with GAN
ProgressiveGAN(PGAN) [27]	Training GANs for high-resolution images
Self-AttentionGAN(SAGAN)	Self-attention mechanism for image generation
WassersteinGAN(WGAN) [28]	Enhanced stability and training by Wasserstein loss

Even though there are different types of GANs, the focus is only on the Conditional Generative Adversarial Networks because it is used in the research work, which is going to be explored in the next chapter.

The CGAN is used because, in traditional GAN, the user has no control over the modes of the data to be generated. But conditional GAN changes that by adding the label 'y' as an additional parameter to the generator to generate specified images. Also, the label is added to the discriminator to distinguish real images better. See the mathematical equation differences:
The general loss function equation, same equation as given above (equation(1)) for GAN is given by,

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{\text{data}}(x)}\left[\log D(x)\right] + \mathbb{E}_{z\sim P_z(z)}\left[\log(1-D(G(z)))\right] \; - \; (1)$$

The equation for CGAN is given by,

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{\text{data}}(x)}\left[\log D(x|y)\right] + \mathbb{E}_{z\sim P_z(z)}\left[\log(1-D(G(z|y)))\right] \; - \; (6)$$

In equation (6) **y** term is a new addition, which is used to control the output of CGAN.

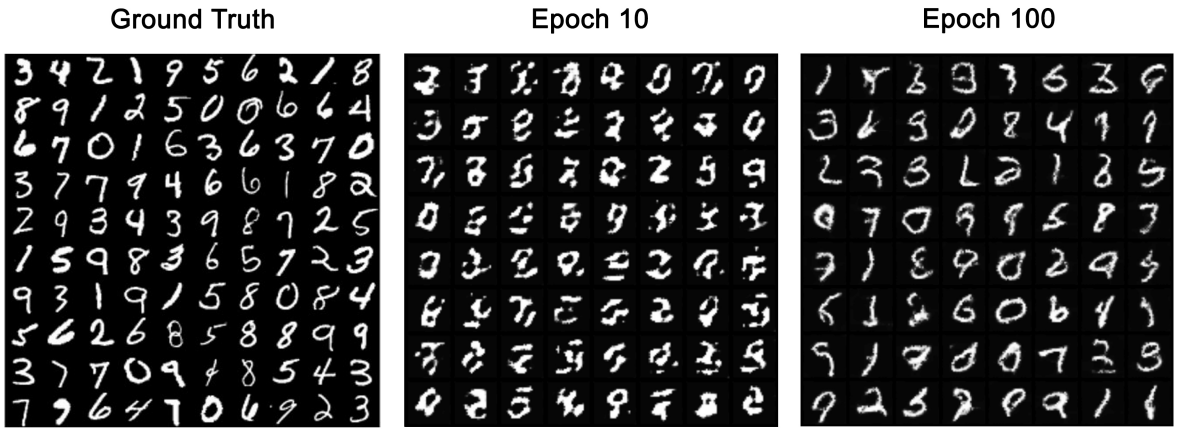


Figure 3.10: GAN working on MNIST Data [1]

3.2.10 Uses of GANs

Table 3.2: Uses of GANs

Application	GAN Type	Description
Image Generation	Vanilla GAN	Realistic images from noise
Style Transfer	CycleGAN	Apply artistic styles to photos
Face Aging	Face Aging GAN [29]	Simulate the aging process of faces
Data Augmentation	DCGAN	Training data for ML models
Video Generation	VideoGAN [30]	Synthetic video sequences
Drug Discovery	ChemGAN [31]	Molecular structures for drug design

In addition to Table 3.2, Figure 3.11 will give a pie-chart representation of the use of GAN in the industry.

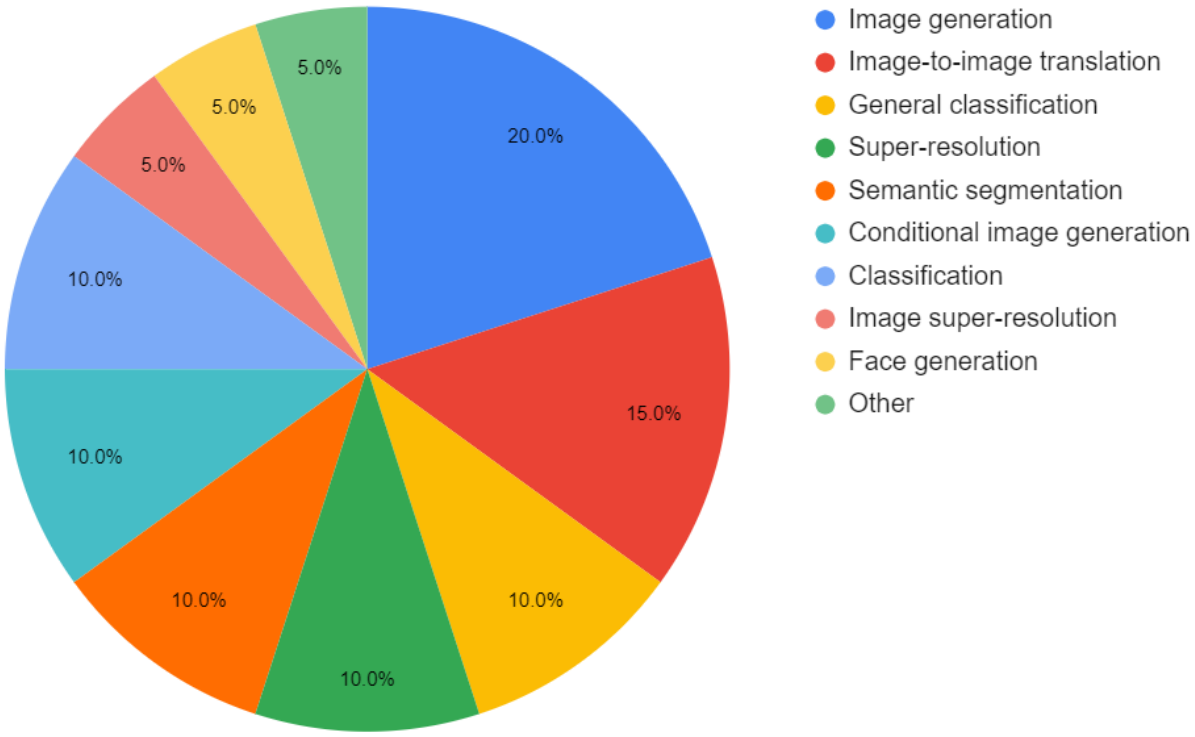


Figure 3.11: Uses of GAN as Pie-Chart

In a nutshell, Generative Adversarial Networks are a powerful type of deep learning model that can be used to generate new data, such as images, text, and music [11] [5] [23]. The GAN is trained using two neural networks: a generator and a discriminator. The generator creates new data, and the discriminator tries to distinguish between real and generated data. [9] The generator and discriminator are trained in a competitive manner, where the generator tries to fool the discriminator, and the discriminator tries to become better at distinguishing between real and generated data. [1]
Figure 3.11 illustrates a classical example usage of GANs to regenerate Modified National Institute of Standards and Technology(MNIST) Data. MNIST is a dataset of handwritten digits, commonly known as *Hello World to Deep Learning*. The figure shows the GAN at two different epochs: 10 and 100. At epoch 10, the generator is still learning how to generate realistic MNIST digits. The generated digits are often blurry and deformed. However, by epoch 100, the generator has learned to generate much more realistic MNIST digits. The generated digits are now sharp and well-defined.

There are many specialized GANs that have been developed for specific tasks, such as DCGAN(Image Generation) designed for generating high-quality synthetic images, StyleGAN(Human Face Generation) generating high-resolution human faces with a high degree of variability and realism, StarGAN(Style Transfer) [32], which is capable of performing style transfer across multiple domains. For instance, it can convert the style of an image from one domain (e.g., one type of facial expression) to another (e.g., a different facial expression)., and SemanticGAN (Semantic Segmentation) [33], it is oriented toward semantic segmentation, a task in computer vision that involves classifying each pixel in an image into a specific category, such as object detection or image segmentation. These models are also used to develop new types of applications like Artbreeder, DeepFace, Prisma, FacePlay, MidJourney, and DALL·E.

Chapter 4

Research Opportunities And Challenges

As far as we came with the theories and mathematics behind GANs, now the study is directed towards novel research work in the field of GANs which is *Feature Interpretation Using Generative Adversarial Networks: A Framework for Visualizing a CNN's Learned Features*, which is an innovative strategy to address post hoc explainability issues while using CNNs for medical imaging, using GANs. The contents of the research paper are Introduction, Commonly and Currently Used Explainability Methods, Proposed Methodology, Applications, Results and Discussion, Experiments, and Conclusion.

4.1 Introduction

In medical imaging, CNNs have emerged as powerful tools for a diverse array of classification tasks. However, their growing utility has exposed a critical need for improved methods of explainability. Current techniques of explainability often focus on highlighting specific regions within an image that contribute to a classification decision. Yet, this approach is insufficient in cases involving co-localized or diffuse conditions, such as pulmonary edema or fibrosis, where traditional localization alone may lead to ambiguity. FIGAN leverages CGAN to synthesize the CNN's core embedded features. The author's efforts have demonstrated how the resulting feature interpretations can effectively disambiguate attention areas highlighted by existing explainability methods.

4.2 Case Study

The realm of Artificial Intelligence (AI) is undergoing a profound transformation as the quest for explainability becomes an increasingly pivotal challenge, profoundly influencing the adoption and trust in AI systems. This case study embarks on a journey deeply rooted in the importance of explainability within AI systems. The study highlights the challenges that hinder these conventional techniques from effectively unveiling the black-box nature of AI decision-making processes. Thereby rendering AI systems understandable, especially given their growing impact across diverse sectors like healthcare, and finance, where the accuracy and transparency of AI-driven decisions hold significant weight.

Unveiling Innovative Approaches:

As an integral part of this inquiry, the study shifts focus towards innovative AI explainability methods that transcend the boundaries of traditional approaches. It delves into groundbreaking techniques and frameworks, seeking to address the identified limitations and complexities. Among these innovative methodologies, the study pays particular attention to the FIGAN architecture, an emerging paradigm that shows promise in unraveling the opacity of Medical AI systems.

Proposed Methodology and Goal:

The primary objective of this case study lies in exploring the potential of innovative approaches, such as the FIGAN architecture, to tackle the challenges of AI explainability. By synthesizing insights gathered from analyzing limitations in conventional methods and exploring cutting-edge approaches, the study aims to contribute solutions and insights that advance the frontier of AI explainability. Ultimately, the case study endeavors to address the imperative need for transparent and understandable AI systems, shaping the future landscape of AI research and application across various domains.

