CATCH GPT: AI ART DETECTION

Harry Leung
29279838
hgleung@uci.edu

Alex Jen
14601673
ajen2@uci.edu

ABSTRACT

With the rise of AI-generated art, it is important to have reliable ways to tell whether the artwork was created by a human or an AI. This project builds a deep learning system to classify images as either human-made or AI-generated, comparing two types of models: Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). Using the ArtBench and GenImage datasets (more than 180,000 images), we trained both models with custom designs. The CNN reached 95.42% accuracy on validation data, while the ViT (DeiT-Tiny) performed even better at 99.84%, showing its strength in capturing image details. Our key contributions include comparing these model types, improving data processing techniques, and exploring the challenges of AI art detection. This work helps build better tools for verifying and curating digital art.

1 PROJECT INTRODUCTION

The rapid advancement of artificial intelligence (AI) in creative fields has led to a surge in AI-generated artwork, produced by models such as Stable Diffusion and DALL-E. These AI systems can generate high-quality images that closely resemble human-created art, often making it difficult to distinguish between the two. This blurring of boundaries raises critical concerns about artistic authenticity, intellectual property rights, and the integrity of art markets and competitions. In addition, museums and archives responsible for historical preservation must now face the challenge of identifying and verifying the origin of digital artwork.

Currently, the primary method for distinguishing AI-generated art is based on manual inspection by experts or art enthusiasts. However, this approach is subjective, time-consuming, and prone to human error. Given the rapid evolution of AI-generated art, manual methods are becoming increasingly inadequate. Therefore, there is a pressing need for automated, scalable, and reliable techniques to classify artwork as either AI-generated or human-created.

To address this challenge, we propose a deep learning-based system for automated art classification. Our approach systematically compares two widely used neural network architectures. Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). CNNs excel at extracting fine-grained local features from images, while ViTs utilize global attention mechanisms to capture broader contextual relationships. Using these two architectures, we aim to determine which model is more effective in distinguishing AI-generated art from human-created art.

Our key contributions in this work are as follows:

- **A Reproducible Classification Pipeline:** We develop a structured and scalable framework for training and evaluating deep learning models on AI art detection.
- **Empirical Comparison of CNNs and ViTs:** We conduct a systematic analysis of CNN-based and ViT-based classifiers to assess their accuracy, efficiency, and generalizability across different AI-generated art styles.
- **Insights into Dataset Requirements:** We analyze the impact of dataset composition, including variations in resolution, style, and diversity, on the performance of deep learning models in AI art detection.

By providing an in-depth comparison of these two model architectures and offering insights into dataset requirements, our work aims to advance automated AI art detection. This research con-

tributes to the broader field of digital art authentication and helps establish foundational tools for identifying AI-generated content in artistic domains.

2 Related Works

Previous research has explored various methods for detecting AI-generated content, focusing on different aspects of the problem. Silva, Lotfi, Ihianle, Shahtahmassebi, and Bird (2024) developed **ArtBrain**, a toolkit designed to analyze and attribute artistic styles. Their approach emphasizes explainability, helping users understand why a piece of art is classified in a certain way. However, their work primarily focuses on stylistic attribution rather than AI-generated detection.

Zhu, Chen, Yan, Huang, Lin, Li, Tu, Hu, Hu, and Wang (2023b) introduced **GenImage**, a dataset containing AI-generated versions of ImageNet images. This dataset is valuable for studying how AI-generated content differs from real-world images and has been used to benchmark various classification models. However, it does not focus specifically on fine art or creative works, leaving room for further research in AI-generated art detection.

Kholy (2024) examined **CNN attention mechanisms** in the context of art classification. His study analyzed how CNN models focus on different regions of an image when making predictions, shedding light on the decision-making process of these models. While his work improves our understanding of CNN behavior, it does not compare CNNs to newer architectures like Vision Transformers (ViTs).

Our research extends these efforts by **directly comparing CNN and ViT architectures** for AI-generated art detection. Unlike previous work, we focus on larger and more diverse datasets, incorporating both fine art and AI-generated images from multiple sources. Our study also explores the advantages of **global attention mechanisms in ViTs** over the localized feature extraction of CNNs, providing deeper insights into which architectures are best suited for this task.

