Milestone 3 - Deepfake Facial Imagery Detection

Alex Zhang, Apple Jin, Alex Xiang

1. Methodology

In this project, we employed the MobileNet V1 network initialized with ImageNet-pretrained weights, followed by retraining with various hyperparameters or on different datasets. Our baseline MobileNet model was initially trained on a dataset comprising GAN-generated images. While the model achieved satisfactory performance on its test set, with only minor overfitting, it demonstrated poor generalization on image datasets produced by other synthetic engines. For instance, when evaluated on a dataset consisting of Stable Diffusion-generated images, the model showed low accuracy, yielding a score of just over 0.5.

To tackle this generalizability issue, we proposed two potential solutions: First, we could expand the diversity of the training dataset by incorporating images generated by different models. Second, inspired by a relevant paper[1], we implemented a data preprocessing technique called Error Level Analysis (ELA), which detects manipulated regions or artifacts by revealing inconsistencies in compression artifacts between real and fake images.

For the first solution, we augmented the training dataset with an additional 9,000 fake images generated using three versions of Stable Diffusion, along with 6,000 GAN-generated images. These were then fed into the pretrained MobileNet model. We unfroze the last 10 layers and trained two variations: one with no dropout and another with a dropout rate of 0.5. The models were trained using the Adam optimizer with a learning rate 0.0001.

For the second solution, we applied ELA preprocessing to a dataset consisting solely of GAN-generated images. In this configuration, the model was retrained with the last 20 layers unfrozen, using the Adam optimizer at a learning rate of 0.0001.

Finally, we combined both techniques by training the model on a dataset processed with ELA that included GAN- and Stable Diffusion-generated images. The model was retrained with the last 20 layers unfrozen, using the Adam optimizer at a learning rate of 0.0001.

2.Results:

In this project, we evaluated the detection accuracy of four distinct models against datasets of AI-generated images. The models differ in their training configurations:

1. trained exclusively on GANs
2. trained on GANs with Error Level Analysis
3. trained on both GANs and Stable Diffusion (SD) and
4. trained on both GANs and SD with ELA.

The results, delineated in two tables, illustrate the models' performance across varied datasets. Table 1, assessing equal distributions of real and fake images (50:50 ratio), revealed that the model trained solely on GANs exhibited the highest accuracy on GAN-generated images (0.941). However, incorporating ELA or additional training on SD reduced accuracy. Conversely, the model trained on both GANs and SD with ELA achieved perfect accuracy (1.000) in detecting SD-generated images. Table 2 focuses on datasets composed entirely of SD-generated fake images. Here, models solely trained on GANs performed poorly on all SD dataset. In contrast, models trained on both GANs and SD, with or

[1] Martin-Rodriguez, F., Garcia-Mojon, R., & Fernandez-Barciela, M. (2023). Detection of AI-created images using pixel-wise feature extraction and convolutional neural networks. *Sensors*, *23*(22), 9037.

without ELA, flawlessly identified all SD versions (1.000 accuracy). These outcomes highlight the importance of diverse training regimes for effective AI-generated image detection.

3.Analysis

To improve the performance of deepfake detection models, employing a diverse dataset can be a preferable approach to using ELA as a preprocessing tool. This recommendation is substantiated by the observation that the incorporation of ELA did not lead to an increase in accuracy in our tests.

The efficacy of a diverse dataset is clearly demonstrated by our experimental results. First, the model trained only on GAN-generated images showed effective performance only on other GAN-generated images, showing a possible dependence on specific datasets. This limited applicability

	Trained on solely GAN	Trained on GAN with ELA	Trained on GAN & Stable Diffusion	Trained on GAN & SD with ELA
GANs	0.941	0.834	0.860	0.793
Stable Diffusion	0.516	0.50	0.98	1.000

Table 1. Test Accuracy on Different Model

	Trained on solely GAN	Trained on GAN with ELA	Trained on GAN & Stable Diffusion	Trained on GAN & SD with ELA
SD1.5	0.015	0.000	1.000	1.000
SD2.1	0.015	0.000	1.000	1.000
SD XL1.0	0.030	0.000	1.000	1.000

Table 2. Test Accuracy on Different Model

suggests that the model was potentially overfitting to the characteristics unique to GAN outputs. However, substantial improvements were observed after we expanded the datasets to include images generated by Stable Diffusion models. The model trained using this diversified dataset can be used on Stable-Diffusion images that are significantly deviate from those in the training set (some examples are shown below), showcasing its robustness and adaptability to varied types of synthetic imagery.

To understand the inefficacy of detecting deepfake while using models with GAN-exclusive datasets and ELA, we also applied Grad-Cam's heatmap to diagnostics. We deduced some hypotheses on how our model decides to differentiate fake images from real images.