4.3 Explainability in AI Systems

Explainable AI addresses the opaque nature of AI models, aiming to clarify their decision-making processes. The XAI has prime importance as the world becomes more technologically advanced, especially in the medical field for making interpretable AI. Therefore, This section delves into both the limitations of common explainability methods and other useful explainability methods.

Amitojdeep Singh et al. wrote a paper in June 2020 titled *"Explainable Deep Learning Models in MedicalImage Analysis"*. In which research is done from a practical side of approaches, challenges in deployment, and the areas that need further research. The majority of other explainability studies were focused on features that influence the decision of a model, taxonomy, ethics, and the need for explanations. [34]

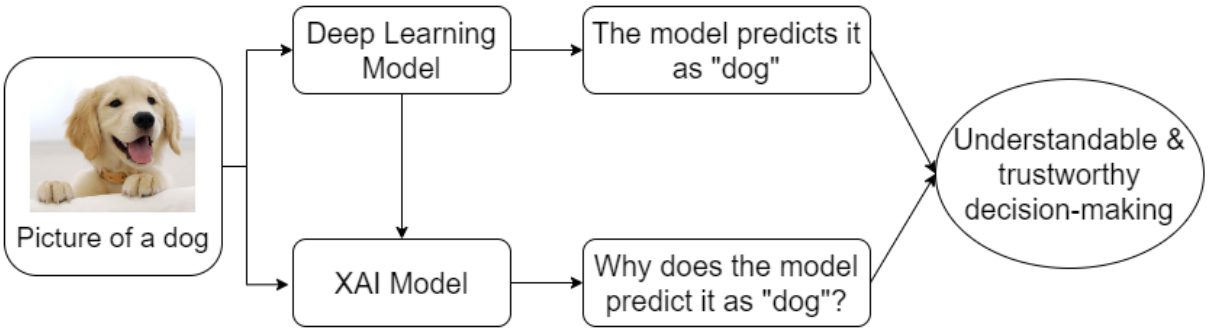


Figure 4.1: The Schematic represents a fundamental way of getting explanation for a decision made by Deep Learning Models. [34]

Table 4.1: Importance of Explainability in AI Systems: Key Needs Addressed by XAI

Need	Explanation
Trust	Explainability build trust by showing AI reasoning.
Bias Detection	Identifies and addresses biases in AI systems.
Compliance	Meets regulatory requirements by explaining decisions.
Debugging	Pinpoints issues in AI for improvement.
Human-AI Collaboration	Enables better collaboration by understanding AI actions.

4.3.1 Limitations in Common Explainability Methods

Even though there has been a huge amount of work done for interpreting medical images, in which many are attribution methods. The problems with AI method's explainability are:

- Attribution methods highlight important areas in an image for CNN Prediction, but not the specific characteristics within the region of focus of the image.
- Texture characteristics within the region of activation are not transferred properly, as large parts of image are important for classification.

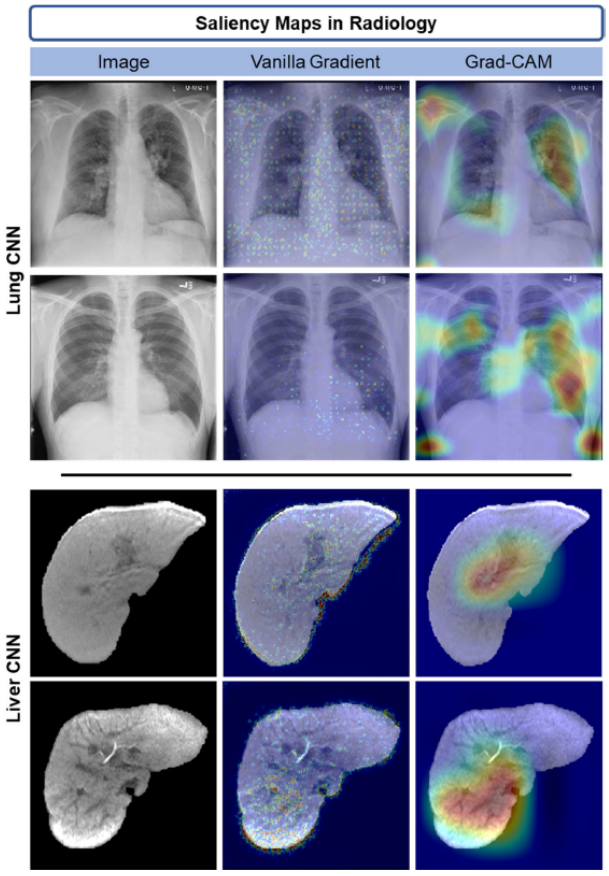


Figure 4.2: Example saliency maps from lung regression and liver classification CNN [20]

The attribution maps in Figure 5.1 highlight attention areas, yet they lack clarity on the specific relevant characteristics within these regions for diagnosis. Within these highlighted areas, various anatomic or texture features could potentially be responsible. In the chest radiograph algorithm, attention might focus on the vasculature, air spaces, or the cardiomediastinal silhouette edge. Similarly, in the liver MRI algorithm, attention might center on hepatic veins, encompassing aspects like the veins themselves, contrast with hepatic parenchyma, or texture.

