AI-Generated Synthetic Images Differentiation

Efe Eboreime, Steve Hudgson, Weixiao Liao, Minh Nguyen

Rice University

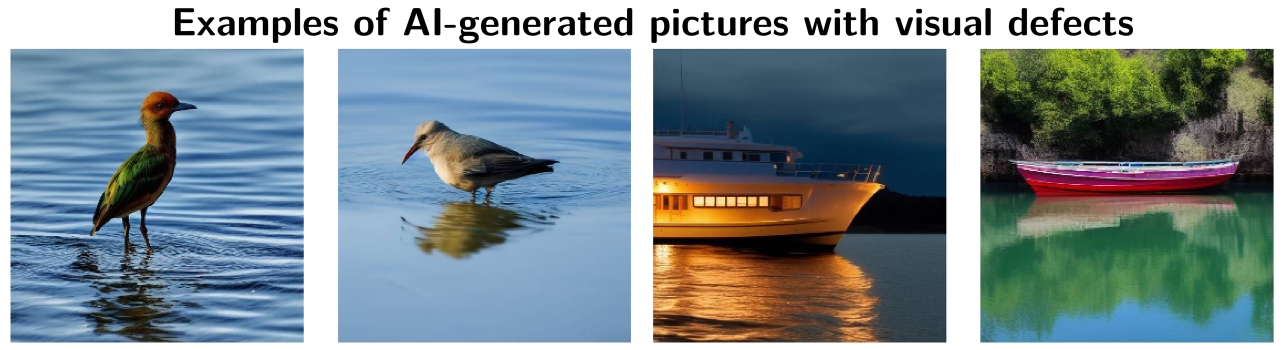
Abstract

The purpose of this research is to investigate the application of advanced neural networks in distinguishing real from AI-generated images. The study will leverage the CIFAKE dataset, which collects 60,000 synthetically-generated images and another 60,000 real images. Our methodology will focus on developing and training a conventional neural networks (CNNs) with mutiple dense layers that can detect AI-generated image. We plan to compare the performance of our custom CNN with that of the established Resnet-50. The evaluation of the models performance will be based on precision and accuracy metrics. The intention of this research is to enhance the capability of neural networks to determine the genuineness of images and to broaden our understanding of the role AI plays in the realm of image analysis.

Motivation & Background

Distinguishing authenticity from AI-generated images is pivotal for technological ethics and security. The widespread use of AI-generated visuals introduces risks in societal, informational, and cybersecurity domains. Of concern is the creation of synthetic evidence for criminal activities, exemplified by models like Stable Diffusion Models (SDMs) crafting deceptive visuals to cast doubt on legal proceedings. Machine-generated images infiltrating fake news compound the challenge of discerning truth, posing a formidable threat to public opinion. In cybersecurity, lifelike synthetic personas in false acceptance attacks jeopardize digital authentication, with generative models adapting to overcome signature verification systems. Techniques for identifying synthetic elements include optical flow methods achieving 81.61% accuracy in detecting synthetic human faces and efficient systems using EfficientNets and Vision Transformers with an F1 score of 0.88 and an AUC of 0.95. The rapidly evolving landscape necessitates staying current on methodologies and challenges in synthetic image detection for effective navigation.

Examples of AI-generated pictures with visual defects



Examples of AI-generated pictures with complex visual attributes

References

Barua, Sukarna, Sarah Monazam Erfani, and James Bailey. "FCC-GAN: A Fully Connected and Convolutional Net Architecture for GANs." *arXiv preprint arXiv:1905.02417*, 2019, pp 3-8.

Bird, J. J. "CIFAKE: Real and AI-Generated Synthetic Images." *Kaggle*, 2023, www.kaggle.com/datasets/birdy654/cifake-real-and-ai-generated-synthetic-images.

Krizhevsky, Alex, and Geoffrey Hinton. "Learning Multiple Layers of Features from Tiny Images." 2009.