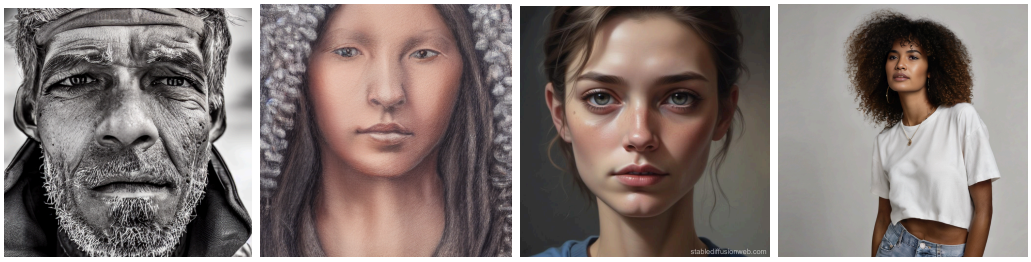

Figure 1. Different Stable Diffusion generated fake images that our model is able to detect

After incorporating Stable Diffusion images into the training dataset, the Grad-CAM heatmaps reveal distinct edges and lines when analyzing the GAN-generated test data. In contrast, for the Stable Diffusion-generated data, the model's decision-making regions on the heatmap are often blurred, as presented in the left two images in Figure 2. This could be due to the fact that the model might be more

familiar with the types of artifacts and patterns generated by GANs due to their prevalence in the training dataset. Thus, it recognizes these features more confidently, highlighting distinct edges and lines. However, recognizing edges and lines seems to make minor improvements in the model accuracy on the GAN images (Accuracy of 0.86 on GAN images and accuracy 0.98 on Stable Diffusion images).

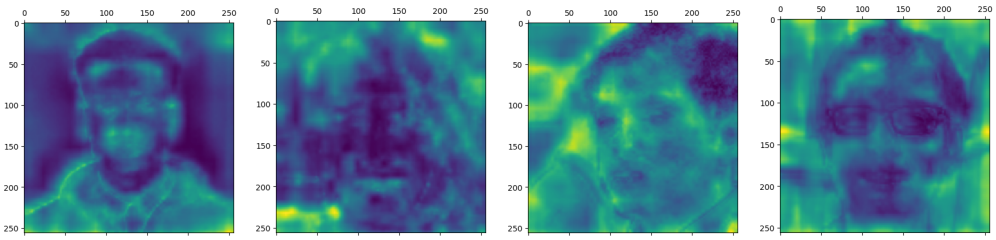

Figure 2. Grad-Cam heatmap on sample images generated by GAN (left first) and by Stable Diffusion (left second); False positive (false real) image (right second) and false negative (false fake) image (right first)

In falsely detected images, whether real or fake, the Grad-CAM heatmaps show higher gradients (appearing greener) in the background portion (non-human-face regions). This suggests that the model relies heavily on the image background for decision-making, indicating its high importance in predictions, as shown in the right two images in Figure 2. This observation implies that the background confuses the model during inference and overshadows the regions that should be the primary focus, such as the human face. However, it is worth noting that the background may reveal unnatural elements characteristic of generated images, so emphasizing the background is not inherently incorrect.

4.Conclusions

Our project was designed with the aim of developing a sophisticated classifier capable of discerning genuine images from those fabricated by AI technologies, irrespective of the origin of the dataset. Throughout the development, we opted to utilize MobileNet due to its efficiency and suitability for deployment in mobile devices, achieving significant improvements in the accuracy of detecting AI-generated images. Notably, our model reached an accuracy of 100% on images from Stable Diffusion and 85% on those generated by GANs, supplemented by successful validation tests on images sourced externally from online AI-image generators. This suggests that our model possesses a commendable degree of robustness. However, the classifier's dependency on the characteristics of the training data may remain a limitation, since its ability to generalize across truly novel or diverse datasets has not been fully tested.

A pivotal strength of our project is the integration of Grad-CAM, which significantly bolstered the interpretability of our model. This was particularly effective when applied to images that had undergone Error Level Analysis, which tends to produce predominantly darkened images. Grad-CAM provided us with deeper insights into the decision-making processes of our neural network, identifying specific areas within the images that were pivotal in determining their authenticity. This not only helped in understanding why certain predictions were made but also contributed immensely to the trustworthiness and transparency of the model's outputs.

The future trajectory of our project involves several promising directions. First, there is a clear need to expand our dataset to include a broader array of images from emerging AI technologies like MidJourney and DALL-E. This expansion is expected to enhance the model's robustness and adaptability significantly. Additionally, refining the classification thresholds to suit specific application needs will help

optimize the model's performance across various contexts, acknowledging that the default threshold of 0.5 may not be ideal in all scenarios. Finally, utilizing insights from Grad-CAM to identify frequently highlighted regions in fake images. We suggest developing targeted classifiers to focus on these specific image attributes, which enhance robustness of our model to distinguish diverse AI-generated content.

5.Workflow:

Milestone 1: Setup and Data Preparation

The project began with defining its scope and gathering a diverse dataset. The team set up initial models like DenseNet-121 and MobileNet, and established a training and validation framework.

Milestone 2: Model Training and Evaluation

Training commenced with ongoing adjustments to hyperparameters. The team also conducted initial evaluations to measure the classifier's performance metrics.

Milestone 3: Model Refinement and Final Evaluations

The final milestone involved refining the models. This included further tuning of hyperparameters and adjustments to the model architecture. An extended qualitative evaluation was conducted to include more images and validate the effectiveness of highlighted areas. The final week was dedicated to a comprehensive evaluation of the models, focusing both on quantitative metrics and qualitative feedback. Detailed documentation of the project was compiled, and a presentation was prepared to showcase the project's objectives, methodology, findings, and implications.

At the conclusion of Milestone 3, all team members contributed equally to the project's success, each bringing their unique skills and dedication to the table. Apple Jin invested considerable time in testing and fine-tuning different models, ensuring optimal performance. Alex Xiang was pivotal in implementing advanced techniques such as ELA and Grad-CAM, enhancing the model's interpretability. Alex Zhang led the direction of the research, meticulously analyzing data and preparing presentations.