AIRA GENERALIZED DETECTION OF AI-GENERATED IMAGES LEVERAGING EXPLAINABLE AI.

Devarn Paraneetharan

A dissertation submitted in partial fulfilment of the requirement for

Bachelor of Engineering (Honours) degree in Software Engineering

School of Computing

Informatics Institute of Technology, Sri Lanka

in collaboration with

University of Westminster, UK

2025

ABSRACT

Artificially generated images have rapidly risen in both quality and quantity, largely driven by their popularity across various domains due to their relative ease of access and use. This poses threats of misinformation and lack of trust in completely artificially generated images that pass along as real, particularly in news and social media. Current research focuses on the development of novel methods to enhance detection, improving generalization, or on the addition of interpretability factors using explainable artificial intelligence (XAI). This project addresses the need for a detection system that allows for efficient prediction of "AI-generated" images across a variety of image generators and of various content types over "Real" images, whilst providing insights and explanations for the prediction.

In this research, a novel approach is proposed for the detection of AI-generated images while integrating Explainable Artificial Intelligence techniques to facilitate interpretability. The proposed solution implementation was developed using a Convolutional Neural Network (CNN) classifier built on top of the DenseNet121 architecture, which leveraged pre-trained weights from ImageNet, allowing for efficient feature extraction. The model was trained on a curated dataset consisting of images from four AI-image generators, which included Midjourney, Stable diffusion, Dall-E, ProGAN, and other images from State-of-The-Art image generators for the classification of AI-generated imagery.

The proposed solution was evaluated using a range of data science metrics, including the F1 score, precision, accuracy and recall on an unseen portion of the dataset. The initial results prove promising, achieving on average and accuracy of 93% while maintaining an F1-score of 93% and loss of 0.22.

Keywords: AI-generated image detection, Image recognition, Explainable AI, Synthetic image detection, Deep learning.

Subject Descriptors: