Proposal on AI Image Detector

LAM, Tsz Kit
SID: 1155194085
The Chinese University of Hong Kong
1155194085@link.cuhk.edu.hk

CHOI Ho Yan
SID: 1155194468
The Chinese University of Hong Kong
1155194468@link.cuhk.edu.hk

Abstract

This proposal outlines a project plan for an AI image detector. We introduce background motivation and recent research related to image classification. We explain approaches, implementation and expected conclusions of proposed solution.

1 Background

Recently, Ghibli-style AI images have aroused heated discussion about the use of AI in artwork. Ghibli, a Japanese animation studio, is well-known for its unique anime style behind Spirited Away, My Neighbor Totoro, and other classical films. However, the artistic style has been transferred to the image generator in the latest GPT-4o model from OpenAI, with 11% and 5% increases in the global ChatGPT app downloads and weekly active users [1]. While people are amazed by the realistic imitations, artists, including Ghibli's legendary director Hayao Miyazaki, have criticized its contempt for art [1].

The over-success of the generative models has raised not only artwork copyright infringements but also information security challenges. Due to the high similarity and indistinguishability between AI-generated and human-designed images, there are potential threats of information misuse and misleading content, especially considering an average of 2 million images are generated daily [2]. As people are more sensitive to images when receiving information, this further highlights the influences of AI images. For example, misleading information and photos in political campaigns or celebrity reputations could frame public opinion. The research found that over 80% of Americans expressed concerns regarding the misuse of AI in the U.S. presidential election [3]. The technology can also be manipulated with deepfake technologies to achieve identity theft.

To mitigate the underlying concerns about AI images, this project proposes a classification model that classifies between AI-generated and human-designed images. We hope that the AI image detector can be used to effectively recognize and highlight the use of AI images on media platforms, so that people would be cautious about the source of images.

2 Problem Statements

The main challenge in distinguishing AI-generated images from human-crafted designs lies in the increasingly sophisticated visuals produced by current generative models, which often blur the line between manmade artwork and automated outputs. Previous studies on AI image detection have primarily focused on analyzing pixel-level artifacts, watermark patterns, or statistical anomalies. Then developing a foundation classifier accordingly, these solutions usually employ traditional CNN-based structures like ResNet-50. Although these techniques often achieve moderate accuracy on standard datasets, providing a foundation for broader misuse detection frameworks that suggest the lower bound accuracy [4] [5], they might struggle with newer generator architectures that employ sophisticated diffusion or transformer-based strategies [5].

Our proposed solution is to conduct literature reviews on different classification models, and build a particular classifier for AI image detector to distinguish AI-human image imitations.

3 Approaches

To develop an AI-human image classifier, we will adopt the following approaches:

1. Data Collection: We first prepare two datasets of images, labelled "AI-generated" and "human-crafted" respectively, which serve as ground-truth data for supervised learning. Each dataset is divided into training data and testing data sets for training and evaluating the model.
2. Data Processing: We normalize the images by resizing and scaling them to a consistent size, which helps reduce variations in image formats. We could also augment the datasets by flipping, rotating, and adding noise to images to improve model robustness.
3. Model Selection: After conducting literature reviews on different model advantages, we will try and select a suitable model architecture from a range of popular image-processing models.
4. Model Training: We write Python programs to train the selected model by adjusting hyperparameters such as the number of layers and neurons, the learning rate, and possibly L2 regularization to achieve better model performance.
5. Model Evaluation: The model performance is measured by standard metrics including accuracy, precision, recall, and F1 score on the test dataset. Generally, we hope to achieve an overall accuracy of 95% or above.
6. Application Presentation: We would build a simple web application that allows users to upload an image and receive classification output information.

4 Implementation and Experimentation

Following the approaches we proposed, we explain the initial implementation details of the AI-human image classifier.

4.1 Data Manipulation

To ensure accurate training of deep learning models, we expect to collect any image data with high diversity to reduce classification bias. There will be around 25,000 training data and 5,000 testing data for each "AI-generated image" and "human-generated image" class. Each image should be processed and outputted with a consistent resolution of 224x224 pixels. We will primarily use the dataset available on Kaggle [6].

4.2 Model Implementation

We will write Python programs to build and compare two or more classification models with popular architectures, including the Convolutional Neural Network (CNN) and Vision Transformer (ViT).

CNNs utilize convolutional layers, which allow extracting edge and texture features from images and learn these spatial features by layers; ViTs utilize a self-attention mechanism, which allows them to capture long-range dependencies and global features effectively and differentiate authentic content from synthetic imagery [4][7].

We would try different hyperparameters on training these models and evaluate their accuracy, precision, recall, and F1 score using test dataset.

4.3 Deployment Implementation

After validating the model's performance, we would build a web application using HTML, CSS, NodeJS for front-end web server, and Python for back-end application server. The webpage will classify the uploaded images using the trained model, and display the classification class result. Both servers will be deployed and hosted on Microsoft Azure.

5 Expected Conclusions

1. **Accuracy and Robustness**: The classification model is expected to achieve high accuracy (95% or higher) in distinguishing AI-generated images from human-crafted designs, especially when trained on a diverse dataset.
2. **Generalizability**: The model should demonstrate good coverage across different styles of AI-generated images, including those produced by newer architectures.

References

[1] "South China Morning Post," *South China Morning Post*, Apr. 02, 2025. https://www.scmp.com/tech/big-tech/article/3304828/chatgpt-usage-hits-record-studio-ghibli-style-ai-images-go-viral?module=perpetual_scroll_0&pgtype=article (accessed Apr. 04, 2025).

[2] N. Tiku, "AI can now create any image in seconds, bringing wonder and danger," *Washington Post*, Sep. 28, 2022. https://www.washingtonpost.com/technology/interactive/2022/artificial-intelligence-images-dall-e/

[3] H. Y. Yan, G. Morrow, K.-C. Yang, and J. Wihbey, "The origin of public concerns over AI supercharging misinformation in the 2024 U.S. presidential election," *The origin of public concerns over AI supercharging misinformation in the 2024 U.S. presidential election*, Jan. 2025, doi: https://doi.org/10.37016/mr-2020-171.

[4] C. Tan, Y. Zhao, S. Wei, G. Gu, and Y. Wei, "Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection." Available: https://openaccess.thecvf.com/content/CVPR2023/papers/Tan_Learning_on_Gradients_Generalized_Artifacts_Representation_for_GAN-Generated_Images_Detection_CVPR_2023_paper.pdf

[5] "SKDU at De-Factify 4.0: Vision Transformer with Data Augmentation for AI-Generated Image Detection," *Arxiv.org*, 2025. https://arxiv.org/html/2503.18812v1 (accessed Apr. 05, 2025).

[6] "Search | Kaggle," *Kaggle.com*, 2025. https://www.kaggle.com/search?q=ai+image+in%3Adatasets (accessed Apr. 05, 2025).

[7] Nouar AlDahoul and Y. Zaki, "Detecting AI-Generated Images Using Vision Transformers: A Robust Approach for Safeguarding Visual Media Integrity," Jan. 2025, doi: https://doi.org/10.2139/ssrn.5029893.