Rahul Khichar
23110264

Technical Report: Classifying AI-Generated vs Human-Generated Images

1. Approach and Methodology

The task of distinguishing between AI-generated and human-generated (real) images presents a unique challenge due to the subtle artifacts and sophisticated generation techniques of modern generative models. My approach involved a mix of empirical experimentation and a structured evaluation pipeline. I began with a subset of 2,000 images to quickly test model performance and iterate faster before scaling up.

I initially explored basic CNN architectures to establish a performance baseline. Once I confirmed that a simple CNN achieved only around 40% accuracy, I moved toward more robust, pre-trained architectures to leverage transfer learning. I experimented with ResNet50 and a frequency-domain adapted ResNet (“ResNetFreq”), but their performance plateaued at ~57% accuracy.

Seeking further improvements, I transitioned to the EfficientNet family, known for its balance between performance and computational efficiency. EfficientNet-B0 significantly outperformed previous models, reaching 70% accuracy out-of-the-box, and after fine-tuning and hyperparameter optimization (learning rate scheduling, augmentation, weight decay), the model reached 73% on the validation subset. I then trained the model on the full dataset, achieving a final accuracy of **91%** on internal validation and **89.3%** on the official test submission.

---

2. Model Architecture and Design Decisions

- **Initial Baseline**: Custom CNN with 4 convolutional layers and dense layers, trained from scratch. This model served to test the feasibility of the classification task but underperformed due to insufficient capacity to capture complex visual features.

- **ResNet50**: Utilized transfer learning by loading pre-trained weights. While it improved results (~57% accuracy), it failed to robustly generalize to both classes.

- **ResNetFreq**: This variation incorporated Fourier Transform-based preprocessing to amplify subtle generation artifacts. Although a novel approach, its performance was only marginally better.

- **EfficientNet-B0**: I adopted EfficientNet-B0 due to its compound scaling efficiency, which balances width, depth, and resolution. The model demonstrated strong generalization and high accuracy, making it the final choice.

Training Details:

- Optimizer: Adam
- Loss Function: CrossEntropyLoss
- Learning Rate: 1e-4 with cosine annealing scheduler
- Batch Size: 128
- Epochs: 10
- Input Size: 224×224
- Augmentations: Random flips, rotations, color jitter

---

3. Explainability Methods Implemented

To understand and interpret the decisions made by the final model, I used **Grad-CAM (Gradient-weighted Class Activation Mapping)**. Grad-CAM visualizes the most important regions of the input image that influence the model's prediction.

- **Implementation**: I applied Grad-CAM on the final EfficientNet-B0 model using the last convolutional block (model.features[-1]) as the target layer.
- **Observations**: The CAM heatmaps revealed that the model focused on regions with distinct textures or synthetic patterns in AI images, while focusing on natural object boundaries and color distributions in real images.

This visualization confirmed that the model learned meaningful and human-comprehensible patterns rather than overfitting to random noise or irrelevant features.

4. Performance Analysis

Model	Accuracy (Subset)	Final Accuracy	Test Accuracy
Custom CNN	~40%	-	-
ResNet50	~57%	-	-
ResNetFreq	~58%	-	-
EfficientNet-B0	73%	**91%**	**89.3%**

EfficientNet-B0 clearly outperformed the other models across all evaluation phases. The performance gain can be attributed to its architectural efficiency, better feature representation, and effective regularization.

---

5. Known Limitations and Potential Improvements

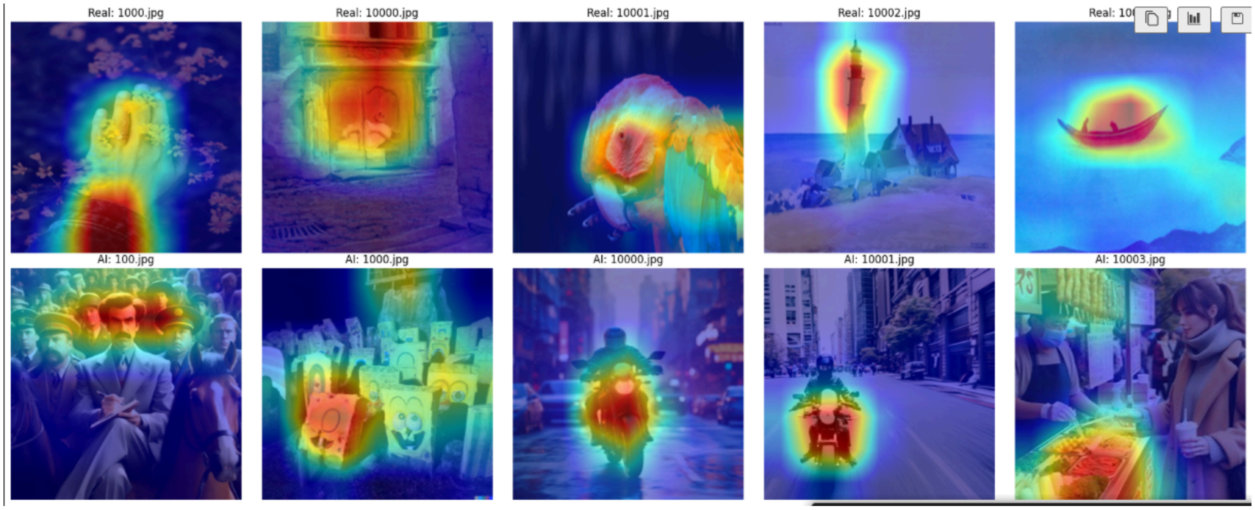
Limitations:

- The current model is limited to the dataset provided and might not generalize well to other forms of AI-generated content (e.g., newer GANs or image styles).
- High-quality AI images with minimal artifacts may still confuse the classifier.
- The dataset imbalance (if any) and the quality of images might affect performance.

Potential Improvements:

- Introduce frequency-domain features (e.g., DCT coefficients) alongside spatial features in a multimodal model.
- Use ensemble techniques combining EfficientNet and ResNetFreq.
- Incorporate adversarial training to improve robustness against subtle image perturbations.
- Fine-tune a larger variant of EfficientNet (like B3 or B4) with more data or synthetic augmentations.
-

Real Image Visualization



The Grad-CAM visualizations clearly highlight how the model focuses on meaningful regions within each image to differentiate between AI-generated and real content. Real images show attention to fine details and natural structures, while AI images trigger focus on unnatural textures and artifacts. This confirms that the model is learning interpretable and relevant features, enhancing its trustworthiness.