AI vs. Human-Generated Image Detection

Image Classification

JI SUNG HAN

DAVID HUYNH

SHAHZAD JALIL

LUKE SHEN

Data Description

Objective: Develop a machine learning model capable of accurately distinguishing AI-generated images from human-created ones.

Dataset Overview

- Real images paired with their AI-generated counterparts.
- Labels: 0 = Human, 1 = AI-Generated.
- 79,950 training images, 19,986 test images.

Training Data Sample:

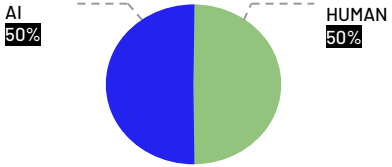
	Unnamed: 0	file_name	label
0	0	train_data/a6dcb93f596a43249135678dfcfc17ea.jpg	1
1	1	train_data/041be3153810433ab146bc97d5af505c.jpg	0

1 (AI-Generated)

0 (Human-created)

Class Distribution

- AI-Generated (50%) | Human (50%) (Balanced dataset)
- No missing values or outliers detected.
- Images vary across **74 different sizes → Resized for consistency**.

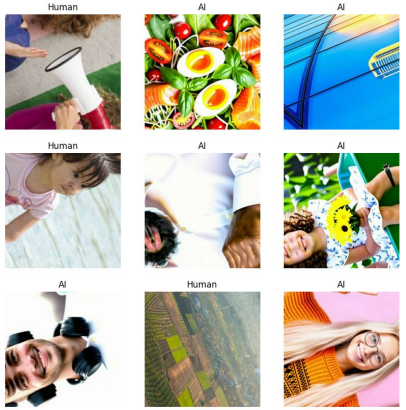


Preprocessing & Augmentation

- Resized images based on the model architecture:
  EfficientNet-B4: 380 × 380, ResNet-50: 224 × 224, (.tfRecord)
- Swin Transformer: 224 × 224 (.pt)

Hard Data Augmentation techniques applied:

- Rotation, Flipping, Color Jitter, and Random Gaussian Blur
- Prevents overfitting & improves generalization across diverse patterns.



Key Challenges

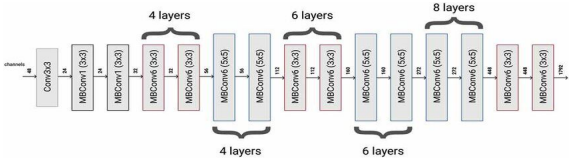
- Test data is more diverse & complex than training, impacting performance
- Consists of tens of thousands of high-resolution images, causing RAM overload and kernel errors during .pt, tfRecord file conversion and loading.

Methodology

Objective: Build a robust ensemble model combining Swin Transformer, EfficientNet, and ResNet to maximize accuracy in AI image detection.

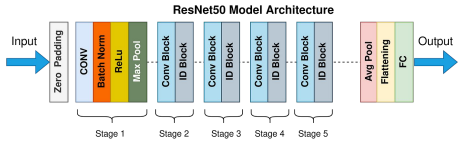
EfficientNet-B4

- Strong in fine texture recognition
- Loss Function: Binary Focal Cross Entropy.
- Optimizer: Adam



ResNet-50

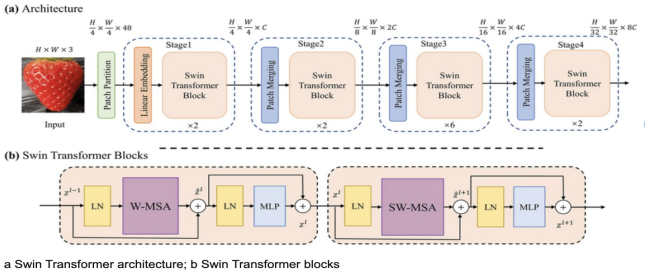
- Traditional CNNs that are robust to high-level structural patterns
- Loss Function: Binary Focal Cross Entropy.
- Optimizer: SGD with Momentum



Swin Transformer

(Swin-Base-Patch4-Window7-224)

- Excellent for detecting global patterns and expressing long-range dependencies.
- Loss Function: BCE with LogitsLoss
- Optimizer: AdamW
- Learn local information with shifted window self-attention and construct global representation hierarchically.



a Swin Transformer architecture; b Swin Transformer blocks

Key Challenges

- If use a large learning rate without warmup, the initial loss will explode or NaN will occur, and on the contrary, if it is too small, learning will be slow and it will be easy to get stuck in local minima.
- Due to the long training time of individual models, full training and testing of the ensemble model could not be completed within the time constraints.

TRAINING OVERVIEW

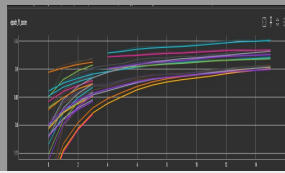
We trained each model using transfer learning with pre-trained weights, applying data augmentation and optimized with AdamW or SGD and appropriate learning rate schedulers.

TRANSFER LEARNING

- Remove and Add a new Classification head to utilize transfer learning of pretrained CNN and ViT
- Minimize Categorical (Focal) Cross Entropy for hard negative mining.

MODEL SELECTION

- Hyperband Tuning or GridSearch was used to optimize and explore hyperparameter space quickly
- Early Stopping on the validation F1 Score using a training, validation split(80 : 20)



ENSEMBLE

- Ensemble structure was designed using late fusion through logistic regression
- Results are unoptimized but are around

COMPARISON

- CutMix improved train/val scores but hurt test performance, showing poor generalization.
- ResNet50 and EfficientNet were faster and more efficient to train than Swin Transformer due to their relatively smaller model sizes, but Swin performed better

Performance

	Training – F-1 score	Validation – F-1 score	Test – F-1 Score
ResNet50	0.7948	0.7963	0.56105
Efficient NetB4	0.9254	0.9198	0.52611
Swin Transformer	0.9879	0.9923	0.68474
Ensemble (ResNet+ B4, logistic)	0.9444	0.9459	0.4848

Conclusion

- Swin Transformer showed the best performance in this task due to its structure that can efficiently learn local/global features of high-resolution images.
- ResNet50 and EfficientNet trained quickly, but their sophisticated pattern recognition performance was relatively low.
- The performance drop on the test set may stem from domain shift or overfitting, and highlights the need for model-specific feature engineering and generalization strategies.
- The ensemble structure has been completed, but final training and evaluation cannot be performed due to time and resource constraints.
- We aim to reliably learn and evaluate not only individual models but also entire ensemble models by further researching how to optimally utilize Colab's GPU resources.

THANK YOU