Bird, J. J., and A. Lotfi. "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images." *arXiv preprint arXiv:2303.14126*, 2023.

Singh, B., and D. K. Sharma. "Predicting Image Credibility in Fake News over Social Media Using a Multi-Modal Approach." *Neural Computing and Applications*, vol. 34, no. 24, 2022, pp. 21503-21517.

Khosravy, M., et al. "Model Inversion Attack: Analysis under Gray-Box Scenario on Deep Learning-Based Face Recognition System." *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 15, no. 3, 2021, pp. 1100-1118.

Bird, J. J., A. Naser, and A. Lotfi. "Writer-Independent Signature Verification; Evaluation of Robotic and Generative Adversarial Attacks." *Information Sciences*, vol. 633, 2023, pp. 170-181.

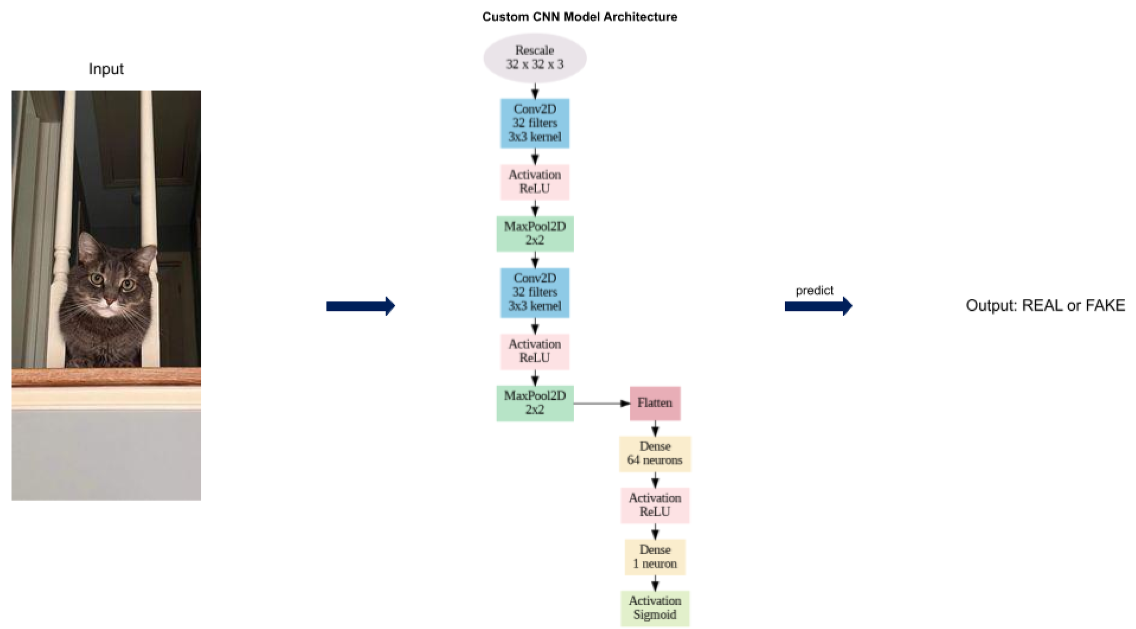
Sha, Z., et al. "De-Fake: Detection and Attribution of Fake Images Generated by Text-to-Image Diffusion Models." *arXiv preprint arXiv:2210.06998*, 2022.

Amerini, I., et al. "Deepfake Video Detection through Optical Flow-Based CNN." *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

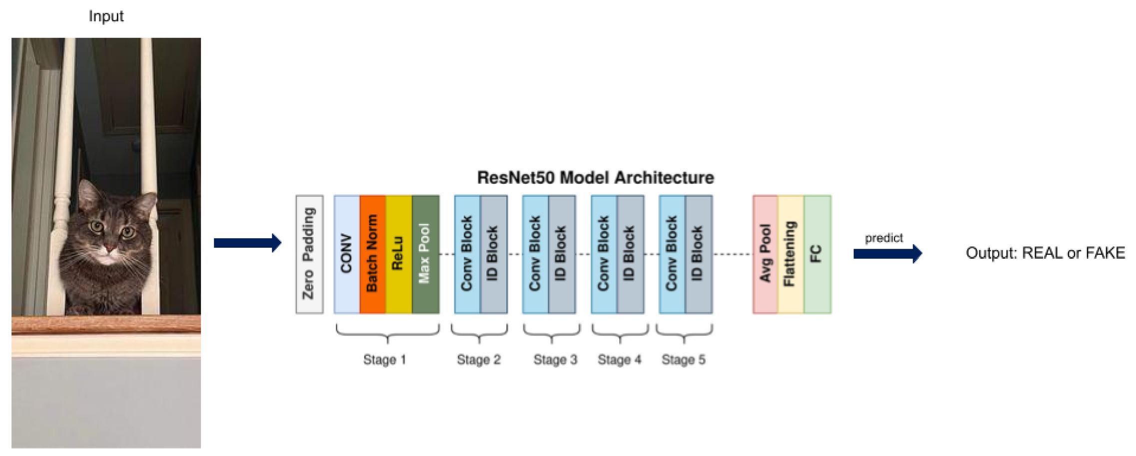
Wang, J., et al. "M2TR: Multi-Modal Multi-Scale Transformers for Deepfake Detection." *Proceedings of the 2022 International Conference on Multimedia Retrieval*, 2022, pp. 615-623.

Model

- **Method 1**: Custom Convoluted Neural Network (CNN)



- **Method 2**: ResNet50



- **Dataset**: CIFAKE

	Number of REAL images	Number of FAKE images
Train	50,000	50,000
Test	10,000	10,000

- **Metrics**:
  - **Precision**: This is the ratio of true positive predictions to the total number of positive predictions made. It is a measure of the accuracy of the positive predictions, accounting for false positives. The formula for precision is given by:

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

  - **Recall**: This is the ratio of true positive predictions to the actual number of positive cases. It measures how many actual positive cases were correctly identified and is crucial for analyzing false negatives. The formula for recall is:

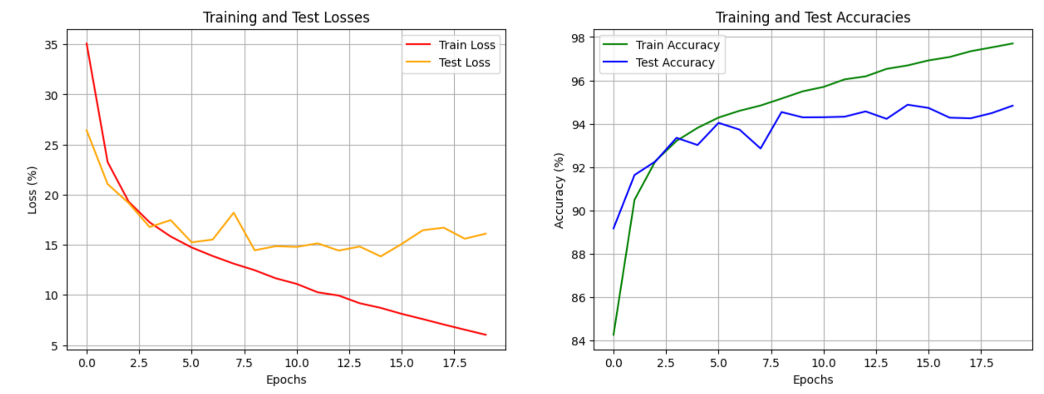
$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

  - **F1-Score**: This score is a harmonic mean of precision and recall, providing a balance between them. It is especially important in situations like fraud detection, where a false negative could have serious implications. The F1-score is calculated as:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