3 Dataset Description and Analysis

In our project, we use two primary datasets to help us distinguish between human-created and AI-generated art. These datasets provide a diverse collection of images, and we process them in a standardized way to ensure our models work effectively. The datasets are referenced in the footnote.[1]

3.1 ArtBench

- **Size and Composition**: ArtBench contains over 180,000 images. Out of these, 60,000 images are created by humans, while 120,000 images are generated by an AI model (Stable Diffusion).
- **Purpose**: This dataset is designed to provide a wide range of artistic styles and sources, making it ideal for training and testing our models to identify the differences between human and AI art.

3.2 GenImage

- **Size and Composition**: GenImage consists of AI-generated images based on the ImageNet dataset. It serves as an additional source of AI-created artworks.
- **Purpose**: GenImage is used to further evaluate our models. By testing on images derived from a well-known dataset like ImageNet, we can assess how well our models generalize to different styles of AI art.

3.3 Preprocessing Steps

Before feeding the images into our deep learning models, we preprocess the data with the following steps to ensure consistency and to improve model performance:

[1] ArtBench: , GenImage:

- **Augmentation:**
  - **Resizing:** All images are resized to 224x224 pixels to maintain a uniform input size.
  - **Random Flips:** We apply random horizontal flips to introduce variability in the images.
  - **Rotation:** Images are randomly rotated by up to $\pm 15°$, which helps the model learn to handle slight changes in orientation.
  - **ColorJitter:** Adjustments in brightness and contrast (both set to 0.2) are applied to simulate different lighting conditions and color variations.
- **Normalization:**
  - Pixel values are scaled to the range [0,1].
  - We normalize the images using a mean of [0.485, 0.456, 0.406] and a standard deviation of [0.229, 0.224, 0.225]. This standard normalization helps the model converge more efficiently during training.
- **Data Splitting:**
  - **Training Set:** 70% of the images are used for training the models.
  - **Validation Set:** 15% of the images are used to validate the model performance during training and fine-tune the hyperparameters.
  - **Test Set:** The remaining 15% are reserved for testing the final model performance.

4 APPROACH

Our project focuses on designing two deep learning models that can automatically distinguish between human-made and AI-generated art. In this section, we explain our approach in detail, from the overall design down to specific components and mathematical formulas used in our models.

4.1 CNN ARCHITECTURE

Our Convolutional Neural Network (CNN) model is designed to extract local features from images. The model is built from several layers that gradually reduce the image dimensions while capturing important patterns. The key steps in our CNN architecture are:

- **Convolutional Blocks:** We use four convolutional blocks. Each block contains:
  - **Conv2D:** This layer uses a 3x3 kernel to scan the image and detect local features.
  - **Batch Normalization:** This layer normalizes the output from the convolution, helping to stabilize and speed up training.
  - **ReLU Activation:** This function introduces non-linearity, allowing the network to learn complex patterns.
  - **Max Pooling:** This layer reduces the spatial dimensions (width and height) by selecting the maximum value in each region, thereby highlighting the most important features.
- **Adaptive Average Pooling:** After the convolutional blocks, we apply adaptive average pooling. This step reduces the dimensions of the feature maps to a fixed size, regardless of the input image size.
- **Flatten and Dropout:** The pooled feature maps are then flattened into a one-dimensional vector. A dropout layer with a probability of 0.5 is applied to prevent overfitting by randomly turning off some neurons during training.
- **Fully Connected (FC) Layer:** The flattened vector is passed through a fully connected layer with 512 units. This layer helps to combine features from different parts of the image.
- **Output Activation:** Finally, a Sigmoid activation function produces a binary output, indicating whether the image is AI-generated or human-created.

- **Loss Function:** We train the CNN using the binary cross-entropy loss function:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

where y_i is the true label and $\hat{y}_i$ is the predicted probability for image i.

