Classifying Real V.S. AI-Generated Artwork Images With ExplainableAI Analysis

Eugenie Choi

Advisor: Professor Xiaoyan Li

Submitted in fulfillment

of the requirements for the degree of

Bachelor of Arts

Department of Computer Science

Princeton University

April 2024

I hereby declare that I am the sole author of this thesis.

I authorize Princeton University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

_____

Eugenie Choi

I further authorize Princeton University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

_____

Eugenie Choi

Abstract

The currently existing artwork image dataset ArtBench was combined with images artificially generated by Stable Diffusion and DALLE-2 containing similar visual content. Two datasets were created, with half of each dataset consisting of real artwork images from the ArtBench-10 dataset, which contains 60K real artwork images from seven artistic styles. The other half of each dataset consisted of artwork images artificially generated by a text-to-image diffusion model under the same styles, depicting very similar content to that of the real images in the dataset. A discriminative convolutional neural network was then trained on the datasets to differentiate whether an image chosen from the test dataset was either a real or generated artwork image. The results of the discriminative model's ability to classify real vs. fake images were then evaluated through multiple experiments using precision, recall, and accuracy metrics. The classifier's evaluation metrics were high for both datasets, suggesting an overall success for the classifier's ability to differentiate between classes. The classifier's results were also visually interpreted using the ExplainableAI architecture LIME, which creates a heatmap of the features in each image that contribute to the class output label. Overall, the classifier results were very promising and LIME provided explanations for the classifier's decisions, proving that the classifier can be used as a successful and faithful detector against synthetic art fraud or fabrications.

Acknowledgements

I would like to thank my advisor, Professor Xiaoyan Li, for advising me throughout the semester and providing invaluable suggestions and feedback for my work. I would also like to thank my secondary reader, Professor Christopher Moretti, for agreeing to read and provide feedback on my thesis as well.

To me

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The advent of increasingly popular AI models and applications such as ChatGPT[1], which has been used for a variety of tasks such as schoolwork, information, and problem-solving, has raised many questions about how to differentiate human work from artificially generated work. This question is further extended concerning image generating models. These models can generate images based on a variety of input types, such as text, audio, or even other images. At a high level, diffusion models learn a denoising function by being trained at multiple noise levels to minimize error between predicted noise and the actual noise. Eventually, the function is able to generate a denoised image [7]. There are a variety of image synthesizing models that exist, including Generative Adversarial Networks and Variational Autoencoders as the earliest types. In the past, GANs have been the highest performing models for image generation tasks, as they are able to produce high fidelity images. However, there are some notable issues with using GANs, such as the models being extremely hard to train and scale and their inability to generate a diversity of images. More recently, diffusion models have surpassed these models in terms of accuracy and quality with the ability to generate images at higher qualities. These diffusion likelihood-based models can capture more diversity in images, are able to cover the entire training

[1]https://chat.openai.com/

distribution, and are much easier to train and scale [7]. For example, Dhariwal and Nichol discovered that, when conditional diffusion models were trained on the ImageNet dataset, they performed better when generating images in terms of FID score than the GAN models they tested as well [7].

Figure 1.1: **Stable Diffusion Generated Image (left) and Real Artwork (right).**

Figure 1.2: **A series of artwork-like images generated by DALLE-2.**

A few such popular text-to-image diffusion models are Stable Diffusion[2], Midjourney[3], and DALLE-2[4], which can artificially generate images and even artwork from

[2]https://stability.ai/news/stable-diffusion-public-release
[3]https://www.midjourney.com/home
[4]https://openai.com/dall-e-2

a simple text prompt. These images can be extremely realistic to the point where they are, from a human perspective, almost indistinguishable from real-world images. Because of this, researchers have been exploring the role of diffusion models in image synthesis, as well as attempting to discover the features that can help differentiate the two types of images.

Figure 1.3: **A series of artwork-like images generated by Stable Diffusion.**

While generating synthetic images has been helpful in allowing artists and creative users to create artwork and designs quickly and easily, this newfound ability for anyone to create has led to some disputes. For example, the use of AI generated artwork can be used for malicious purposes, such as copying an artist's work and claiming it as one's own. As Baraheem and Nguyen note, the widespread use of diffusion models may give users the ability to improve their media online, but it raises issues to security and authenticity. For instance, as artificially-generated images become increasingly more photo-realistic, this could be an issue for forensic/crime cases for which false evidence could be generated or planted [1].

For example, between the sets of images in 1.1, it may not be clear to the average person which image is real and which is artificially generated. This is an even bigger question considering Figures 1.2 and 1.3, as the images generated appear to be in-

credibly similar to real life artworks. Thus, it can be especially helpful in the realm of digital art to have a method for differentiating between real and fake artwork, such as when selling or buying artwork or digital work. Therefore, this thesis aims to effectively train a discriminative network to distinguish between real and fake artwork images online and explore the feature segments that are used to differentiate them through the use of Explainable AI.

Chapter 2

Background and Related Work

There is much quantitative work that has been done on determining how to classify real versus artificially generated artwork images, as well as using artificial intelligence to detect artificial intelligence. Similarly, much research has been conducted on classifying the artistic styles of artwork images using convolutional neural networks (CNNs) [5]. The problems for both subjects are similar: both utilize convolutional neural networks to make these decisions. For the problem of classifying artistic styles, many researchers employ the use of convolutional neural networks to output one of many classification labels for the input artwork image, while the issue of using AI to detect AI should theoretically also use CNNs as a binary discriminator to output one of two labels for an input image.

Previous research conducted on the latter subject mostly involves the use of human participants to identify whether artwork images are real or fake, discussing the ethical or philosophical implications of generating artificial art based on real training images depicting the work of human artists, or evaluating the artistic quality of artificially generated art. For example, Chen et al. defined artistic qualities that artificially generated art shares with real artwork [6], while Samo and Highhouse conducted a survey of human participants to gauge their ability to differentiate between real and

generated art and their preferences between the two [15].

There is also research done on utilizing machine learning models to conduct the process of differentiating between real and fake artwork, which is essential for determining if AI can detect the difference between real and synthetic artwork more accurately than humans can. For example, Baraheem and Nguyen detected images generated by Generative Adversarial Networks (GANs). These GANs are made up of a generative and discriminator model, which work collectively to generate fake images that cannot easily be detected by the discriminator. The generative model begins by passing a sample of random noise to the discriminator, which will classify the image as fake. The discriminator penalizes the generative model every time it classifies an image as fake, which the discriminator learns from and, over time, the generative model begins to produce more plausible images that trick the discriminator [1]. Tan et al. proposed a specific type of GAN architecture, called ArtGAN, to generate high quality artwork images by including "back-propogation of the loss function with respect to the labels to the generator from the discriminator [16]." Baraheem and Nguyen utilized pre-trained CNNs to classify images in a dataset they generated utilizing a variety of GAN architectures, with large successes [1]. However, there is the question of whether a similar result of classification can be achieved with images generated by diffusion models, which are another type of image-generating AI model that have been shown to generate higher quality images than GANs.

Diffusion models are a type of generation model that have recently gained popularity among researchers and the public alike. Broadly speaking, diffusion models generate images by sampling from a distribution and reversing a gradual noising process. The sampling starts from noise $x_T, x_{T-1}, x_{T-2}, ...$ to the denoised final sample x_0. In order to reach a less noisy step x_{T-1} from x_T, the model $\epsilon_\theta(x_t, t)$ is used to calculate how noisy a sample is. This function is then trained on the objective $||\epsilon_\theta(x_t, t) - \epsilon||^2$, to minimize the mean-squared error loss between the true noise and

calculated noise [7]. This denoising model, trained on multiple noise levels, can then be used to produce high-quality images. Diffusion models have been shown to be superior to GANs, with a higher Frechet-Inception Distance score compared to GANs on unconditional image generation [7]. Therefore, training a classifier on a dataset of higher-quality images generated by diffusion models instead of GANs could lead to different results, especially as diffusion models are more publicly widespread than GANs to the average user.

When classifying images as either real or fake, a classification method like the discriminator convolutional network model found in generative adversarial models can be used. A discriminator model's function is to build upon the features and patterns of the images that the model utilized in its decision-making to classify the artwork image as either real or fake [4]. Castellano and Vessio provided an overview on multiple deep learning approaches to pattern extraction in paintings and drawings, including convolutional neural networks, generative adversarial networks, and recurrent neural networks, which all mainly work to learn from a sample distribution and output a result after calculating appropriate weights or distributions [4]. A similar approach involves that of research done on classifying artistic styles of artwork images, which also utilizes convolutional neural networks. For instance, Gultepe et al. utilized a method of unsupervised feature learning with K-means to extract the features of digitized paintings and then a support vector machine algorithm to classify the style of new test paintings (eventually grouping them into styles such as Baroque, Impressionism, Romanticism, etc.) [8]. Similarly, Chen et al. utilized convolutional neural networks to extract the features of Chinese classical paintings and group the paintings based on the similarity of their features [5]. However, these classifiers still remain largely uninterpretable in their decisions to users. In order to provide solutions for this issue, a series of models labeled as Explainable AI have been created.