Attribution: deciding how each feature value in an instance is important in obtaining specific outcomes

For the explainability of AI systems in medical imaging, the 'what' is as likely as important as the 'where'. Current methods provide localization information but their effectiveness is limited in medical imaging, where often multiple objects are spatially co-localized.

4.3.2 Other AI Explainability Methods

There are several special methods applied to medical imaging that use CNNs for Explainable AI methods. Some commonly used methods are Attribution Methods and Non-Attribution Methods.

Attribution Methods

Most explainable AI algorithms are attribution methods, which are a class of algorithms that highlight important areas in an image for a CNN's prediction. Different types of Attribution Methods are:

- Gradient-based Saliency Maps: The most popular form of attribution is gradient-based saliency, which computes the gradient of a prediction with respect to the pixels of the input image [20]. The visualization of these gradients allows for the interpretation of how input pixels or specific regions impact the CNN's predictions. The gradient-based attribution methods mostly differ in how the gradient is computed. The most popular such methods are Vanilla Gradient maps and Gradient-Weighted Class Activation Mapping

- Perturbation-based Saliency Maps: Perturbation maps are another form of saliency that visualize the effect of input feature perturbations on a CNN's prediction. Perturbed areas of the input image showing a relatively large effect on the CNN predictions are highlighted for interpretation [20]. Popular approaches are Shapley values and local interpretable model-agnostic explanations.

  Furthermore, while saliency maps effectively highlight clinically significant structures, but interpretation of these can be challenging, particularly in texture and morphology. And these maps are locally interpretable and do not offer a global understanding of CNN's predictive features since they're generated at the image level.

Non-Attribution Methods

Non-attribution methods prioritize the creation of alternative models or explanations without explicitly assigning significance to specific features. Rather than isolating the importance of individual features, these approaches aim to provide a comprehensive understanding of the model's behavior, offering a global view.

- Attention Networks Attention Networks integrate attention modules directly into the CNN architecture, aiding in both explaining CNN predictions and boosting CNN performance. Rather than employing attribution methods post-training, these modules operate within intermediate CNN layers as feature selectors, amplifying crucial features for prediction and dampening irrelevant ones. Visualizing these augmented feature maps reveals the vital features influencing CNN's predictions.

  However, this approach provides only a localized understanding of the features used for prediction. Attention networks, and CNNs in general, also contain a large number of feature maps within each layer of the network, many of which show little or no activation, making exploration of this feature space less tractable. [20]

- Generative Methods A novel method to enhance CNN explainability involves creating synthetic images to boost the activation of an output neuron or altering existing images to simulate a different output class. In initial activation maximization techniques, gradient descent was employed to produce an image that maximized a particular neuron's activation, effectively revealing the crucial imaging features influencing a CNN's prediction.

  However, the synthetic images produced by these methods do not appear realistic, which makes their interpretation quite difficult, especially in the context of medical images. The proposed model is a GAN-based explainability method that facilitates global interpretation of embedded features important for CNN prediction and shows the resulting embedded feature interpretations that can clarify the ambiguities observed in commonly used attribution maps [20].

4.3.3 Problems Statement

The research in the text addresses the problem of limited explainability in medical imaging classification tasks that use convolutional neural networks (CNNs). While CNNs have shown great performance in these tasks, their complex architectures make it difficult to understand the specific features used by the network to make its decisions. This lack of transparency limits the clinical application of these models and hinders the discovery of new biomarkers for diseases.

Current methods for post hoc explainability, such as attribution maps, provide only limited visibility into the characteristics used by CNNs during inference. These methods can highlight areas of attention in an image, but it is not clear what specific characteristics within those areas are relevant to diagnosis. Any of the anatomic or texture characteristics within the highlighted regions could be responsible for CNN's decision.

This problem is particularly relevant for diffuse disease processes where large portions of the image are potentially relevant for classification. For example, in the case of a chest radiograph algorithm designed to infer log NT-pro B-type natriuretic peptide (BNPP) to assess the severity of pulmonary edema, attention could be attributed to any of the following—the vasculature, the interstitium, air spaces. Similarly, for a liver MRI algorithm designed to classify images as having adequate or suboptimal contrast uptake for cancer screening and surveillance to assess the severity of liver fibrosis, attention is centered on hepatic veins but could be attributed to any of the following—the veins themselves, contrast with hepatic parenchyma, or texture.

Therefore, the research proposes a new framework called Feature Interpretation using Generative Adversarial Networks that allows for the dynamic visualization of the features used by a CNN. This framework provides a new level of understanding and explainability in medical imaging classification tasks, improving the transparency of these models for clinical application and potentially uncovering new biomarkers for disease.

4.3.4 Proposed Methodology

Generative Adversarial Networks are used in the proposed FIGAN framework to address the problem of limited explainability in medical imaging classification tasks. FIGAN uses CGAN to synthesize images that span the range of CNN's principal embedded features. The GAN is trained to generate images that are similar to the input images but with variations in the specific features of interest. These generated images can then be used to visualize the specific features used by the CNN for classification or regression.

By generating synthetic images that highlight the specific features used by CNN, FIGAN is expected to provide a new level of understanding and explainability in medical imaging classification tasks.

Overall, FIGAN leverages the power of GANs to generate synthetic images that help to clarify ambiguities within attention areas highlighted by existing explainability methods. This approach provides a new tool for interpreting medical imaging-related tasks.

