Detecting AI-Generated Synthetic Images using Deep Learning

Swathi Ashok Kumar
Department of Computing and Mathematics
Manchester Metropolitan University
23623012@stu.mmu.ac.uk

***Abstract*—The rise of AI-generated synthetic images, represented by the CIFake dataset, poses significant risks, including the potential for malicious use in creating deepfakes and spreading misinformation. This study aims to address this issue by developing an effective deep learning-based detection system. Utilizing convolutional neural networks (CNNs), we preprocess and analyze the CIFake dataset to differentiate between real and AI-generated images. Our methodology includes detailed data curation, rigorous model training, and extensive experimentation. The results demonstrate that our approach achieves high accuracy and reliability in identifying synthetic images. This work underscores the critical role of deep learning techniques in safeguarding digital media integrity and provides a foundation for future enhancements in detection capabilities.**

Keywords—AI-generated images, Convolutional neural networks (CNNs), Image classification.

I. Introduction

The rapid advancement of AI-generated synthetic images has created a significant challenge in identifying authentic images from real ones. These AI-generated images, often nearly identical to real ones, pose threats to the credibility of digital media and can be exploited for disinformation and fraud. To address this issue, we focus on using the CIFAKE dataset, which contains images labeled as "Real" or "Fake."

The input to our algorithm is an image, and we use a Residual Network (ResNet-18) to output a prediction indicating whether the image is real or AI-generated. Previous research has shown the effectiveness of deep learning methods like CNNs and DenseNet, but these often require high computational resources and complex architectures. There is a need for a more efficient yet accurate approach.

Our aim is to develop an efficient and accurate model using ResNet-18 to distinguish AI-generated images from real ones. We preprocess the CIFAKE dataset, implement and fine-tune the ResNet-18 model, and evaluate its performance using metrics such as accuracy, precision, recall, and F1 score. This study aims to provide a robust solution for maintaining the integrity of digital media.

II. Related Work

Research on detecting AI-generated images employs both traditional image processing and deep learning methods. Traditional approaches, such as manual feature extraction combined with Support Vector Machines (SVM), offer simplicity and lower computational costs but suffer from limited scalability and adaptability to complex data (Hearst et al., 1998). In contrast, deep learning methods like Convolutional Neural Networks (CNNs) provide high accuracy due to automatic feature extraction but are computationally intensive and prone to overfitting. Bird and Lotfi (2023) used a CNN with Gradient Class Activation Mapping (Grad-CAM) for explainability, achieving 92.98% accuracy by focusing on visual imperfections (Bird & Lotfi, 2023). Advanced architectures such as ResNet, VGGNet, and DenseNet further enhance feature extraction and depth handling capabilities. Wang, Hao, and Cong (2023) found DenseNet to be the most effective with a 97.74% accuracy, followed by ResNet at 94.95% (Wang, Hao, & Cong, 2023). Our approach utilizes ResNet-18, balancing depth, and efficiency with residual connections to address the vanishing gradient problem, making it less computationally intensive than DenseNet while maintaining high performance. This choice leverages the strengths of advanced architectures, providing a robust solution for distinguishing real from fake images.

III. Dataset

The CIFAKE dataset, designed to tackle the challenge of distinguishing real images from AI-generated ones, consists of 120,000 images equally divided into real and fake categories. Real images are sourced from the CIFAR-10 dataset by Krizhevsky and Hinton (2009), featuring classes such as automobiles, airplanes, birds, deer, cats, dogs, frogs, ships, horses, and trucks. Fake images are generated using Stable Diffusion version 1.4, mimicking CIFAR-10 classes.

Dataset	Real Images	Fake Images	Total Images
Training	50000	50000	100000
Testing	10000	10000	20000
Total	60000	60000	120000

This table clearly shows the distribution of real and fake images across the training and testing sets.

A. Data Preprocessing

In this study, the CIFAKE dataset undergoes several preprocessing steps to prepare the images for training and testing using the ResNet-18 model. The preprocessing involves resizing, converting to tensors, and normalizing the images. These steps ensure that the data is standardized and suitable for input into the deep learning model.

Each image is resized to a standard dimension of 224x224 pixels to match the input size expected by ResNet-18.

