Capstone Project Proposal:

Detecting AI-Generated Images

Problem

Given the recent developments in the creation of images using AI, detecting AI-generated content has become increasingly difficult for the average human. Humans failing to detect fake images can result in ethical, social, and political risks. Anywhere from the spread of misinformation and fake news to faking legal evidence can sway the public's opinion in a negative way.

Objective

The objective of this project is to build a deep learning model that can detect whether an image is real or AI-generated.

Dataset

Source: The dataset I will be using to train and test the model on is the CIFAKE dataset on Kaggle: https://www.kaggle.com/datasets/birdy654/cifake-real-and-ai-generated-synthetic-images

Distribution: This dataset contains 120,000 images, half of which have been collected from CIFAR-10, and are split as follows:

1. Training
    a. 50,000 real images from CIFAR-10
    b. 50,000 AI-generated images using Stable Diffusion on CIFAR-10 images
2. Testing
    a. 10,000 real images from CIFAR-10
    b. 10,000 AI-generated images using Stable Diffusion on CIFAR-10 images

Image Size: Same as CIFAR-10, images are 32x32 pixels.

Class Labels: REAL and FAKE

Model Architecture

Base Model: The base model that will be used is the ResNet-18, fine-tuned such that the final layer is a single neuron (i.e., Binary Classification) to classify images as real or fake. This model is pretrained on ImageNet, so it has learned a wide variety of image features (e.g., edges, textures, and objects) that will be useful to identify fake images.

Loss Function: Binary Cross Entropy (will use BCEWithLogitsLoss instead of BCELoss to avoid having to apply sigmoid before passing to loss function).

Optimizer: will attempt the problem with SGD and Adam with different learning rates and use the one that yields best results.

Performance Evaluation

Performance of the model will be evaluated using, but not limited to, the following metrics:

- Training/Validation/Test Loss (incl. plots)
- Training/Validation/Test Accuracy (incl. plots)
- Confusion Matrix Heatmap
- Precision/Recall
- F1 score

Interpretation of Model

I will use explainable AI tools learnt in the classroom, such as CAM, to understand what the model focuses on to identify fake images.

Additional Sources

The following is an article of the study for which the CIFAKE dataset was created. This article can be used as a guide or benchmark to compare my model results to.

https://www.researchgate.net/publication/377538637_CIFAKE_Image_Classification_and_Explainable_Identification_of_AI-Generated_Synthetic_Images