4.2 Vision Transformer (ViT) Architecture

Our second approach uses a Vision Transformer (ViT), which is known for its ability to capture global relationships in an image using self-attention mechanisms. The ViT we employ is based on a pretrained DeiT-Tiny model from Hugging Face. The main components are:

- **Input Processing:** The image is divided into fixed-size patches of 16x16 pixels. Each patch is then flattened into a vector.
- **Positional Embeddings:** To retain the spatial relationship between patches, we add positional embeddings to each patch vector.
- **Transformer Encoder:** The patched inputs with positional embeddings are fed into a Transformer encoder that consists of 12 layers with 3 attention heads per layer. Within these layers, the self-attention mechanism computes:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Here, Q, K, and V represent the query, key, and value matrices, and d_k is the dimension of the key vectors.

- **Classification Head:** After the Transformer encoder, a Layer Normalization is applied, followed by a fully connected layer that outputs a binary prediction.

4.3 Key Components and Benefits

The design of our models incorporates several important components that contribute to improved performance:

- **Local Feature Extraction in CNNs:** The CNN's multiple convolutional layers help in identifying fine details and textures that distinguish human art from AI-generated art.
- **Global Context in ViTs:** The Vision Transformer captures long-range dependencies across the image, providing a broader context that is particularly useful for recognizing overall styles and structures.
- **Adaptive Preprocessing:** Both models benefit from standardized preprocessing steps, such as image resizing, normalization, and data augmentation, which help in generalizing across diverse art styles.
- **Robust Training Techniques:** The use of dropout in the CNN and pretrained weights in the ViT helps in reducing overfitting and accelerating convergence during training.

5 Evaluation Results and Quantitative Analysis

We conducted our approach using PyTorch, among other libraries essential for machine learning using Python, and wrote a custom art classifier class easily interchangeable between CNN and ViT. We used a 4-layer CNN and facebook/deit-tiny-patch16-224, a compact data-efficient image transformer pre-trained on ImageNet-1k and optimized for time and space constraints.

5.1 Evaluation Metric

As the goal is to classify AI images from real art, we measure the performance of our models based on accuracy, or the percentage of correct predictions. To do so, we have established a confusion

matrix per (Kholy, 2024) to establish a benchmark we can use to evaluate our models. There are four categories a classification can belong to: True Positive, True Negative, False Positive, and False Negative, with the positive referring to AI generation and negative referring to human work.

Classification	Negative	Positive
True Negative		
True Positive		

In addition to the confusion matrix, we can also evaluate our model's performance by calculating the receiver-operating characteristic area under the curve, or the ROC AUC score, derived from the true positive rate compared to the false positive rate. In the context of our models, the ROC AUC score can be interpreted as the percentage of correctness of the model. The ROC AUC is set on a scale of $\{0,1\}$, with 1 meaning a guaranteed probability of correctness, and 0.5, 50% correct and 50% incorrect, meaning that the model's predictions are no better than random guesses.

5.2 CNN Results

We trained the CNN on a subset of the GenImage dataset, specifically 18,000 images split between real images from ImageNet and their AI-generated equivalents using Stable Diffusion 1.5. We trained twice on this subset; the first run with a batch size of 32, a learning rate of 0.001, and 50 epochs. The second run retains the same parameters with the exception of a learning rate of 0.0001.

Learning Rate	Training Accuracy	Validation Accuracy	Training Loss	Validation Loss
0.001	98.08	95.43	0.0016	0.0056
0.0001	98.41	93.5	0.0013	0.0063

The results revealed excellent accuracy in both facets, which often appears to be a common sign of overfitting. It is possible that the subset of the images has proven too small and the resulting small sample size is insufficient to generalize outside of the training data.

Plotting the accuracy and loss over the 50 epochs also reveals a possible symptom of overfitting.