The pixel values are normalized using the mean and standard deviation of the ImageNet dataset. This step standardizes the input distribution, improving model training stability and convergence.

Mean: [0.485, 0.456, 0.406]

Standard Deviation: [0.229, 0.224, 0.225]

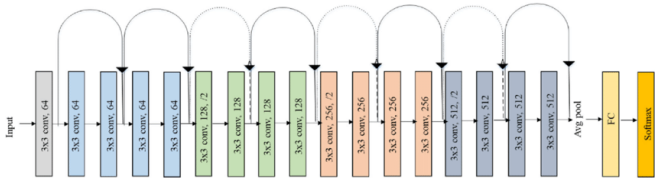
IV. Methodology

The objective of this study is to differentiate between real and AI-generated synthetic images by employing the ResNet-18 model, a type of Convolutional Neural Network (CNN). The CIFAKE dataset, which includes equal numbers of real and fake images, was preprocessed, and divided into training, validation, and test sets. The approach includes data preprocessing, model initialization, training, validation, and checkpointing to ensure the best model is saved.

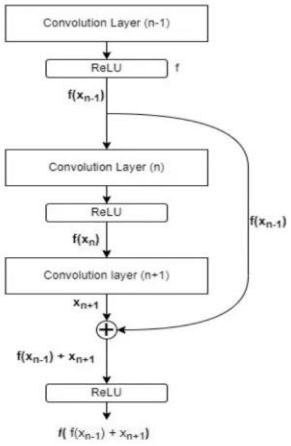
A. RestNet18 Architecture:

For this project, we utilized the ResNet-18 architecture to determine if images from the CIFAKE dataset are real or generated by AI. ResNet-18 is part of the ResNet family, which uses shortcuts or skip connections to improve the training of deep networks. These shortcuts help the network learn better by skipping some layers and reducing the vanishing gradient problem.

In our approach, we leveraged transfer learning by fine-tuning a pretrained ResNet-18 model to classify the CIFAKE images. This way, we could use the knowledge the model had already gained from being trained on a large dataset like ImageNet and apply it to our specific task of distinguishing real images from AI-generated ones.



The ResNet-18 architecture comprises 18 layers, beginning with a 7x7 convolutional layer with 64 filters and a stride of 2, followed by a max pooling layer for initial feature extraction. It includes residual blocks, each containing two 3x3 convolutional layers, batch normalization, and ReLU activation. Skip connections add each block's input to its output, addressing the vanishing gradient problem. These blocks are organized into four layers with progressively increasing filter sizes: 64, 128, 256, and 512, each containing 2 blocks. The final output passes through an average pooling layer and a fully connected layer, adjusted to output two classes for our binary classification task on the CIFAKE dataset, distinguishing real images from AI-generated ones.



Loss Function:

The Cross Entropy Loss function was selected for this binary classification task. This loss function is ideal for classification problems as it evaluates the performance of a model that outputs probabilities ranging from 0 to 1. It measures the disparity between the true class labels and the predicted probabilities, with a higher penalty for predictions that deviate significantly from the actual labels. The formula for Cross Entropy Loss is:

$$-\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p(y_i)) + (1-y_i) \cdot \log(1-p(y_i))$$

where Yi denotes the actual class and $\log(p(Yi))$ represents the probability of that class.

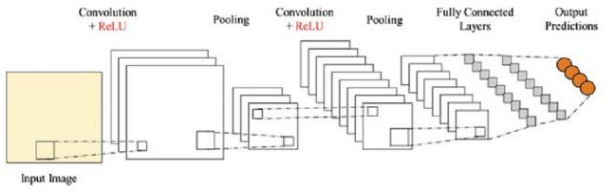
Hypothesis: By fine-tuning a pretrained ResNet-18 model on the CIFAKE dataset, it is expected that the classification performance will significantly improve compared to training a model from scratch. The pretrained model's ability to transfer learned features is anticipated to effectively identify the subtle differences between real and AI-generated images.

Optimizer:

The Adam optimizer was selected, and its implementation is shown as
optim.Adam(model_restnet18.parameters(), lr=0.001)
Adam is a first-order gradient-based optimisation technique that is selected due to its computational efficiency and adaptability. During training, a learning rate of 0.001 is chosen for ensuring a slow and consistent convergence.