4.3.5 FIGAN Architecture

FIGAN is a framework that uses a conditional generative adversarial network to synthesize images that is trained to generate images that are similar to the input images but with variations in the specific features of interest. These generated images can then be used to visualize the specific features used by the CNN for classification or regression.

The researchers apply FIGAN to two previously developed CNNs and show that the resulting feature interpretations can clarify ambiguities within attention areas highlighted by existing explainability methods. They also perform a series of experiments to study the effect of auxiliary segmentations, training sample size, and image resolution on FIGAN's ability to provide consistent and interpretable synthetic images.

The process (refer to fig 4.3) starts with an input image, and this image is given to a pre-trained CNN, the CNN extracts many features from the input image. These features can be extracted from any intermediate layer within the CNN, but the researchers focused only on the high-level features from the final fully connected layer, which are more likely to explain a CNN's decisions.

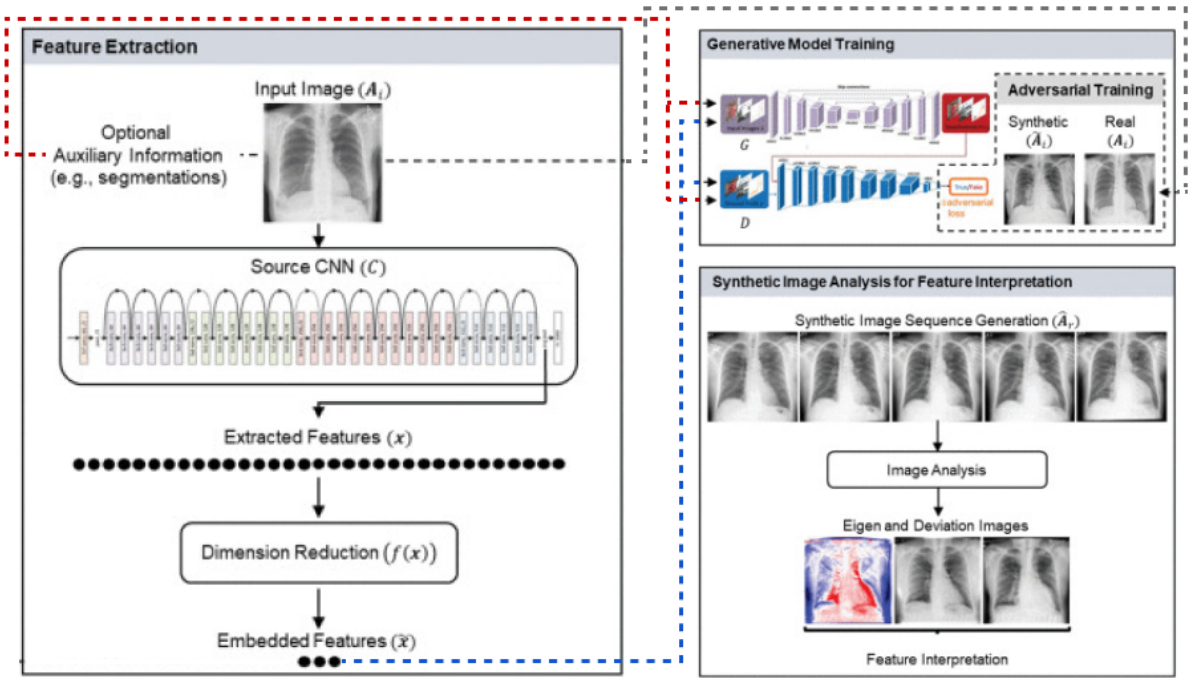


Figure 4.3: FIGAN Architecture

As the number of features are more in count, the unwanted features are reduced or deleted using a dimensionality reduction method just like Principal Component Analysis(PCA), but here the researchers have used a method called Partial Least Squares(PLS).

***PCA**: PCA is an unsupervised technique primarily used for reducing the dimensionality of data while preserving as much variance as possible. It doesn't consider any specific outcome or target variable.*

***PLS**: PLS is a supervised technique used for dimensionality reduction in the context of regression and predictive modeling. It focuses on finding the directions in the feature space that are most related to the target variable.*

The selected features are then given to the CGAN Generator and Discriminator for training the GAN to make realistic images. Now the images made by the CGAN or the synthetic images have most features that actually contribute to CNN's decision-making process, these images are then analyzed, and can be viewed as a Graphics Interchange Format (GIF) or under a medical image viewing device to understand and analyze whether the CNN has taken most relevant features into consideration or not. Thereby getting explainability for the CNN AI system.

4.4 Results and Discussion

Understanding the intricate mechanisms within Convolutional Neural Networks has been a long pursuit in computer vision and deep learning. In this section, the focus is directed towards unwind the output produced by the proposed FIGAN Model. This investigation not only promises insights into the inner workings of these models but also holds the potential to utilize these types of AI Systems for visual data analysis in the medical field. The researchers have applied FIGAN to two disease-analyzing CNNs, namely Lung CNN and Liver CNN. They chose three common steps to come up with results from the regenerated images(using FIGAN). They are:

- Feature Extraction and Selection
- Feature Interpretation, and
- Comparison to Visual Geometry(VG) and Grad-CAM

Feature Extraction and Selection

The methodology to get to result was to choose three to five feature maps for regeneration using FIGAN at training steps of 470,000 and the feature map with minimum Fréchet Inception Distance(FID) was chosen for subsequent analysis. And is reported that PLS components have variations in log BNPP. But this result of log BNPP should be interpreted with caution, as Lung and Liver results should be finally interpreted by an expert as per inference table *Applications of explainability in medical imaging [18].*
The Fréchet Inception Distance is a metric used to evaluate the quality of images generated by generative models, especially Generative Adversarial Networks. It measures the similarity between real and generated images by comparing feature representations extracted from a pre-trained Inception network.

Feature Interpretation

Lung CNN Feature Interpretation: The authors analyzed the synthetic image sequence and the leading eigen and deviation images for each feature. They found that feature 1 and feature 2 emphasized the importance of cardiomegaly (enlarged heart) and increased perihilar vascularity for predicting log BNPP. Feature 2 also highlighted the importance of the chest wall soft tissues. However, the authors note that feature 3 only explained 1.6% of variability in log BNPP and must be interpreted with caution.

Liver CNN Feature Interpretation: The authors noted that feature 1 emphasized the importance of vessel-liver contrast, which is the primary imaging feature used by radiologists to determine contrast uptake adequacy. Feature 2 highlighted the importance of heterogeneous texture, which is a known correlate of liver pathology. Feature 3 emphasized the importance of liver brightness, which is also a known correlate of liver pathology. The authors note that features 2 and 3 are secondary features of importance for contrast uptake adequacy.
In radiology, "contrast uptake adequacy" refers to the extent or quality of contrast agent absorption or enhancement in a specific organ or tissue region during imaging procedures like CT scans, MRI scans, or other imaging modalities.

Comparison to VG and Grad-CAM

The authors compared the results of using FIGAN to visualize learned features in lung and liver CNN tasks to the results obtained using VG and Grad-CAM. In the lung CNN task, VG showed diffuse network attention scattered throughout the radiograph, while Grad-CAM showed attention toward the left lung and right hilum but could not resolve which structures in these areas drove the regression of NT-proBNP. In contrast, FIGAN images provided insight into variations in heart size and hilar vascular fullness. In the liver CNN task, both VG and Grad-CAM showed attention toward the portal vein, but the meaning of activations scattered throughout the liver was unclear. FIGAN images highlighted liver texture and poor liver vessels, which are associated with inadequate contrast uptake and liver fibrosis.

Overall, the results of applying FIGAN to the lung CNN for a regression task provided insights into the imaging features used to predict log BNPP. On the other hand, the results of applying FIGAN to the liver CNN for a classification task provided insights into the imaging features used to determine contrast uptake adequacy. The authors were able to identify primary and secondary imaging features, as well as correlations with the body's physical characteristics of an individual, which may be useful for improving the accuracy and interpretability of medical imaging classification tasks.

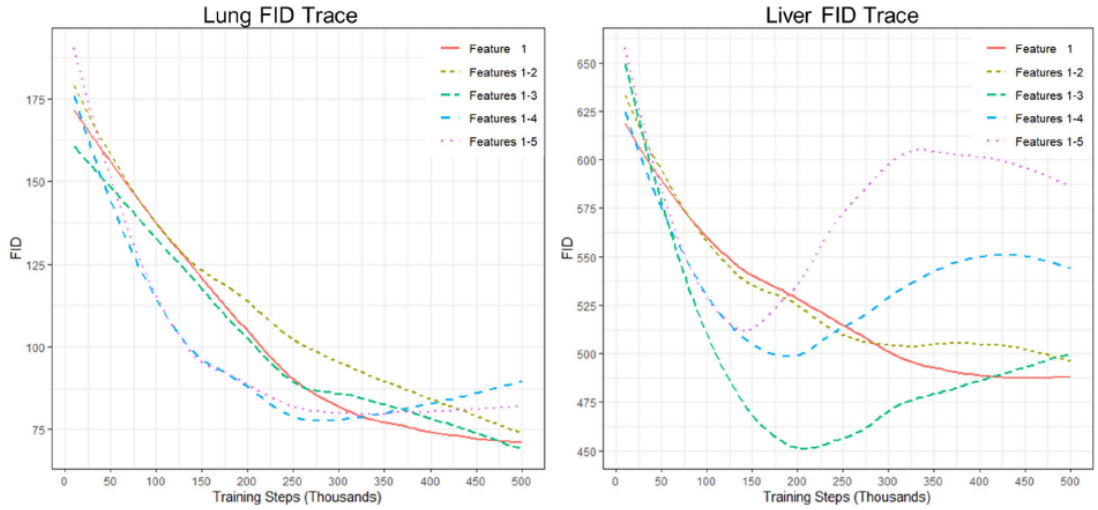


Figure 4.4: Top three features of Lung CNN shown best FID in FIGAN training. While liver CNN, the first three features.