Explainable AI aims to explain a Deep Neural Network model's results and pro-

cesses, especially as these DNNs tend to have little explanation for their outputs. For example, Bird and Lotfi utilized a specific method in classifying real versus fake images of real-world objects. Using the XAI heatmaps of the images, they discovered that the small visual imperfections or inconsistencies in the synthetic images likely caused the model to classify an image as fake, rather than the actual subject entity of the image itself, as the fake images were able to replicate the entities themselves almost perfectly [2].

Using this research, we built our dataset generation and classification methods based on previous approaches. Similar methods to the those mentioned above concerning image generating models, classifier methods, and Explainable AI models were used to generate and classify artworks.

Chapter 3

Approach

Two datasets were created, one with a mix of fake artworks generated by Stable Diffusion [14] and real artworks form the ArtBench dataset [9], and the second with a mix of fake artworks generated by DALLE-2 [12] and real artworks form the ArtBench dataset. For each real artwork in ArtBench, a fake image based off of the content in the artwork was generated from either Stable Diffusion or DALLE-2, providing a 1:1 ratio of fake artworks and real artworks in each dataset (except for the DALLE-2 dataset, which had significantly fewer generated artworks, explained further in the limitations section of this thesis). Then, a convolutional neural discriminative network was created that could take a test artwork as input and, trained on either dataset, could determine whether the artwork was real or fake. To evaluate the classifier model's performance, a variety of experiments were carried out, such as comparing the model's accuracy when trained and tested on the Stable Diffusion dataset versus the DALLE-2 dataset. The model was also evaluated using Explainable AI, which is a term that encompasses multiple packages that are able to provide visual explanations for model results. This provided a more qualitative, human-readable analysis of the model's results.

3.1 Stable Diffusion

Stable Diffusion is a text-to-image model that is conditioned on a CLIP ViT-L/14 text encoder and pre-trained on a subset of the LAION-5B database [3]. We utilized the Stable Diffusion V1-2, which is a checkpoint from v1-1 trained on "515,000 steps at resolution 512x512 on 'laion-improved-aesthetics' (a subset of laion2B-en, filtered to images with an original size $\geq$ 512x512, estimated aesthetics score $>$ 5.0, and an estimated watermark probability $<$ 0.5. The watermark estimate is from the LAION-5B metadata, the aesthetics score is estimated using an improved aesthetics estimator) [14]."

Other than text-to-image generation, the Stable Diffusion model is also able to perform image modification tasks such as inpainting, outpainting, and image translation [3]. For the purposes of this study, we utilized the HuggingFace Stable-Diffusion-V1-2 model pipeline[1].

3.2 DALLE-2

DALLE-2 is another text-to-image model that is trained on 650 million image-text pairs scraped from the Internet. The researchers invovled with creating DALLE-2 and OpenAI proposed a new kind of text-to-image generation using unCLIP, in which a CLIP text embedding is fed to a diffusion prior, producing an image embedding, which is then conditioned on a diffusion decoder, resulting in a final image [12]. When compared to other text-to-image models such as the previous iteration of DALLE-2 and GLIDE, Ramesh et al. discovered that the final image samples are comparable in quality to those generated by the other models and more diverse in content generation [12].

Other than text-to-image generation, DALLE-2 is also able to produce image

[1]https://huggingface.co/CompVis/stable-diffusion

variations, providing interpolations between two seemingly unrelated images [12]. For the purposes of this study, we utilized EdenAI, a third-party API that allows users to run batch jobs using diffusion model APIs, including batch text-to-image generation[2].

3.3 Dataset

In this section, two datasets that were generated and utilized for training and testing are introduced. For the purposes of my thesis, two new datasets were created using existing artwork images from ArtBench, which is a corpus of around 60K real artwork images spanning seven artistic styles/eras, including Realism, Romanticism, Renaissance, Art Nouveau, Expressionism, Baroque, Impressionism, and Surrealism. These images are high quality, taken from a pre-existing dataset WikiArt and other websites housing artwork images. They have been cleanly annotated, filtered, preprocessed, and formatted in an accessible format for machine learning purposes. The license is available for use under a Fair Use license.

3.3.1 Artwork Styles

As mentioned previously, the DALLE-2 and Stable Diffusion datasets mimic artworks from eight different artwork eras/styles that are included in the ArtBench dataset, including Realism, Romanticism, Renaissance, Art Nouveau, Expressionism, Baroque, Impressionism, and Surrealism. These categories come from different time periods of art history, expressing a unified painting style for each period.

Renaissance

Spanning from 1420 to 1520, the Renaissance era consisted of paintings that focused on humans and the natural, realistic world surrounding them, inspired by Greek

[2]https://docs.edenai.co/reference/start-your-ai-journey-with-edenai

and Roman art. Influential artists from this period include Leonardo da Vinci and Michelangelo [10].

Baroque

This era (1590-1760) consists of people of power and royalty, like kings and popes, utilizing art to celebrate their wealth and almost God-like status. These paintings included materials of wealth such as gold and marble, and they depicted magnificent and over-exaggerated, unrealistic scenes [10].

Romanticism

From 1790 to 1850, the Romanticism era consisted of paintings that were emotional, sentimental. The natural world was a large focus of Romantic artists during this era. This was one era that had less of a determined style compared to the other eras [10].

Impressionism

Impressionism marks the beginning of the modern art of art in the 20th century, stemming from 1850-1895. This era was characterized by expressive, almost furious brushstrokes in paintings, with the scenes depicted in paintings becoming more blurred and less distinct. Artists like Vincent Van Gogh and Claude Monet were hallmark artists of this era [10].

Realism

From 1850 to 1925, Realism focused on expressing the reality of the world, including the good, bad, and ugly. Artists focused on revealing the true nature of people and animals, rather than depicting something idealistic or necessarily beautiful [10].

Art Nouveau

The Art Nouveau era, spanning from 1890 to 1910, boasts a particular distinct style with curved lines, floral depictions, and stylized human figures. Many paintings within this era include symmetry and an air of youthfulness [10].

Expressionism

This era heralds a return to subjective, sentimental emotions in art. Expressionist paintings reflected political messages and an almost violent, aggressive brushstroke style, hinging on the abstract from 1890 to 1914 [10].

Surrealism

Surrealism focused on psychosomatic depictions of desires and hidden concepts, critiquing the beliefs of the upper class. Artists such as Salvador Dalí painting from their dreams, forming psychoanalytical images from 1920 to 1930 [10].

3.4 Classifier

Instead of using a pretrained classifier, a classifier was trained from scratch on the two datasets to test whether the classifier could accurately differentiate between the real and synthetic images given in the dataset. A convolutional neural network was trained to classify the artwork images as either "real" or "fake" utilizing the tensorflow Keras package and Sequential() model[3].

Using a similar implementation to the code outlined here[4], a network consisting of three sets of 2D convolutional layers and max pooling layers was created, including activation layers with relu and sigmoid functions at the end. The model's details are shown in Figures 3.1 and 3.2. This model framework was created as it was a framework

[3]https://www.tensorflow.org/guide/keras/sequential_model
[4]https://github.com/2spi/ai-v-real/blob/master/ai_real.ipynb

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 conv2d (Conv2D)              (None, 253, 253, 16)      784