B. Custom CNN Architecture:

In addition to ResNet-18, a custom Convolutional Neural Network (CNN) was implemented to compare performance.



The CNN architecture included:
- An initial convolutional layer with a 3x3 kernel, followed by ReLU activation and max pooling.
- Multiple convolutional layers to extract features, each followed by ReLU activation and pooling.
- Fully connected layers to aggregate the features and output the final class probabilities.

The custom CNN model was trained using the same dataset and preprocessing techniques as ResNet-18, allowing for a direct comparison of their performance.

V. Experimental Results and Discussion

A. Aim:

The aim was to develop and evaluate a deep learning system using ResNet-18 and a custom CNN to distinguish between real and AI-generated images in the CIFake dataset. The objective is to determine which model performs better based on metrics such as accuracy, precision, recall, F1 score, and AUC.

B. Hypothesis:

ResNet-18 Hypothesis: The ResNet-18 model, due to its advanced architecture and pretrained weights, will outperform the custom CNN in distinguishing real from AI-generated images.
Custom CNN Hypothesis: The custom CNN will achieve reasonable accuracy but will perform slightly worse than ResNet-18 due to its simpler architecture.

C. Models Evaluation Metrics:

To thoroughly evaluate the performance of a classification model, several metrics are used. Below are the key evaluation metrics used in this study:

Accuracy: Accuracy is measured as the ratio of correctly predicted values to the total values.

$$\text{Accuracy} = \frac{TP+TN+FP+FN}{TP+TN}$$

Precision: Precision is calculated as the ratio of correctly predicted positive values to the total predicted positive values.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall: Recall is defined as the ratio of correctly predicted positive values to the actual positive values.

$$\text{Recall} = \frac{TP+FN}{TP}$$

Here,
TP ="True Positive"
FP ="False Positive"
FN = "False Negative"
TN = "True Negative"

F1 Score: The F1 Score is computed as the harmonic mean of precision and recall.

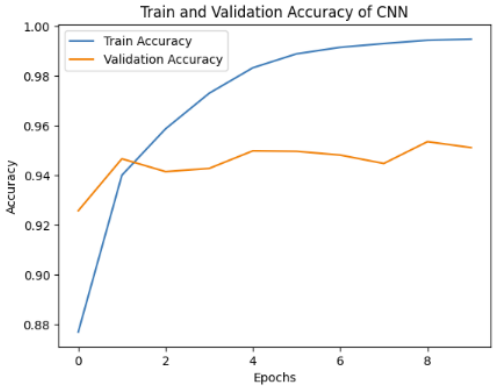
$$\text{F1 Score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision} + \text{Recall}}$$

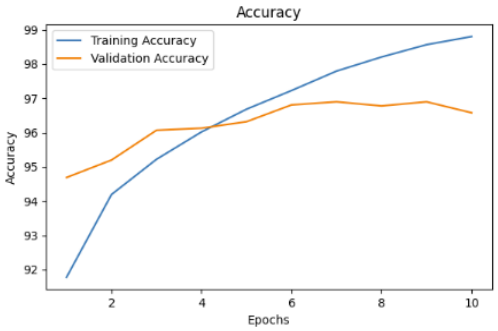
AUC-ROC: AUC-ROC represents the area under the Receiver Operating Characteristic curve, which plots the true positive rate against the false positive rate across different threshold settings.

Confusion Matrix: The confusion matrix offers a detailed breakdown of correct and incorrect classifications, showing the counts of true positive, true negative, false positive, and false negative predictions.