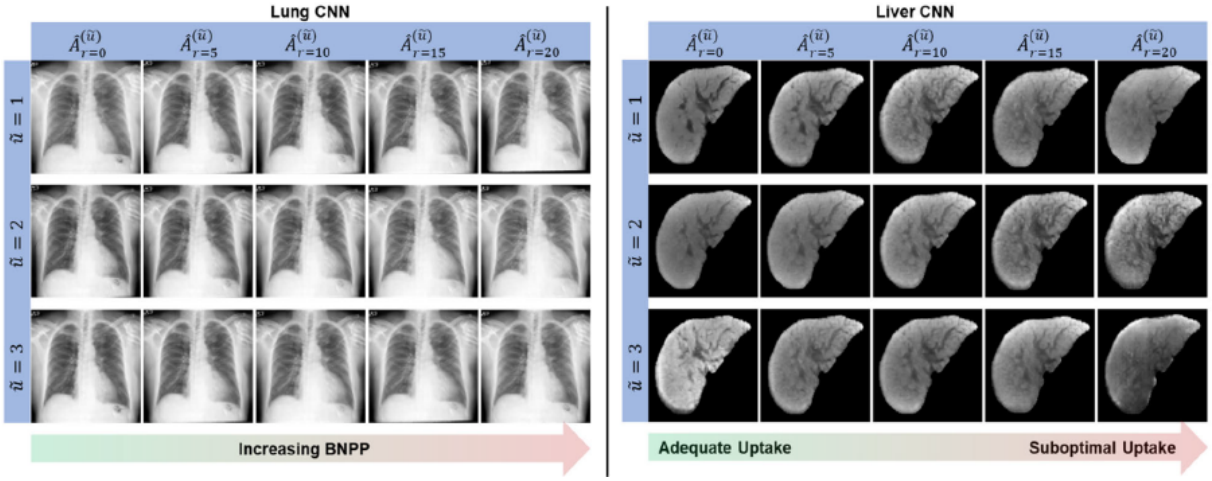


Figure 4.5: In chest x-rays, variations emerge in heart size and hilar vascular fullness. While liver MRI shown changes in vessel-liver tissue contrast and liver texture.

4.5 Applications

The overall work presented in this text has several potential applications in the world of medical imaging and beyond.

- The proposed FIGAN framework provides a new level of understanding and explainability in medical imaging classification tasks. This improved transparency of CNN models can lead to better clinical application of these models and potentially uncover new biomarkers for disease.
- Secondly, the use of GANs in FIGAN has broader implications for the field of explainable AI. The ability to generate synthetic images that highlight specific features used by a CNN could be applied to other domains beyond medical imaging, such as natural language processing or computer vision.
- The experiments conducted in this work provide insights into the effect of auxiliary segmentations, training sample size, and image resolution on FIGAN's ability to provide consistent and interpretable synthetic images. These findings could inform the development of future explainability methods and improve the interpretability of CNN models in various domains.

Overall, the work presented in this text has the potential to improve the transparency and interpretability of CNN models, leading to better clinical application and potentially uncovering new biomarkers for disease. The use of GANs in FIGAN also has broader implications for the field of explainable AI.

4.6 Challenges

The researchers in this text highlight several challenges faced in the field of explainable AI, specifically in the context of medical imaging classification tasks.

One challenge is the limited explainability of CNN models due to their increasingly complex architectures. While these architectures model highly nonlinear relationships in imaging tasks, they also make it difficult to explain the imaging features used in each image evaluated by the CNN.

Another challenge is the limited effectiveness of current methods for CNN explainability in medical imaging, where multiple objects are spatially co-localized. Current methods provide localization information, but the relative importance of individual characteristics within the localized region is unknown.

The researchers also note that existing explainability methods, such as attribution maps, highlight areas of attention but do not clarify what specific characteristics within those areas are relevant to diagnosis. This ambiguity can limit the clinical application of CNN models.

Overall, the challenges faced by the researchers in this text relate to the limited transparency and interpretability of CNN models in medical imaging classification tasks. These challenges highlight the need for new methods, such as the proposed FIGAN framework, to improve the explainability of CNN models and enable better clinical application.

Chapter 5

Conclusion

Generative Adversarial Networks have revolutionized the field of Deep Learning by introducing a novel framework for Generative Modeling. Their ability to generate data that is indistinguishable from real samples has opened up numerous possibilities in fields such as computer vision, natural language processing, and data generation. GANs have been employed in creating lifelike images, enhancing image super-resolution, generating realistic human-like text, and even in drug discovery.

However, it's important to acknowledge that GANs are not without challenges. The training of GANs can be notoriously unstable, and achieving convergence is often a complex task. Moreover, GANs can be susceptible to ethical concerns, including deepfake generation, data privacy issues, and potential misuse. Safeguarding against these challenges is vital to ensure responsible and ethical use of GAN technology. As the field of GANs continues to evolve, the possibility of the development of more stable training techniques is very high, as well as innovations to address ethical concerns. GANs are here to play a pivotal role in the future of AI, providing new ways to generate and manipulate data, and their potential applications are vast and ever-expanding. Nonetheless, it is crucial that the research and implementation of GANs go hand in hand with ethical considerations and responsible AI practices to fully harness the transformative power of this technology.

‘

5.1 Future Scope

The very introduction of GANs has created a boost for Generative AI, which has been a storm for the AI world, that still going on with astonishing Foundation Models and Transformers. So the future of GANs is really bright, appending some possible future scopes for GANs:

- Training GANs on fewer data points: GANs can be made to be trained on fewer data points than other generative models, making them useful for creating virtual environments where data is limited.
- Simpler but robust architectures for GAN to train on: Researchers can develop simpler but robust architectures for GANs, which will make them easier to use and more widely applicable.
- Creating virtual worlds for Movies and Metaverse: GANs can be made to generate high-quality landscape and scenery images in the metaverse, enabling the creation of realistic and immersive natural scenes. They can also be used to create virtual worlds for movies and other media.
- Making multi-modal content: GANs can be used to create multi-modal content that includes video, images, and sounds which can be used to express the creativity of an individual.
- Brainstorming creative ideas: GANs can be used to generate unique and creative artwork in the metaverse, such as virtual art galleries filled with AI-generated masterpieces. They can also be used to generate new building or room designs

5.2 Limitations

Even if GANs can do wonders in the world of text and image data, they come with their own challenges, let us explore two main challenges of GAN in detail and see other challenges in short.

