DeepFake Detection

Abhinav Madabhushi, Pierce Ohlmeyer-Dawson

1 Objective

1.1 Application

The goal of this project is to predict whether a given image is AI-generated or Real. This has significant applications, especially in the advancement of AI. 0.1 percent of all images on the internet are currently AI-generated images, and there is current research suggesting that repeated use of these AI-generated images (synthetic data) can lead to unrecoverable model collapse in Generative Models (e.g., ChatGPT). To prevent the collapse of such Generative Models, it is important to classify what images are Real and what images are AI-generated. This task is getting increasingly complex as AI-generated images become closer to Real data with advancing AI models.

1.2 Dataset

The dataset used in this project is a Kaggle dataset containing 60,000 real images from the CIFAR-10 dataset and 60,000 AI-generated synthetic images generated by Stable Diffusion version 1.4, a text-to-image generation model. The images are equivalent to those in the CIFAR-10 dataset, making them difficult to distinguish as Real or Synthetic. The dataset is divided into 100,000 images for training and 20,000 images for testing.

2 Model

The plan for the model is as follows: We will create a baseline binary classification CNN model with convolution layers, batch normalization layers, a ReLU layer, a pooling layer, a fully connected layer, and a final softmax layer for binary output. Additionally, we will experiment with different pre-trained models like ResNet-18, ResNet-50, and EfficientNet-B0 to achieve better results.

3 Project Steps

The project will follow these steps: First, we will load and preprocess the dataset to meet the model requirements. Next, we will build a baseline CNN model using PyTorch or Keras (TensorFlow). After building the model, we will train

and validate the baseline model. Once the model is trained, we will evaluate its performance on test data using accuracy, precision, recall, and F1-score. To improve model robustness and generalization, we will tune the hyperparameters. Following this, we will experiment with pre-trained models (e.g., ResNet, EfficientNet) for enhanced detection capability. We will then compare the performances of these models using statistical tests, ROC curves, and AUC to assess classification effectiveness. Finally, we will compile the results and analyses into the final report.

4 Project Distribution

Abhinav will oversee data preprocessing, which involves loading the dataset, applying data augmentation, and analyzing class distributions. He will also fine-tune pre-trained models such as ResNet and EfficientNet to enhance performance through transfer learning. Pierce will be responsible for building and training the baseline CNN model, managing hyperparameter tuning, and evaluating performance using metrics like accuracy and F1-score. Both will collaborate on testing and result analysis, including confusion matrices, ROC curves, and statistical tests, and will equally contribute to the report.

5 References

[1] J. J. Bird and A. Lotfi, "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images," *IEEE ACCESS*, vol. 12, pp. 15642–15650, 2024, doi: 10.1109/ACCESS.2024.3356122.

[2] C. M. Bishop, *Pattern recognition and machine learning / Christopher M. Bishop.* in Information science and statistics. New York: Springer, 2006.

[3] J. Fang, L. Zheng, C. Liu, and C. Su, "A Data-Driven Case Generation Model for Transient Stability Assessment Using Generative Adversarial Networks," *TII*, vol. 20, no. 12, pp. 14391–14400, 2024, doi: 10.1109/TII.2024.3452211.

[4] I. Goodfellow *et al.*, "Generative adversarial networks," *ACM CACM*, vol. 63, no. 11, pp. 139–144, 2020, doi: 10.1145/3422622.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep Residual Learning for Image Recognition," in *CVPR*, IEEE, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[6] P. Korshunov and S. Marcel, "Deepfakes: a new threat to face recognition? assessment and detection," *arXiv (Cornell University)*, 2018, doi: 10.48550/arXiv.1812.08685.

[7] Krizhevsky A, "Learning Multiple Layers of Features from Tiny Images," *Master's thesis, University of Tront*, 2009.

[8] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv (Cornell University)*, 2015, doi: 10.48550/ARXIV.1511.08458.

[9] A. Torralba, P. Isola, and Freeman William T., *Foundations of computer vision / Antonio Torralba, Phillip Isola, and William T. Freeman.* in Adaptive computation and machine learning series. Cambridge, Massachusetts: The MIT Press, 2024.