Capstone Project:

Detecting AI-Generated Images

Yaqin Albirawi

Artificial Intelligence and Machine Learning, Fanshawe College

INFO-6147: Deep Learning with Pytorch

Dr. Mohammed Yousefhussien

July 31, 2025

1. Introduction

With the rapid advancements in generative AI tools like Stable Diffusion, the line between real and synthetic visual content has become increasingly blurred. This project addresses the challenge of detecting AI-generated images through a binary image classification task using deep learning. By applying a ResNet-18 convolutional neural network to the CIFAKE dataset, we investigate whether a model can effectively distinguish between real and fake images.

2. Dataset Description

Source: Kaggle CIFAKE Dataset

Size: 120,000 images

50% Real: Extracted from CIFAR-10

50% Fake: Generated using Stable Diffusion conditioned on CIFAR-10

Characteristics:

Image Size: 32x32 pixels (low-resolution)

Balanced classes (60,000 real, 60,000 fake)

Challenge: The fake images closely mimic the style and resolution of real CIFAR-10 images, making them hard to distinguish visually or through shallow models.

3. Data Preprocessing

Resizing: All images retained their native 32x32 dimensions.

Normalization: Pixel values normalized to [0, 1].

Splitting:

Training: 70%

Validation: 15%

Test: 15%

Data Augmentation:

Random Horizontal Flip

Random Rotation

Color Jitter (for robustness)

4. Model Architecture

Base Model: ResNet-18 (trained from scratch)

Loss Function: Binary Cross-Entropy

Optimizer: Adam

Learning Rate: 0.001 (default; tuned manually)

Regularization:

Early stopping (patience = 3)

Data augmentation

Activation: Sigmoid (output layer)

Batch Size: 32

The model was trained without pre-trained weights to prevent overfitting to unrelated domains like ImageNet.

5. Training Process

Training spanned multiple epochs (≈20)

Loss and accuracy were monitored across training and validation sets

Early stopping halted training if validation performance plateaued for 3 epochs

Final model weights were saved based on best validation accuracy

6. Hyperparameter Tuning

Limited manual tuning was conducted due to time constraints. Key experiments included:

Varying learning rate (0.0001, 0.001, 0.01)

Adjusting batch sizes (16, 32, 64)

Comparing performance with/without augmentation

Observation: Smaller batch sizes led to slightly slower but more stable convergence; learning rate of 0.001 performed best.

7. Evaluation Metrics

Evaluated on validation and test sets using:

Accuracy

Precision

Recall

F1-score

Confusion Matrix

Metric Value (%)

Accuracy ~91%

Precision ~90%

Recall ~92%

F1-score ~91%

The classifier showed strong performance in distinguishing between real and fake samples in the CIFAKE dataset.

8. Interpretability: Grad-CAM

To interpret the model’s decisions, Grad-CAM visualizations were generated:

Real Image: Focused on key object edges and textures

Fake Image: Highlighted smoother regions and blurred artifacts

Insight: The model detects subtle generative inconsistencies, such as overly smooth textures or unrealistic object contours

9. Limitations and Improvements

Limitation: The model does not generalize well to real high-resolution images or fakes generated by other tools (e.g., DALLE, MidJourney)

Domain Shift: ChatGPT and human viewers also struggled to classify CIFAKE-style real images due to pixelation

Improvement Suggestions:

Include fake images from diverse generative sources

Train on higher resolution datasets or multi-resolution formats

Employ transfer learning and fine-tune pretrained models like EfficientNet or Vision Transformers

10. Final Testing & Generalization

When tested on high-resolution images not from CIFAKE:

Performance dropped significantly

Downscaled real images were misclassified as fake due to blur and loss of fidelity

This reveals a major challenge in detecting fake content under domain shift and resolution mismatch conditions.

11. Conclusion

This project demonstrates a practical approach to detecting AI-generated images using a CNN model trained on CIFAKE. While the model performs well within its domain, it highlights critical issues around generalization and the importance of diverse training data. Future work could explore multi-resolution training, adversarial training, and the use of multimodal cues (text+image) to enhance robustness.

12. Appendix

Latent Diffusion: Diffusion in compressed latent space instead of raw pixel space

Stable Diffusion: Generates high-quality images from text prompts using latent diffusion

Generative Models: Create images/text by learning data distributions

Diffusion Models: Learn to denoise a sample to generate a new data instance