- Mode Collapse: GAN fails to capture and generate diverse samples from the entire data distribution. Instead, it may focus on generating samples from only a few modes or patterns in the data. This results in a lack of diversity in the generated outputs. It is a challenge as it makes the generated data less useful for tasks like data augmentation or creative content generation.

- Non-Convergence: The discriminator performance gets worse when the generator outperforms it, thus lowering the accuracy of the discriminator. It is a challenge as discriminator feedback gets less meaningful over time and gives random feedback to the generator, by which the generator starts to train on junk feedback, reducing its own quality.

- Computational cost: GANs can require a lot of computational resources and can be slow to train, especially for high-resolution images or large datasets.

- Overfitting: GANs can overfit the training data, producing synthetic data that is too similar to the training data and lacking diversity.

- Bias and fairness: GANs can reflect the biases and unfairness present in the training data, leading to discriminatory or biased synthetic data.

- Interpretability and accountability: GANs can be opaque and difficult to interpret or explain, making it challenging to ensure accountability, transparency, or fairness in their applications.

References

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[2] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," 2020.

[3] V. Bilgram and F. Laarmann, "Accelerating innovation with generative ai: Ai-augmented digital prototyping and innovation methods," *IEEE Engineering Management Review*, vol. 51, no. 2, pp. 18–25, 2023.

[4] J. Liang and J. Chen, "Data augmentation of thyroid ultrasound images using generative adversarial network," in *2021 IEEE International Ultrasonics Symposium (IUS)*, 2021, pp. 1–4.

[5] G. Marti, "Corrgan: Sampling realistic financial correlation matrices using generative adversarial networks," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8459–8463.

[6] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2022.

[7] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," *Parallel Distributed Process*, vol. 1, 01 1986.

[8] S. Brooks, "Markov chain monte carlo method and its application," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 47, no. 1, pp. 69–100, 1998. [Online]. Available: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9884.00117

[9] I. Goodfellow, "Nips 2016 tutorial: Generative adversarial networks," 2017.

[10] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014.

[11] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *CoRR*, vol. abs/1511.06434, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:11758569

[12] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," 2019.

[13] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 7354–7363. [Online]. Available: https://proceedings.mlr.press/v97/zhang19d.html

[14] Y. Jiang, S. Chang, and Z. Wang, "Transgan: Two pure transformers can make one strong gan, and that can scale up," 2021.

[15] Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan, and Y. Zheng, "Recent progress on generative adversarial networks (gans): A survey," *IEEE Access*, vol. 7, pp. 36 322–36 333, 2019.

[16] D.-L. Way, C.-H. Lo, Y.-H. Wei, and Z.-C. Shih, "Twingan: Twin generative adversarial network for chinese landscape painting style transfer," *IEEE Access*, vol. 11, pp. 60 844–60 852, 2023.

[17] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1558–1566. [Online]. Available: https://proceedings.mlr.press/v48/larsen16.html

[18] S. S Band, A. Yarahmadi, C.-C. Hsu, M. Biyari, M. Sookhak, R. Ameri, I. Dehzangi, A. T. Chronopoulos, and H.-W. Liang, "Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods,"

Informatics in Medicine Unlocked, vol. 40, p. 101286, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914823001302

[19] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that?" 2017.

[20] K. A. Hasenstab, J. Huynh, S. Masoudi, G. M. Cunha, M. Pazzani, and A. Hsiao, "Feature interpretation using generative adversarial networks (figan): A framework for visualizing a cnn's learned features," *IEEE Access*, vol. 11, pp. 5144–5160, 2023.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, p. 84–90, may 2017. [Online]. Available: https://doi.org/10.1145/3065386

[22] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," 2019.

[23] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, and Y.-H. Yang, "Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment," 2017.

[24] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," 2017.

[25] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," 2020.

[26] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," 2018.

[27] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," 2018.

[28] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," 2017.

[29] G. Antipov, M. Baccouche, and J.-L. Dugelay, "Face aging with conditional generative adversarial networks," 2017.

[30] N. Aldausari, A. Sowmya, N. Marcus, and G. Mohammadi, "Video generative adversarial networks: A review," *ACM Comput. Surv.*, vol. 55, no. 2, jan 2022. [Online]. Available: https://doi.org/10.1145/3487891

[31] Y. Dan, Y. Zhao, X. Li, S. Li, M. Hu, and J. Hu, "Generative adversarial networks (gan) based efficient sampling of chemical composition space for inverse design of inorganic materials," *npj Computational Materials*, vol. 6, no. 1, Jun. 2020. [Online]. Available: http://dx.doi.org/10.1038/s41524-020-00352-0

[32] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," 2018.

[33] D. Li, J. Yang, K. Kreis, A. Torralba, and S. Fidler, "Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization," 2021.

[34] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *Journal of imaging*, vol. 6, no. 6, p. 52, 2020.