 max_pooling2d (MaxPooling2D)  (None, 126, 126, 16)     0

 conv2d_1 (Conv2D)            (None, 123, 123, 32)      8224

 max_pooling2d_1 (MaxPooling2D)  (None, 61, 61, 32)       0

 conv2d_2 (Conv2D)            (None, 58, 58, 16)        8208

 max_pooling2d_2 (MaxPooling2D)  (None, 29, 29, 16)       0

 flatten (Flatten)            (None, 13456)             0

 dense (Dense)                (None, 32)                430624

 dense_1 (Dense)              (None, 1)                 33

=================================================================
Total params: 447,873
Trainable params: 447,873
Non-trainable params: 0
_________________________________________________________________
```

Figure 3.1: **Logs of the tensorflow keras Sequential model parameters and layers.**

used by a multitude of other researchers when classifying images. For example, Chen et. al utilized a CNN to classify different artstyles of Chinese classical paintings. They utilized a VGG-F model, a feedforward network with 5 convolutional layers and 3 fully connected layers, pretrained on the ImageNet dataset [5]. Similarly, the CNN that was implemented for this study had 3 convolutional layers and a flattening layer, employing a similar use of layers.

In addition, slightly different layers were experimented with to determine if there any differences in the accuracy of the output. For example, different epochs were tried, training the model on 20 epochs v.s. 100 epochs, which resulted in a much higher accuracy in the results. Adding another set of convolutional layer and max pooling layer was also attempted, which had little to no effect on the accuracy of the resulting classifier.

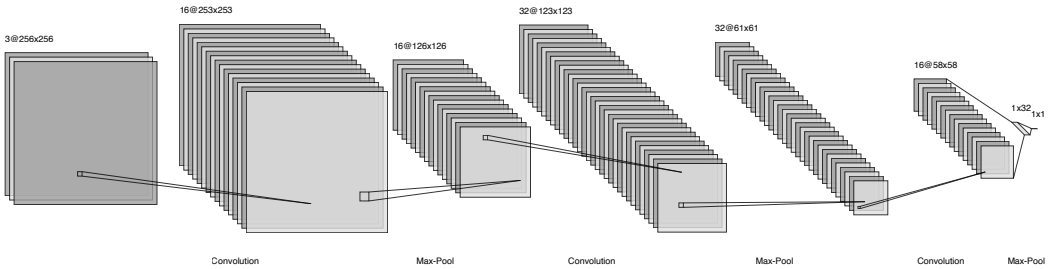

Figure 3.2: **Diagram of tensorflow Keras classifier.**

This approach provided a deeper and more quantitative insight into the question of whether synthetic artwork can be differentiated from real artwork, and if so, what the features that determine this decision are. This is essential research, especially since previous approaches have mostly relied on human evaluation to differentiate between real and synthetic images. It also provided a valuable look into the differences between text-to-image diffusion models such as DALLE-2 and Stable Diffusion and which one could be considered better at generating artwork-related images.

3.5 Experiments

In order to solve the question of whether a trained classifier could differentiate between real and fake images, seven experiments were set up that focused on running a classifier on certain subsets of the datasets. These experiments rely on the accuracy, precision, and recall metrics.

The first experiment involved training and testing the classifier on the full Stable Diffusion/ArtBench dataset and two classes (real and fake), splitting the dataset into a 60/20/20 training/validation/testing split. The accuracy of the classifier after this experiment would show how successful the classifier was in detecting which images in the testing set were real and which ones were generated by Stable Diffusion, after training on the images separated by class in the training set.

The second experiment is the same experiment as the first, using the DALLE-

2/ArtBench dataset instead. Again, there are two classes in this dataset, real and fake, and a second classifier is trained and tested on the entire dataset with a 60/20/20 split. The accuracy of the classifier should also show how accurate the classifier in detecting real and fake images generated by DALLE-2.

The third experiment involves taking the classifier trained on the Stable Diffusion/ArtBench dataset in Experiment 1 and testing it on a subset of the DALLE-2/ArtBench dataset. The goal of this experiment is to determine whether the classifier, when trained on images generated by Stable Diffusion, can also be used to detect fake images generated by other diffusion models. The accuracy of this classifier is a good indication of whether the classifier is a good general classifier, able to detect a wide variety of synthetic AI images, or if the classifier can only detect images generated by the diffusion model it was trained on successfully.

The fourth experiment is the same as the third, but instead taking the classifier trained on the DALLE-2/ArtBench dataset and testing it on the Stable Diffusion/ArtBench dataset. The goal of this experiment is the same as the previous experiment, as well as comparing the accuracy of the classifier in this experiment to the previous experiment to determine if one classifier performs better than the other. This would show which dataset is overall a better training dataset for future work in detecting fake artwork images over real artwork images.

Experiments 5 and 6 involve training and testing new classifiers only using subsets of the Stable Diffusion and DALLE-2 datasets, categorized by art styles. For example, the classifier would be trained and tested on a 60/20/20 split of every real and fake artwork in the style of Surrealism in the Stable Diffusion dataset, and its accuracy would be recorded for that art style. This process is repeated for every other art style in the dataset, such as Art Nouveau, Romanticism, Expressionism, etc. and then repeated for the DALLE-2 dataset. This would provide a good reading of whether any art styles were more difficult for the classifier to detect fake artworks under.

This could be helpful in understanding if some art styles may be harder for AI to emulate/detect. One such style could be Realism, since this many real paintings in this art style are meant to look photo-realistic and lifelike. Since most diffusion models are already very good at generating photo-realistic images, this could make it more difficult for a classifier to detect a fake artwork.

Finally, experiment 7 aims to determine the classifier accuracy v.s. the number of training images the classifier is trained on in both datasets. This provides a good estimate of roughly how many training images are needed for a classifier to provide a baseline accuracy (around 70%, for example). This would also show whether the problem of determining if AI can detect AI art is an inherently simple problem– although it may be hard for the human eye to detect fake AI art, it may not be as hard for AI to detect AI art.

3.6 Explainable AI

With AI beginning to take a more prominent role in important decisions, ranging from medical to business, it is becoming increasingly important for people affected by AI to understand the workings of these models making these decisions. In the past, researchers have used decision trees to explain the logic behind actions. However, as DNNs have become increasingly popular with researchers, these deep learning models do not have the capacity to provide an explanation for their outputs, and remain black-box models–models with internal processes that can't be observed or understood by the user. The higher the accuracy of the model, the more difficult the model becomes to explain [17].

To solve this issue, there are a variety of different Explainable AI (XAI) architectures that have been created to provide a human explanation for DNN outputs.

Some of these architectures include SHAP[5] and LIME[6], which can explain a variety of DNN types, such as multiclass NLP models and multiclass image classifiers. Both of these architectures provide visual explanations for these predictions. For instance, SHAP, or SHapley Additive exPlanations, explains model predictions by calculating every feature's contribution to the prediction. This XAI architecture utilizes Shapley values, or the average marginal contribution for each feature value for all coalitions, which are then mapped as vectors to a feature space, in order to explain model predictions [11].