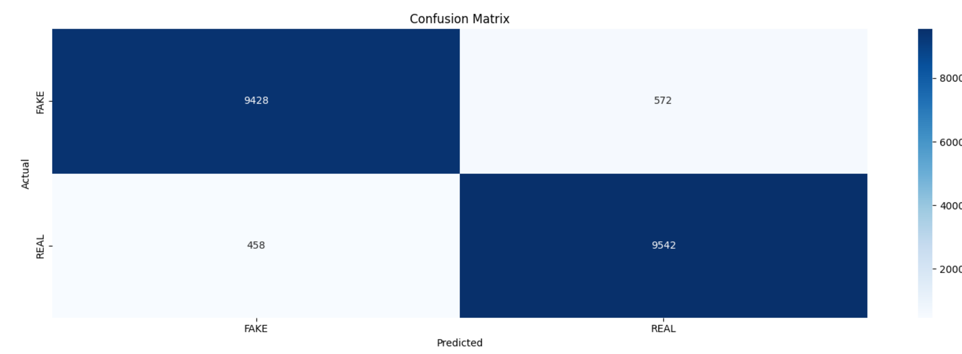
- **Implementation**: In this experiment, we applied a custom CNN to the CIFAKE dataset, which contained an equal mixture of AI-generated and real images. The CNN was meticulously trained and validated to distinguish between synthetic and authentic images, aiming to achieve high precision and recall. The implementation highlighted the network's capability to discern image authenticity, showcasing its potential application in areas such as digital forensics and cybersecurity.

Results

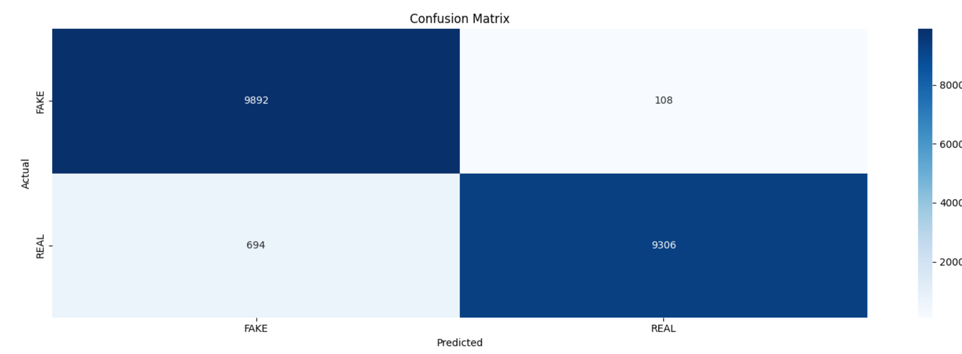


Loss and Accuracy plots for custom CNN

Loss and Accuracy plots for Resnet-50



Confusion Matrix for Custom CNN



Confusion Matrix for Resnet-50

Model	Precision	Recall	F1-Score
Custom CNN	0.9434	0.9542	0.9488
Resnet-50	0.9885	0.9306	0.9587

Performance Metrics

Conclusion

This study has proposed the possibility of differentiating synthetically AI-generated images from real ones. Utilizing the CIFAKE dataset, we employed a Convolutional Neural Network (CNN) architecture to classify images into two categories: authentic and fabricated. The accuracy of the classification was approximately 94%, utilizing a total of 120,000 images, with 20,000 allocated for testing and 100,000 dedicated for training. Notably, in comparison, the ResNet-50 model demonstrated a slightly higher accuracy, which could be attributed to its deeper architecture and sophisticated feature extraction capabilities

Future work for this project could involve implementing Generative Adversarial Networks (GANs) along side with our current CNN architecture in order to see if there will be a difference in performance and scalability.

This type of work will help different areas in computing, such as cyber security and data management, by ensuring data authenticity and trustworthiness when dealing with different datasets that one is not sure what origin they are from.

Github Repository: https://github.com/minhnguyen1406/AI-Generated-Synthetic-Images-Differentiation