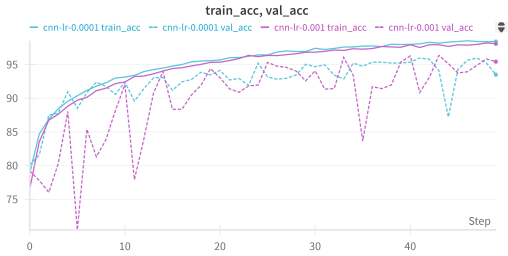

Figure 1: Training and Validation Accuracy

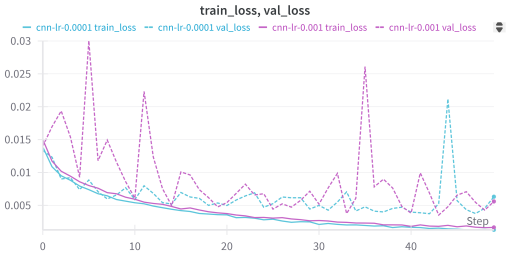

Figure 2: Training and Validation Loss

The spikiness of the graphs may be attributed to a number of factors, including an insufficient sample size and insufficient regularization. A potential theory for the loss is the fact that Adam is a stochastic optimizer, meaning that with each epoch the loss is not guaranteed to decrease, though over time the model does see improvement.

5.3 ViT (DeiT) Results

Due to compute constraints, we opted to apply a ImageNet pre-trained DeiT for its compact space and efficiency. We again train it on an 18,000 image subset of the GenImage dataset over 50 epochs with a batch size of 32 and a learning rate of 0.00003.

Learning Rate	Training Accuracy	Validation Accuracy	Training Loss	Validation Loss
0.00003	99.92	99.84	0.0001	0.0002

Similar to the results of the CNN, the coupling of extremely high accuracy with little to no loss indicates a high likelihood of overfitting. ViTs require more information than CNNs due to their inability to maintain an inductive bias, or a hidden state in the network capable of holding important information. Thus, with the overfitting of the CNN possibly owing to the small dataset available, it is likely that the DeiT is affected by the same issue.

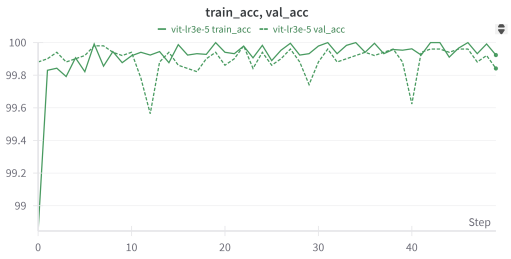

Figure 3: Training and Validation Accuracy

Figure 4: Training and Validation Loss

5.4 Confusion Matrix

Calculating the confusion matrix and ROC AUC for the CNN confirms our belief in overfitting. Clearly the model is trained too closely on the training data, resulting in an almost inability to classify properly. The ROC AUC is below 0.5, meaning it is as accurate, if not less accurate than, as guessing.

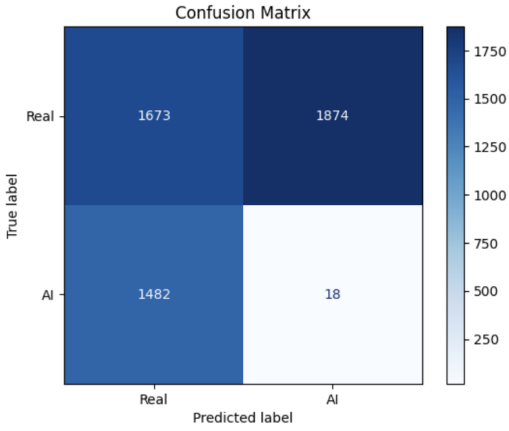

Figure 5: CNN Confusion Matrix

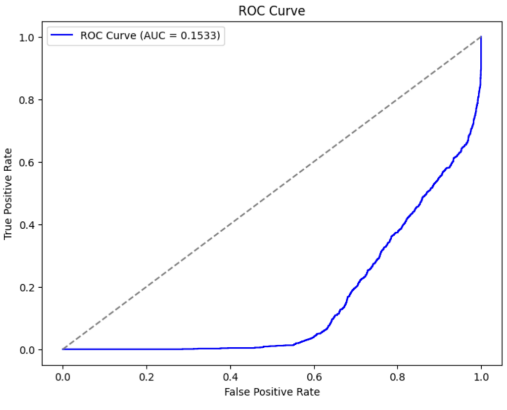

Figure 6: ROC AUC

6 Discussion and Conclusion

6.1 Improvements

The current results are promising but show room for improvement. With stronger computing power, the models could be scaled up to handle larger batch sizes and deeper architectures, making full use of all 180,000 images in the ArtBench dataset. This would help reduce overfitting by better capturing data patterns.