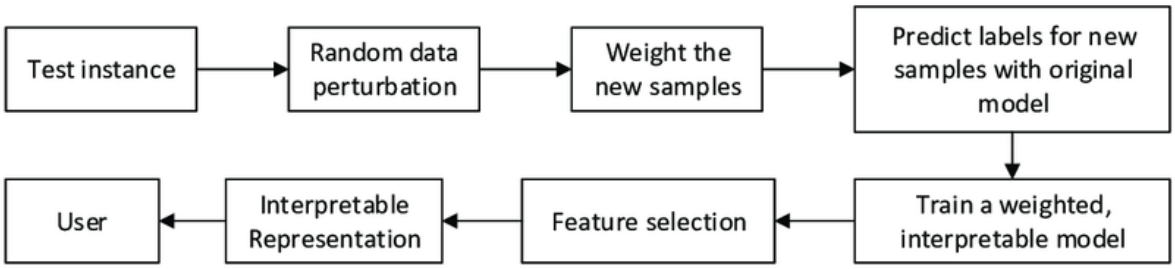

Figure 3.3: **diagram of the LIME framework [18, Figure 1].**

We decided to utilize the LIME architecture over SHAP, however, since it was better suited for the Keras classifier framework and datasets that were being utilized. SHAP's Deep Explainer model[7] required the dataset to be in a list of numpy arrays or torch tensors, while the datasets being used for the purposes of this study were Keras iterable datasets.

In addition, using SHAP's Deep Explainer required training the model over a subset of the training set of both datasets, which would have been time- and resource-consuming. On the other hand, using LIME's LimeImageExplainer meant that the model could be utilized without having to use training data. The LimeImageExplainer generates neighborhood data around features from the instance and learns

[5]https://github.com/shap/shap
[6]https://github.com/marcotcr/lime
[7]https://shap-lrjball.readthedocs.io/en/latest/generated/shap.DeepExplainer.html

linear models on the data, providing an explanation/interpretation for the model's prediction on that instance[8]. This includes providing visual artifacts for qualitative understanding of the relationship between the model's input data and the model's prediction, as seen in Figure 3.3.

LIME provides an explanation g for a classifier as a model within a class of interpretable models G such as decision trees or linear models. It does so by minimizing $L(f, g, \pi x)$, how unfaithful the explanation g is in approximating f, the probability of input x belonging to a certain class, while ensuring the complexity of the explanation $\Omega(g)$ is low enough to be interpretable by humans [13]. Essentially, LIME produces an interpretable explanation that is locally faithful to the classifier through this equation [13]:

$$\xi(x) = \arg\min g \in GL(f, g, \pi x) + \Omega(g)$$

Thus, LIME provides an explanation for individual predictions that should correspond to how the model behaves in the context of the individual prediction, while SHAP provides an explanation within the context of the global features [11]. For the purposes of this study, LIME was used to provide a local fidelity explanation in the context of predictions for individual images.

[8]https://lime-ml.readthedocs.io/en/latest/lime.htmllime.lime_image.LimeImageExplainer

Chapter 4

Implementation

4.1 Dataset

Using the real artwork images from ArtBench, two new datasets were created: one containing the ArtBench images as the real artworks and "fake" artworks artificially generated by the Stable Diffusion text-to-image model, and a second dataset containing the real ArtBench and fake artworks artificially generated by OpenAI's text-to-image DALLE-2 model. For text-to-image generation, we wanted to artificially generate a variety of artworks that would emulate the content and artistic style of the real artworks in the ArtBench dataset without being too similar. To do so, we standardized the text prompt engineering by creating a template for the prompts that would be either put into Stable Diffusion or DALLE-2 to output an image depicting artwork that was artistically similar to those in the ArtBench dataset, aiming for a 1:1 ratio of real/fake artwork in both datasets.

After experimenting with different prompt templates, which generated images we felt did not encapsulate the artistic style and content that I wanted, we ended up utilizing this template: "[Name of artist] [Name of real ArtBench artwork to emulate] in the style of [ArtBench artwork artstyle]." This was the best prompt to use, as not

including the name of the artist or the desired artstyle to emulate would sometimes result in the model generating a photo-realistic version of the content in the work instead of a synthetic artwork.

Figure 4.1: **Original ArtBench artwork image[1].**

Figure 4.2: **Image generated by Stable Diffusion with the prompt "Benjamin Brown view of a garden in the style of impressionism."**

For the Stable Diffusion/ArtBench dataset, a Python script was created (see Appendix A) in which the names, art style, and artist names of each of the real artwork images were first collected from the ArtBench dataset. Then, the prompt was created for each ArtBench image using this information as a text input to the Stable Diffusion

[1]Brown, Benjamin (American, 1865-1942). View of a Garden.

Figure 4.3: **Image generated by DALLE-2 with the prompt "Benjamin Brown view of a garden in the style of impressionism."**

HuggingFace API pipeline, which is available for public use at no additional cost, pre-trained on the HuggingFace CompVis/stable-diffusion-v1-4 model[2]. The generated images output were then combined with the ArtBench images to create the Stable Diffusion/ArtBench dataset.

A similar approach was used to generate the DALLE-2/ArtBench dataset in another Python script (see Appendix A). However, for the DALLE-2/ArtBench dataset, the synthetic images were generated using a third party API called EdenAI[3], which allowed batches of DALLE-2 images to be generated at once but required additional costs. Due to limited costs, a significantly less amount of DALLE-2 images were generated than Stable Diffusion images, which is explored more in the limitations section of this paper. We utilized EdenAI's image generation POST API method in order to send text prompts to their servers and run image generation batch jobs. Both the Stable Diffusion and DALLE-2 images were generated in batches by running Python scripts as jobs on Microsoft Azure. For the Stable Diffusion images, due to the nature of the pipeline, each image was generated separately, while the DALLE-2 images were generated in batches of 500 (with three to five batches running simultaneously).

The Stable Diffusion dataset was therefore generated with 44K synthetic images

[2]https://huggingface.co/CompVis/stable-diffusion
[3]https://docs.edenai.co/reference/start-your-ai-journey-with-edenai

and 54K real images (from ArtBench), while the DALLE-2 dataset was generated with 24K images and 54K real images. The two datasets were split 60/20/20 for training, validation, and testing.

4.2 Classifier

In order to gain the necessary resources for training and testing the model, Microsoft Azure's Machine Learning lab and Princeton's Della compute cluster were utilized, which provided the necessary GPU/CPU resources. Six different experiments on the model were performed: one training and testing the classifier on the Stable Diffusion/ArtBench dataset, the second training and testing the classifier on the DALLE-2/ArtBench dataset, the third training the model on the Stable Diffusion dataset and testing on the DALLE-2 dataset, the fourth training the model on the DALLE-2 dataset and testing on the Stable Diffusion dataset, and the fifth and sixth training and testing the classifier on the Stable Diffusion/ArtBench dataset on 8 different categories of artistic style and the DALLE-2/ArtBench dataset on the same categories, respectively. The seventh experiment involved training the classifier on varying numbers of data for both datasets, while keeping the test size the same. Each of these experiments were conducted on Princeton's Della computer cluster, which allowed for multiple concurrent running jobs that required GPU resources.

4.3 Explainable AI

