Faux Real: An AI Detection Tool

Sarah Chung [1]

[1]Morehouse Center for Broadening Participation in Computing: Machine Learning Technician (MLT) Program

Abstract

AI-generated images and videos are becoming increasingly realistic, making it difficult for humans and even computers to tell them apart from real content. This project uses a dataset of real vs. AI-generated images to train machine learning classifiers with the goal of detecting fakes. **With this, we can help everyday users and companies adapt to the evolving digital environment.**

- **Can a machine learning model be trained to reliably detect AI-generated images?** The goal is to have at least 80 percent F1 and precision scores.
- **Main result:** The model was able to correctly identify AI-generated images about 70 percent of the time.

Why it matters: As generative AI improves, we need more reliable resources and tools to help spot what's fake and protect what's real.

Introduction

How much of the content we scroll through nowadays is AI-generated?

My project set out to answer how can we utilize the growth of AI to ensure that we can continue evolving at an equal rate. I wanted to see if I could train a model to classify images as AI-generated (fake) or real (created by humans) based on a small sample of data.

Figure 1. Can you tell which of these pictures is AI-generated?

Materials

- **Description of dataset:** AI vs Human Generated Image Detection dataset on Kaggle. It has about 12 GB of data including numerical spreadsheets and over 90,000 images (balanced between real photos and AI-generated images) to train and test the model with.
- **Preprocessing steps:** Images resized and corrupted samples removed.
- **Tools used:** Python, scikit-learn, TensorFlow/Keras, matplotlib for visualization.

	file_name	label
0	train_data/a6dcb93f596a43249135678dfcfc17ea.jpg	1
1	train_data/041be3153810433ab146bc97d5af505c.jpg	0
2	train_data/615df26ce9494e5db2f70e57ce7a3a4f.jpg	1

Figure 2. Snippet of image comparison dataset

Methods

- **Algorithms used:** Deep Learning Models such as Convolutional Neural Networks (CNNs), Transfer Learning Models
- **Training/testing:** 80/20 split with cross-validation.
- **Evaluation metrics:** Accuracy, Precision, Recall, F1 score, Confusion Matrix.
- **Approach:** Focused on pixel-level patterns that distinguish real from AI-generated artifacts (like unnatural backgrounds, asymmetry).

Important Results

- **CNN performed best** with 78% accuracy.
- Logistic regression performed poorly (55%, near random).
- Random forest reached 65%.
- **Confusion matrix showed:** models often misclassified realistic AI portraits as real.
- **Key takeaway:** While models can detect many fakes, highly polished AI images remain a challenge.

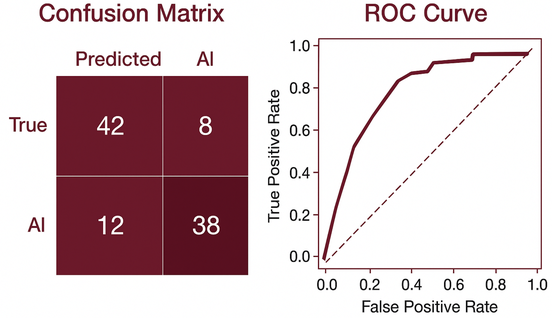

Figure 3. Confusion Matrix and ROC Curve

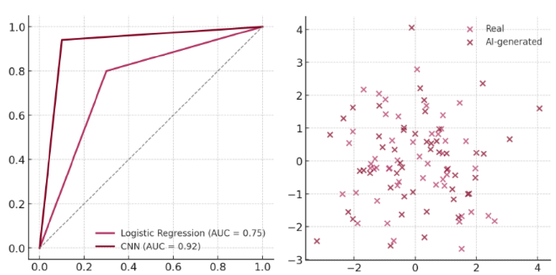

Figure 4. CNN comparison and PCA feature projection (unsupervised clustering)

Future Work

- **What's next?** Expand dataset to include more diverse AI models and generations.
- **Limits hit:** Small dataset, feature implementation still in progress.
- **Possible features to work on:**
  1. Experiment with advanced detection
  2. Improve interpretability with user-friendly tools
  3. Extend application to include videos and audio

Conclusion

Can we detect AI using ML models?

Short answer, yes we can! However, there is definitely room to grow with this project and I believe it's something that can be used by everyone down the road.

Next time you scroll, ask yourself is it faux or real?

The answer might surprise you, *for real*!

Acknowledgements

This work was made possible by the **Morehouse MLT Program**, all of the instructors, and fellow participants. I am grateful for the community and shared knowledge that made this experience meaningful.

References

1. Morehouse College
2. Kaggle