Further improvements include fine-tuning hyperparameters, such as finding the best learning rate for CNN and DeiT using grid search or Bayesian optimization. Adding regularization techniques like dropout and weight decay could also help, along with early stopping to prevent unnecessary training.

Exploring ensemble methods—combining different models—could improve performance, while data augmentation (e.g., geometric transformations, color jittering, adversarial training) would make the model more robust to input variations. These refinements could smooth fluctuations in loss and accuracy while boosting overall predictive accuracy.

6.2 Future Works

Further time can be spent evaluating these models in comparison to the custom models devised by the authors behind the datasets we are interested. For example, Silva, Lotfi, Ihianle, Shahtahmassebi, and Bird (2024) explore the use of AttentionConvNeXT, a proprietary model intended to specialize in the classification of AI art and "get the low, mid and high-level feature maps effectively utilised for the final prediction" (Silva et al., 2024). Essentially, theirs is an ensemble model with a CNN at its core and aided by ConvNeXT, a ViT-inspired model that claims to replicate the superior efficiency of transformers.

In addition, Zhu, Chen, Huang, Li, Hu, Hu, and Wang (2023a) test the capability of teacher-student discrepancy-aware learning and generalized feature augmentation. This method emphasizes the minute details that give away AI art from real art to the human eye by creating an adversarial network similar to GANs. By focusing attention on any "small discrepancy" in the image combined with generators aimed to fool the teacher and student, this system becomes optimized toward identifying the nuances of AI art.

Both the ArtBench and GenImage datasets are not only massive in their library of images but incredibly diverse. Our study only scratches the surface of the different varieties of imagery, real or AI-generated. It is possible that these models may perform differently when trained or tested on specific art movements, as ArtBench categorizes its images by eras such as Baroque and Romanticism. Furthermore, GenImage boasts a wide variety of images categorized by the AI art model. We used Stable Diffusion 1.5 here, but it is worth identifying the key differences when our models are compared to other models like Midjourney, ADM, GLIDE, Wukong, VQDM, BigGAN, and more.

6.3 Conclusion

As Professor Baldi stated, we are approaching one of AI's greatest milestones—perhaps its last—before reaching general intelligence. While AGI is still a distant concept, AI's rapid progress is already transforming industries beyond computer science, raising ethical concerns, particularly in AI-generated art. AI has won art competitions, sparking debates about the future of creative careers and the impact on various other professions. Our project presentations highlight the vast reach of AI, emphasizing the need for responsible oversight. As AI advances, so must our efforts to regulate and guide it, ensuring it remains a tool for progress rather than disruption. Thus it becomes exponentially important every year that artificial intelligence improves that we also improve the means to corral it.

ACKNOWLEDGMENTS

Harry Leung: My contribution to this project was centered on writing most of the model code. I was responsible for designing and implementing the deep learning architectures, including both the Convolutional Neural Network (CNN) and the Vision Transformer (ViT). I developed the code for key components such as the convolutional layers, pooling, dropout, and the classification head for the CNN, as well as integrating the pretrained DeiT-Tiny model for the ViT. I also implemented the training routines, loss functions, and evaluation metrics to monitor model performance and improve accuracy.

Alex Jen: I conducted the research on the current state of AI art, its applications, and the existing classifiers that form the basis of this project. I also finetuned and trained the models created by Harry on the datasets mentioned.

REFERENCES

Adam El Kholy. Ai-artwork, 2024. URL https://www.kaggle.com/dsv/7878124.

Ravidu Suien Rammuni Silva, Ahmad Lotfi, Isibor Kennedy Ihianle, Golnaz Shahtahmassebi, and Jordan J. Bird. Artbrain: An explainable end-to-end toolkit for classification and attribution of ai-generated art and style, 2024. URL https://arxiv.org/abs/2412.01512.

Mingjian Zhu, Hanting Chen, Mouxiao Huang, Wei Li, Hailin Hu, Jie Hu, and Yunhe Wang. Gendet: Towards good generalizations for ai-generated image detection, 2023a. URL https://arxiv.org/abs/2312.08880.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image, 2023b. URL https://arxiv.org/abs/2306.08571.