To utilize the Explainable AI package, another Python script was created (see Appendix A) that would utilize LIME's LimeImageExplainer class[4], which creates an explainer variable that contains an "explain instance" function. Three random images were chosen, two generated by the DALLE-2 dataset and one from the ArtBench

[4]https://lime-ml.readthedocs.io/en/latest/lime.htmllime.lime_image.LimeImageExplainer

dataset to compare. The model was then used to figure out the model's prediction for those images and passed the data to the LimeImageExplainer. This function is able to learn a local model around the classifier's prediction on an image instance and generate an interpretable explanation for the prediction, which in this case was a heatmap of the image segments that showed how much that segment contributed to the model's prediction. Heatmaps were then generated for each image and overlaid them with the original image to determine which segments corresponded to which sections of the artwork.

Chapter 5

Evaluation

5.1 Preliminary Analysis

In this preliminary analysis, the classifier was trained on the dataset for seven different experiments. Three evaluation metrics were defined for these experiments: precision, accuracy, and recall. Precision is a measure of the number of samples a that are correctly predicted to be in the target class divided by the sum of a and the number of samples that are wrongly predicted to be in the target class b. Recall is the metric a divided by N, or the total number of samples of the target class in the test dataset [5]. Accuracy is the measure of the overall correct number of predictions over all classes c over the total number of predictions T over all classes.

$$Precision = a/(a+b)$$

$$Recall = a/N$$

$$Accuracy = c/T$$

Table 5.1: **Experiments 1-4: Testing the classifier on the datasets**

Experiment	Precision	Recall	Accuracy
Experiment 1	0.9967886805534363	0.9991722702980042	0.9969978332519531
Experiment 2	0.9986081719398499	0.9995356202125549	0.9986353516578674
Experiment 3	0.7561335563659668	0.9999070167541504	0.7632368803024292
Experiment 4	0.9988813400268555	0.9994403719902039	0.9987718462944031

Experiments 1-4: Testing the classifier on the datasets

This table displays the different experiments run using the classifer on the two datasets. The first experiment involved training and testing the classifier on the Stable Diffusion/ArtBench dataset, which yielded the respective metrics as shown in Table 5.1. The second experiment involved training and testing the classifier on the DALLE-2/ArtBench dataset. For the third experiment, the classifier was trained on the Stable Diffusion/ArtBench dataset and tested on the DALLE-2/ArtBench dataset. The fourth included training the classifier on the DALLE-2/ArtBench dataset and testing the model on the Stable Diffusion/ArtBench dataset.

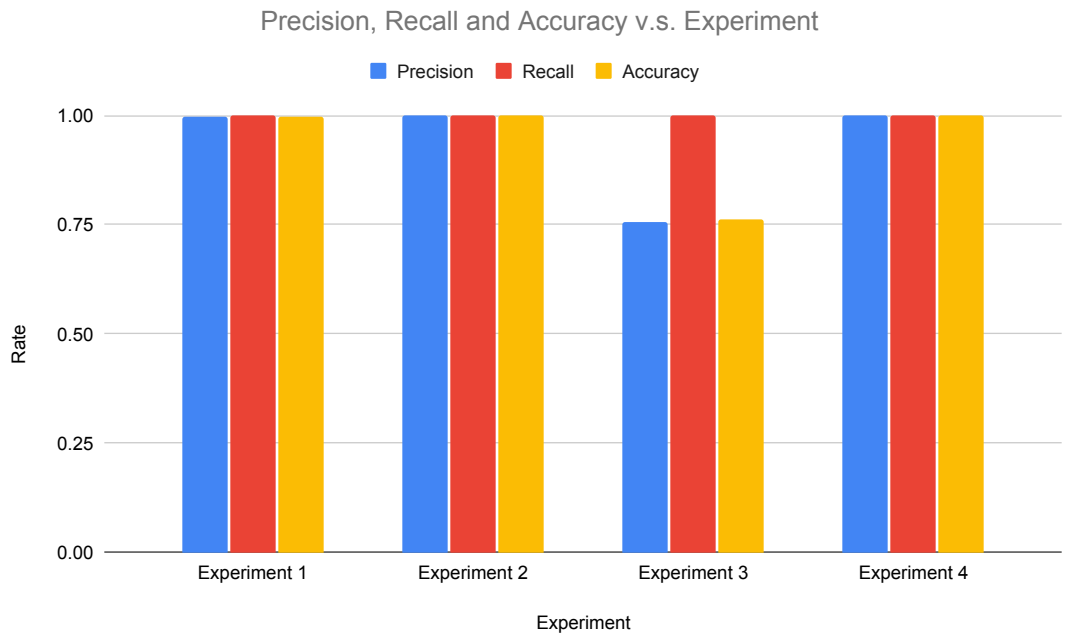

Figure 5.1: **Precision, Recall and Accuracy v.s. Experiment**

When trained and tested on the Stable Diffusion and DALLE-2 datasets, the classifier performed successfully with accuracy, precision, and recall rates at 99%. It is also clear that in the third and fourth experiments, the classifier trained on the DALLE-2 dataset performed better than the classifier trained on the Stable Diffusion dataset, despite the Stable Diffusion dataset containing more images/data than the DALLE-2 dataset. For example, the model trained and tested on the DALLE-2 dataset had a very similar accuracy of 99% to the model trained and tested on the Stable Diffusion dataset. However, the model trained on the DALLE-2 dataset performed significantly better when tested on the Stable Diffusion dataset than the model trained on Stable Diffusion and tested on DALLE-2, with an accuracy of 99% compared to 76% (see Table 5.1). This suggests that the DALLE-2 dataset is a better dataset for training, suggesting that the DALLE-2 model could be more accurate than Stable Diffusion in generating "realistic" artwork.

Experiments 5 and 6: Training and testing the classifier on the Stable Diffusion/ArtBench dataset on eight different categories of artistic style and training and testing the classifier on the DALLE-2/ArtBench dataset on seven different categories of artistic style

Table 5.2: **Accuracy of Classifier Model by Artwork Style v.s. Diffusion Model Used to Generate Fake Artworks.**

	Stable Diffusion/ArtBench Dataset	DALLE-2/ArtBench Dataset
Realism	0.8930457830429077	0.9725610017776489
Romanticism	0.9168074131011963	0.9350479245185852
Renaissance	0.9814189076423645	0.9783018827438354
Surrealism	0.8985655903816223	0.9533811211585999
Art Nouveau	0.9405381679534912	0.96435546875
Expressionism	0.9286157488822937	0.9807180762290955
Impressionism	0.9911184310913086	0.9967350959777832

To illustrate this further, looking at experiments 5 and 6, we see that when the

DALLE-2 and Stable Diffusion datasets are split by artistic style and the classifier is trained and tested on these subsets, DALLE-2 also performs better than Stable Diffusion in terms of accuracy in the same style categories. For example, when trained on the artworks in the style of Realism in the DALLE-2 dataset, the classifier is more accurate (97%) in its fake versus real classification, compared to when it is trained on the Realism artworks in the Stable Diffusion dataset (89%) (see Figure 5.2). In addition, the classifier trained on DALLE-2 artworks was able to classify artworks more accurately in the style of Surrealism, having a higher percentage of 95% compared to the classifier trained on Stable Diffusion artworks having an accuracy of 90% on Surrealism artworks.

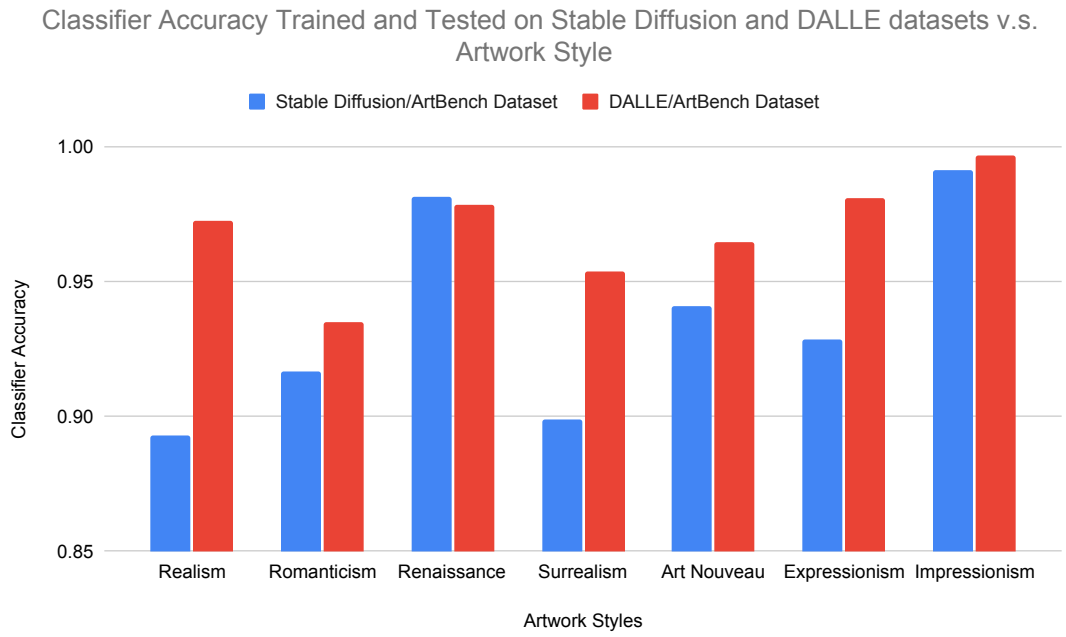

Figure 5.2: **Classifier Accuracy Trained and Tested on Stable Diffusion and DALLE-2 datasets v.s. Artwork Style**

We can also see that the classifier trained on the Stable Diffusion dataset was the least accurate when classifying art under the Realism and Surrealism art styles, while the classifier trained on the DALLE-2 dataset was the least accurate for the Romanticism artworks. On the other hand, both classifiers had the most accurate

performance for classifying artworks under the Impressionism style. Considering the results of the six experiments above, the classifier overall was highly accurate in classifying the "real" versus the artificially generated artwork images, with accuracies in the 90%-99% range.

Experiment 7: Classifier Accuracy v.s. Number of Training Images

Table 5.3: **Classifier Accuracy Trained and Tested on Stable Diffusion images v.s. Number of Training Stable Diffusion Images.**

Number of training images	Accuracy
900	0.6668536067
1830	0.7715640068
2730	0.7900183797
3660	0.7875204086
4590	0.7843596935
5505	0.8249897957

Table 5.4: **Classifier Accuracy Trained and Tested on DALLE-2 images v.s. Number of Training DALLE-2 Images.**

Number of training images	Accuracy
720	0.6319800019
1440	0.7425717115
2910	0.7449411154
4380	0.8420850635
5850	0.8034707904

For Experiment 7, Tables 5.3 and 5.4 display the classifier's accuracy when trained on different subset lengths of the datasets. The testing and validation sets for each dataset remains the same for each training set length, with each set containing the same 20% split of the dataset.

Experiment 7 shows the accuracy of the classifier when trained on different lengths of both datasets. For example, Figure 5.3 suggests that the classifier trained on the Stable Diffusion dataset reaches 70% when trained on a set of images between 1K and

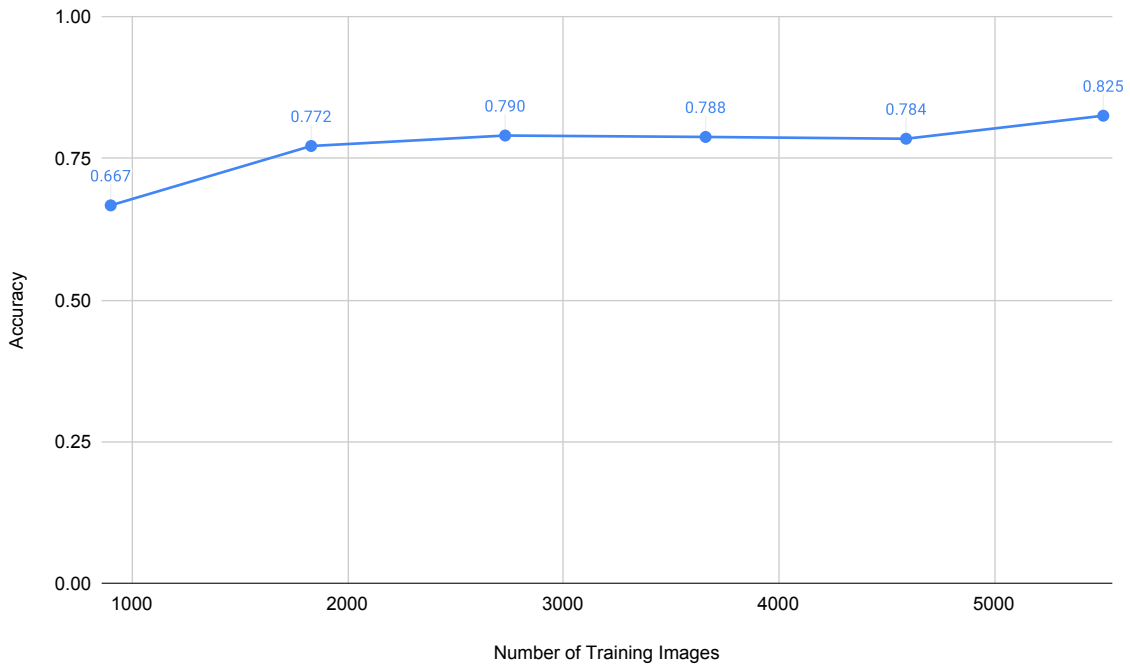

Figure 5.3: **Classifier Accuracy Trained and Tested on Stable Diffusion images v.s. Number of Training Stable Diffusion Images.**

2K. Similarly, in Figure 5.4, the classifier shows that when trained on the DALLE-2 dataset, the classifier can reach 70% between 1K and 2K as well. Trained on either dataset, the classifier is able to reach high levels of accuracy even at smaller training set sizes. This suggests that, for this specific task, it may not be necessary to have such a large training dataset. In addition, it might be an easier task for artificial intelligence models to detect artificially-generated images than expected, only requiring a small sample size.

Overall, this suggests that DALLE-2 could be a better text-to-image generator for artificially generating artworks. There are many reasons that could explain this—for example, Stable Diffusion has been found to generate more accurate photorealistic images of people and objects [3]. As many of the artworks in these datasets involve humans and objects, the classifier may interpret the images in the Stable Diffusion set to be too photorealistic or not similar enough to the less realistic, artistic styles

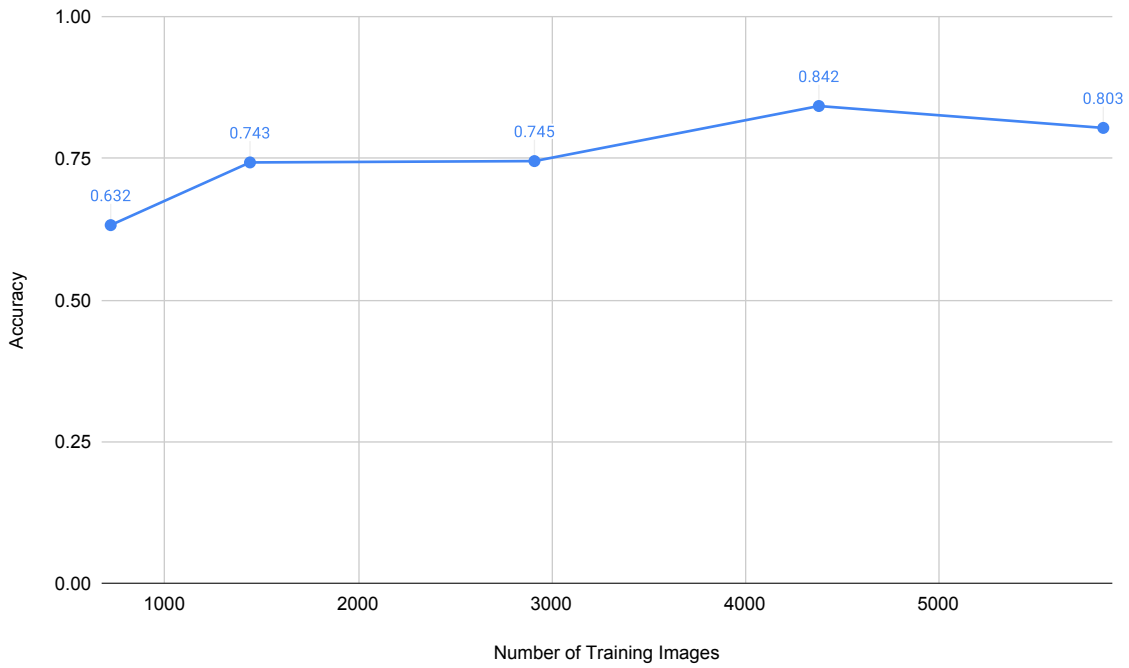

Figure 5.4: **Classifier Accuracy Trained and Tested on DALLE-2 images v.s. Number of Training DALLE-2 Images.**

of the artworks in the ArtBench dataset. On the other hand, Stable Diffusion may produce more artifacts or flaws within generated images than DALLE-2 does, allowing classifiers to easily detect the generated images as synthetic.

Despite these comparisons, it is clear that the classifier trained is a successful detector of synthetic artwork and can be used to accurately label images as real or fake.

5.2 ExplainableAI LIME Architecture

Overall, the ExplainableAI architecture provides a deeper insight into the decisions that the classifier makes when classifying images. As mentioned in previous sections, the LIME Explainable AI architecture has the ability to split images into segments, and then generate heatmaps of the segments. The segments are either colored red

or blue, with the darkness of the color corresponding to a larger contribution of the segment towards the model's prediction. A red segment indicates that the segment contributes negatively towards the model's prediction, while a blue segment corresponds to the segment contributing positively towards the prediction. For instance, if the model predicted an image to be fake, blue segments would indicate that the segment contributes positively to the overall image being a fake image, while red segments would indicate that the segment suggests that the image may be real instead (although not an overall prediction).

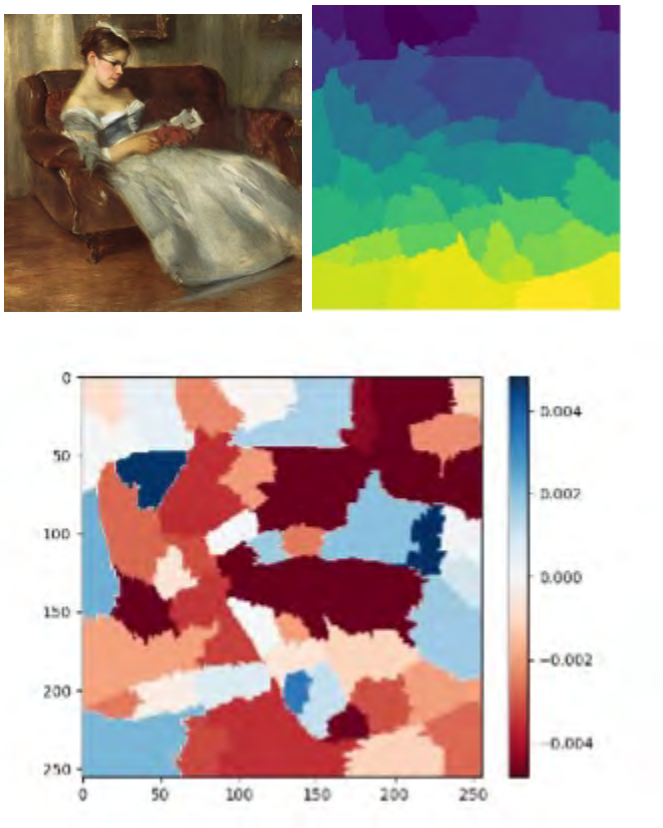

Figure 5.5: **DALLE-2-generated image (left), LIME image segments (right), LIME XAI heatmap of image(bottom).**

Conversely, if the model predicted that an image was real, blue segments would be segments supporting this prediction, while red segments suggest otherwise (that something in the segment suggests that the image may be a fake)[1].

[1]https://github.com/PacktPublishing/Applied-Machine-Learning-Explainability-

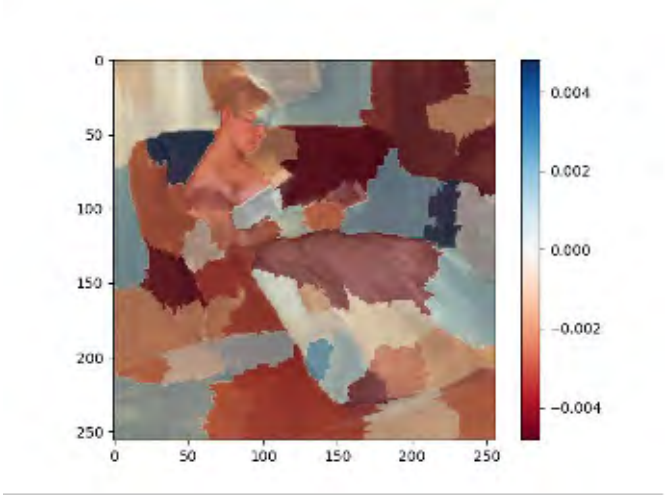

Figure 5.6: **LIME XAI heatmap overlaid with original DALLE-2-generated image.**

We can see that Figure 5.5 shows the segments that LIME splits the image into, which the classifier looks at when making its decision of whether the image is real/fake. In the third diagram, we can see that LIME creates a heatmap, with blue meaning the segment contributes positively to the classifier's correct predicted outcome of FAKE and the red meaning the segment contributes negatively to the outcome. Looking closely, we can see that when overlaid, the blue segment seems to match up with the woman's left hand. Looking at the woman's left hand, we can see that it looks a little off–there is a long line running from the space between her pointer finger and middle finger. Some other strange areas include the leg of the chair, which is positioned awkwardly and not necessarily straight, as well as a patch of her dress near the bottom that looks strangely out of place from the rest of the dress.

A similar example occurs with another AI-generated artwork containing humans, especially in Figure 5.7. Since the classifier predicted that this image was fake/AI-generated, we know the blue segments show sections of the image that contribute positively to the prediction. For example, we can see that the young girl on the left has multiple extra arms. We can see that parts of these extra arms are highlighted

Techniques/blob/main/Chapter05/LIME_with_image_data.ipynb

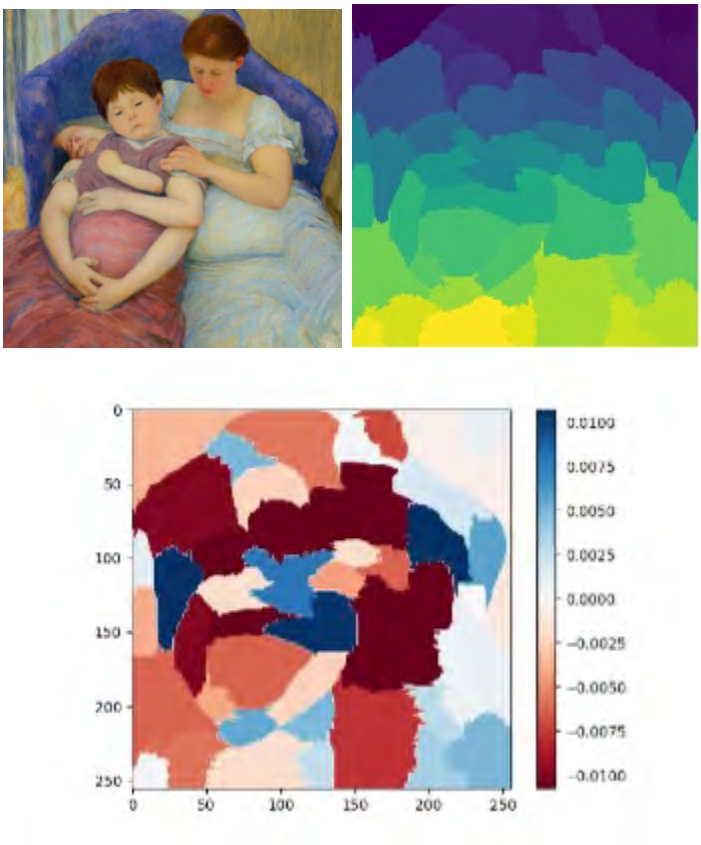

Figure 5.7: **DALLE-2-generated image (left), LIME image segments (right), LIME XAI heatmap of image(bottom).**

as blue by the classifier, showing that the classifier was able to detect these extra appendages as being indicative of the image being AI-generated.

However, there are some sections of the heatmap where the colorings are not as clear. For example, it can be seen that parts of the chair that the women are sitting on are colored blue, suggesting that, for some reason, the classifier flagged those parts of the chair as likely being fake. However, on closer inspection, it is difficult to see what exactly about those sections of the chair triggered that response–to the human eye, it appears to be a normal chair. This suggests that there might be some kind of hidden artifact or flaw that, while imperceptible to the human eye, is evident to the classifier in these particular image segments. Thus, artificial images generated by diffusion models may contain some artifacts that are easily detected by another

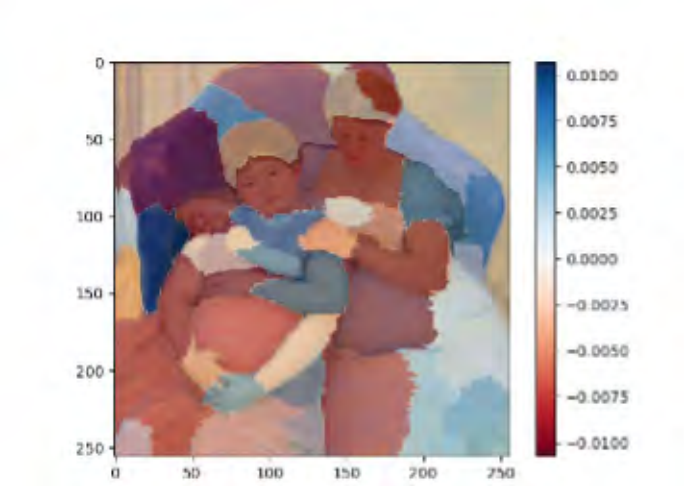

Figure 5.8: **LIME XAI heatmap overlaid with original DALLE-2-generated image.**

artificial intelligence, while being much more difficult to detect by a human.

For an image that the classifier predicts as a real image, such as in Figure 5.9, there are much fewer red sections than blue, suggesting that there are many more sections that contribute positively to the predicted outcome of the image being real by the classifier. The lack of red indicates that there is not much about the image to suggest that the image is AI-generated, which makes sense as it is an image of a painting crafted by a human artist. In addition, while the other fake images have both dark red and blue segments, suggesting that the model sees segments that both contribute greatly to the image being real or fake, this image's heatmap does not contain any dark red segments, suggesting that there aren't any segments to indicate the image is fake. This makes sense logically, as the image is of a real painting and there shouldn't be any indicators that it is fake.

This suggests that the fake images generated by the diffusion models can contain image segments that can emulate real images enough to trick the model, while other segments the model can accurately predict as fake. This could be one of the reasons a model may classify the overall image as fake–if there is sufficient conflicting evidence

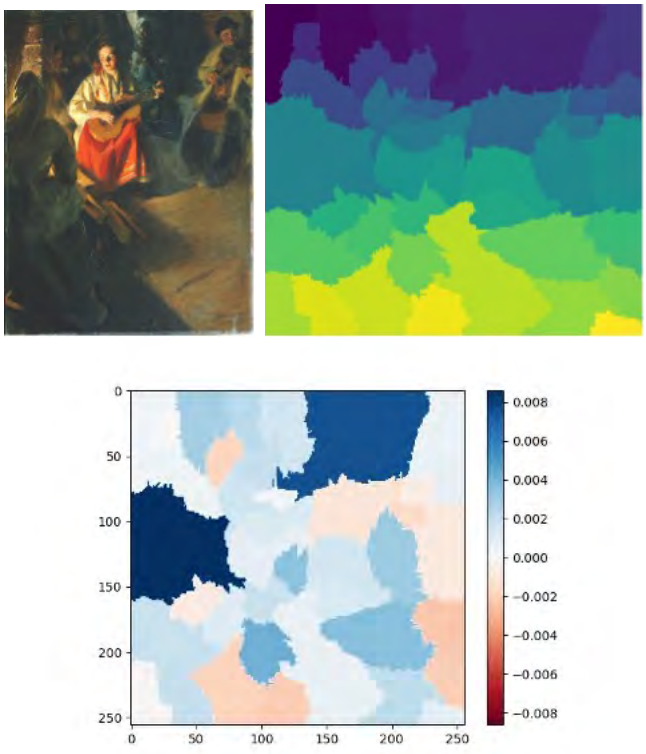

Figure 5.9: **Artbench image (left), LIME image segments (right), LIME XAI heatmap of image(bottom).**

between the two types of segments. On the other hand, real images may be much easier to classify, since they have a wide variety of blue segments and little to no red segments to suggest the image may be fake. This is valuable evidence, as it provides a deeper insight to the inner process of the model that could be interpreted in various ways by humans.

From these examples, the LIME architecture can be incredibly helpful in understanding why the classifier makes its prediction, especially which segments of the image positively and negatively impact the classifier's prediction. Although the classifier/LIME architecture may not pick up on every mistake that the AI generating models make that is noticeable to the human eye, and the image segments defined by LIME do not fully give an exact analysis of what decisions the classifier makes in its overall prediction, the LIME package provides a more comprehensible insight

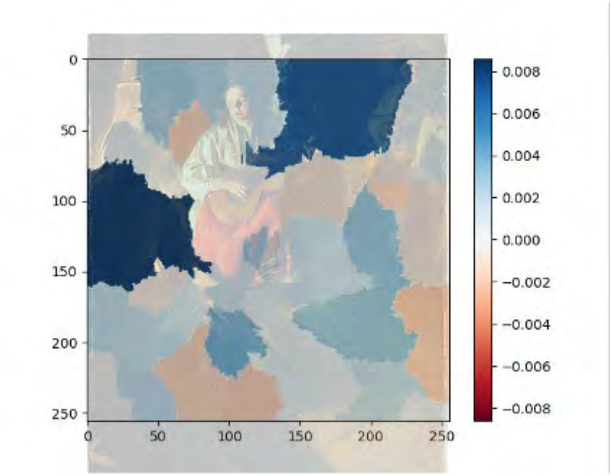

Figure 5.10: **LIME XAI heatmap overlaid with original real Artbench image.**

into how the classifier model makes its predictions, rather than requiring blind trust in the model's decisions. These results also suggest that the images generated by diffusion models may contain some artifacts or imperfections that are imperceptible to the human eye. However, from the examples above, it is clear that the classifier is able to detect on these flaws and accurately classify the images as fake, making this task difficult for humans but much easier for artificial intelligence. This is a testament to how essential Explainable AI architecture can be in AI model analysis.

Chapter 6

Conclusions and Future Work

This paper proposed and carried out the generation of two datasets: one with real artworks from the ArtBench dataset and AI artworks generated by Stable Diffusion, and one with real artworks from the ArtBench dataset and AI artworks generated by DALLE. A classifier was then trained on the datasets, being able to differentiate between the real artworks and the fake, AI-generated artworks successfully with high accuracy, even on datasets the classifier was not trained on. More specifically:

- When trained on the Stable Diffusion dataset, the classifier achieved a 99% accuracy.
- When trained on the DALLE dataset, the classifier achieved a 99% accuracy.
- The classifiers achieved over 75% accuracy when trained on one dataset and tested on the other dataset.
- The classifiers achieved 90% accuracy when trained and tested on subsets of the datasets based on artwork style.
- Even on smaller subsets of training data, the classifiers had reasonable accuracy.

This classifier can thus be used as a detector for fake artwork images on the Internet, which can prevent fraudulent art from circulating or being bought/stolen. Explain-

ableAI architecture in the form of LIME was then used on the model's results to show why the model classified some images as fake and others as real, explaining the model's results in a human way instead of limiting the evaluation to the model's accuracy.

- Real images are likely to have less features detected by the classifier that suggest the image is fake, while fake images tend to contain more of these segments.
- The XAI heatmaps suggest that the diffusion models produce artifacts or hidden flaws in generated images that can be detected by the classifier.

This provided a valuable insight into how XAI can be utilized to provide reasonable human interpretations of a model's results and processes, increasing trust in the model's abilities. It also suggests that the task of detecting artificially-generated artwork is an easier task for a classifier model than it is for humans–there may be be hidden artifacts or mistakes in artificially-generated images that are easily detected by artificial intelligence. Overall, the project was a success in completing its goals.

6.1 Limitations

There were some limitations we encountered throughout the completion of this thesis, especially during the generation of the two datasets. While there were no issues with generating the entire dataset with Stable Diffusion, as there was an open source HuggingFace API pipeline that we were able to utilize, there were many issues with generating the DALLE dataset. The third party API (Eden AI) that was utilized to generate DALLE images in large batches increased their prices to a much larger amount during the year while we were collecting data, which used up all our funding quickly. We were unable to generate every single image to match every artwork in the ArtBench dataset–thus, while the Stable Diffusion dataset contains 99K images,

the DALLE dataset only contains 66K images. With the datasets having different sizes, this could have had a slight effect on the training and testing of the classifier.

- Other limitations included the text-to-image models utilized to generate the datasets–while Stable Diffusion and DALLE-2 were utilized for this study, there are a plethora of other generative models that could have been explored when generating the datasets.

- In addition, rather than using a convolutional neural network to classify the images in the datasets as real or fake, other classification models could have been explored as well.

- Finally, while the XAI architecture LIME was utilized to explain the classifier's predictions, other architectures such as SHAP provide slightly different approaches to providing interpretable explanations for models, which can also be explored.

6.2 Future Work

To extend the results found in this work in the future, it may be a worthwhile endeavor to compare artwork image datasets generated by Generative Adversarial Networks to those generated by diffusion models, or even experiment with other diffusion models not mentioned in this study such as Midjourney. It would also be pertinent to explore other avenues of classifying images, such as feature extraction methods. For instance, Gultepe, Conturo, and Makrehchi proposed the use of unsupervised feature learning in the form of K-means UFLK image feature extraction. These features were then used in a spectral clustering algorithm to to group the paintings based on style groups [8].

In the future, we also plan to expand upon the Explainable AI used to evaluate the classifier model. There are a variety of different XAI packages/architecture that

are available for public use, such as SHAP, LIME, etc. These packages also have the ability to highlight areas of images that contribute to the classifier's decision, as seen with LIME in the previous chapter. Some future work for this project could be to compare the image segments that different XAI architectures highlight, which would likely be different between architectures, and analyze any trends for each architecture. For example, SHAP is another XAI architecture that produces global fidelity explanations compared to LIME's local fidelity explanations, which could have an impact on the image segments highlighted by SHAP.

In addition, another experiment that could be interesting would be to analyze any XAI trends for each art style–more specifically, to determine if there are certain parts of an artwork for a particular art style that provide more contributions to the classifier's decision compared to other parts. For example, for AI-generated artworks mimicking Realism paintings depicting humans, how does the classifier differentiate between fake artworks of people versus real artworks of people? Does the classifier focus on facial features, such as the nose, eyes, mouth? Or is it something else entirely?

Overall, the topic of why and how models make their decisions is just as important as collecting model results, albeit being too broad for the scope of this thesis. Exploring XAI and its ability to explain model decisions in image classification warrants future exploration and research.

Appendix A: Code

```
% generate_stable_dataset.py
%%Generating the synthetic images using Stable Diffusion

def generate_stable_image(text_prompt):
    pipe = StableDiffusionPipeline.from_pretrained
    ("CompVis/stable-diffusion-v1-4")
    # add torch_dtype=torch.float16 for gpu
    pipe = pipe.to("cpu") # change to cuda for gpu
    image = pipe(text_prompt).images[0]
    filename = text_prompt.replace(" ", "_") + ".png"
    image.save(filename)

generate_stable_image("oaks in the mountains of carrara
in the style of realism")

% generate_dalle_dataset.py
%%Generating the synthetic images using DALLE

for i in range(10000, len(prompts)): #change
```

```
        if ((i % 500 == 0 and i != 10000) or i == len(prompts) - 1):
            url = "https://api.edenai.run/v2/image/generation/
            batch/final"
            + str(run) + "/"
            print("test", i, url)
            response = requests.post(url, json=payload,
            headers=headers)

            run += 1
            payload = { "requests": [] }

            if (i % 2500 == 0):
                time.sleep(90 * 60) #1.5 hour

        payload["requests"].append({
            "response_as_dict": True,
            "attributes_as_list": False,
            "show_original_response": False,
            "resolution": "1024x1024",
            "num_images": 1,
            "text": prompts[i],
            "providers": "openai"
        })
```

```
% results.py
%%Training and testing the classifier on the datasets
```

```
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Conv2D, MaxPooling2D, Dense,
Flatten

model = Sequential()

model.add(Conv2D(16, (4, 4), 1, activation='relu',
input_shape=(256, 256, 3)))
#(256, 256, 3)
model.add(MaxPooling2D())

model.add(Conv2D(32, (4, 4), 1, activation='relu'))
model.add(MaxPooling2D())

model.add(Conv2D(16, (4, 4), 1, activation='relu'))
model.add(MaxPooling2D())

model.add(Flatten())
model.add(Dense(32, activation='relu'))
model.add(Dense(1, activation='sigmoid'))

model.compile('adam', loss = tf.losses.BinaryCrossentropy(),
metrics=['accuracy'])
...
hist = model.fit(train, epochs=100, validation_data=cv,
callbacks=[tensorboard_callback])
...
```

```
from tensorflow.keras.metrics import Precision, Recall,
BinaryAccuracy

pre = Precision()
rec = Recall()
acc = BinaryAccuracy()

for batch in test.as_numpy_iterator():
    X, y = batch
    yhat = model.predict(X)
    pre.update_state(y, yhat)
    rec.update_state(y, yhat)
    acc.update_state(y, yhat)

print(f'Precision: {pre.result().numpy()},
Recall: {rec.result().numpy()}, Accuracy: {acc.result().numpy()}')

% xai.py
%%Using the Explainable AI LIME architecture on the images
to generate heatmaps

for image_name in dir_:
    image_path = load_image_data_from_path(path, image_name)
    normalized_img = transform_image(image_path, IMG_SIZE)
    model_prediction = model.predict(normalized_img)
    explainer = lime_image.LimeImageExplainer()
    exp = explainer.explain_instance(normalized_img[0],
```

```
                                model.predict,
                                top_labels=5,
                                hide_color=0,
                                num_samples=1000)
plt.figure(plt_num)
plt.imshow(exp.segments)

dict_heatmap = dict(exp.local_exp[exp_class])
heatmap = np.vectorize(dict_heatmap.get)(exp.segments)
plt.imshow(heatmap, cmap = 'RdBu', vmin  = -heatmap.max(),
vmax = heatmap.max())
```

Bibliography

[1] S. S. Baraheem and T. V. Nguyen. Ai vs. ai: Can ai detect ai-generated images? *Journal of Imaging*, 9(10):199, 2023.

[2] J. J. Bird and A. Lotfi. Cifake: Image classification and explainable identification of ai-generated synthetic images. *IEEE Access*, 2024.

[3] A. Borji. Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2. *arXiv preprint arXiv:2210.00586*, 2022.

[4] G. Castellano and G. Vessio. Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview. *Neural Computing and Applications*, 33(19):12263–12282, 2021.

[5] B. Chen et al. Classification of artistic styles of chinese art paintings based on the cnn model. *Computational Intelligence and Neuroscience*, 2022, 2022.

[6] J. Chen, J. An, H. Lyu, and J. Luo. Learning to evaluate the artness of ai-generated images. *arXiv preprint arXiv:2305.04923*, 2023.

[7] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[8] E. Gultepe, T. E. Conturo, and M. Makrehchi. Predicting and grouping digitized paintings by style using unsupervised feature learning. *Journal of cultural heritage*, 31:13–23, 2018.

[9] P. Liao, X. Li, X. Liu, and K. Keutzer. The artbench dataset: Benchmarking generative models with artworks. 2022, 2023.

[10] I. Meyer. Art periods - a detailed look at the art history timeline. *Art in Context*, Dec 2023.

[11] H. T. T. Nguyen, H. Q. Cao, K. V. T. Nguyen, and N. D. K. Pham. Evaluation of explainable artificial intelligence: Shap, lime, and cam. In *Proceedings of the FPT AI Conference*, pages 1–6, 2021.

[12] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. arxiv 2022. *arXiv preprint arXiv:2204.06125*, 2022.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[15] A. Samo and S. Highhouse. Artificial intelligence and art: Identifying the aesthetic judgment factors that distinguish human-and machine-generated artwork. *Psychology of Aesthetics, Creativity, and the Arts*, 2023.

[16] W. R. Tan, C. S. Chan, H. E. Aguirre, and K. Tanaka. Artgan: Artwork synthesis with conditional categorical gans. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3760–3764. IEEE, 2017.

[17] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu. Explainable ai: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*, pages 563–574. Springer, 2019.

[18] M. R. Zafar and N. Khan. Deterministic local interpretable model-agnostic explanations for stable explainability. *Machine Learning and Knowledge Extraction*, 3:525–541, 06 2021.