D. RESULTS:

The accuracy curves for both the Custom CNN model and the ResNet-18 model across 10 epochs are shown below:



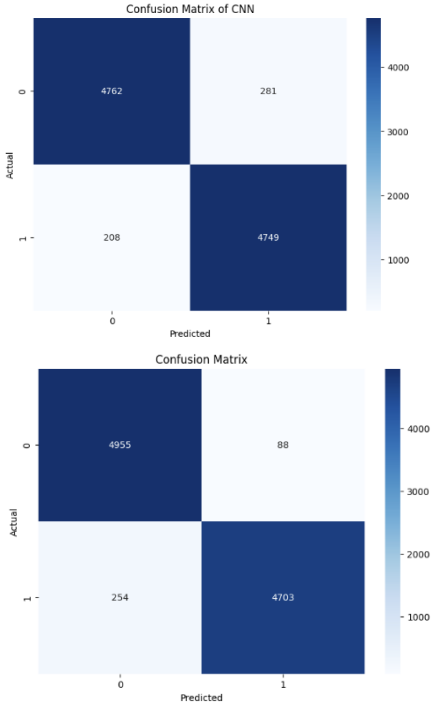


Both models demonstrate increasing training accuracy, with ResNet-18 achieving a higher final training accuracy (~99%) compared to the Custom CNN.
In terms of validation accuracy, Custom CNN shows minor fluctuations around 95%, while ResNet-18 stabilizes at a higher accuracy of around 96%. This indicates that ResNet-18 not only fits the training data better but also generalizes more effectively to unseen data.

The smaller gap between training and validation accuracy in ResNet-18 suggests it handles overfitting better than Custom CNN. Overall, ResNet-18 outperforms Custom CNN, making it more reliable for distinguishing between real and AI-generated images.

E. Confusion Matrix:

Below are the confusion matrices for both the ResNet-18 and Custom CNN models:



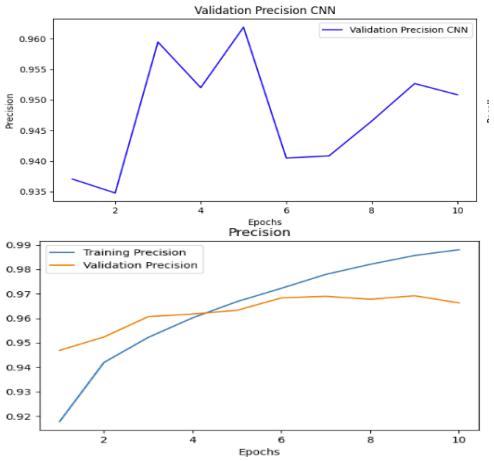
True Positives (TP): Both models have a high count of true positives (ResNet-18: 4703, CNN: 4749), indicating that most AI-generated images are correctly identified.
True Negatives (TN): ResNet-18 shows a higher number of true negatives (4955) compared to the Custom CNN (4762), indicating better performance in correctly identifying real images.
False Positives (FP): ResNet-18 has fewer false positives (88) compared to the Custom CNN (281), which means ResNet-18 is better at avoiding misclassification of real images as AI-generated.
False Negatives (FN): ResNet-18 has more false negatives (254) compared to Custom CNN (208), indicating that it sometimes misses AI-generated images more frequently than Custom CNN.

Metric	**ResNet-18**	**Custom CNN**
True Positives (TP)	4703	4749
True Negatives (TN)	4955	4762
False Positives (FP)	88	281
False Negatives (FN)	254	208

Overall, ResNet-18 provides a balanced performance with high accuracy in distinguishing between real and AI-generated images, while Custom CNN excels slightly more in detecting AI-generated images but with more false positives.

F. Precision:

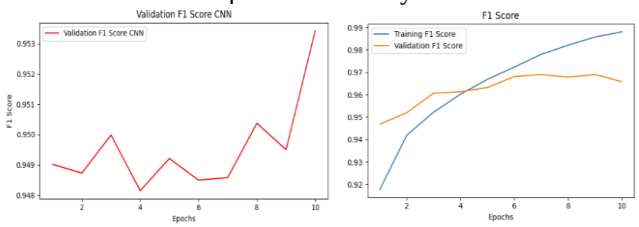
The precision curves for both the Custom CNN and ResNet-18 models across 10 epochs were analyzed.



For Custom CNN, precision exhibited fluctuations across epochs, indicating variability in its ability to correctly identify AI-generated images. Despite these fluctuations, precision remained relatively high, generally above 94%. In contrast, the ResNet-18 model demonstrated a steady increase in precision during training, reaching close to 99% by the final epoch. The validation precision for ResNet-18 also improved and stabilized around 96% over the epochs. The smaller gap between training and validation precision for ResNet-18 suggested better generalization compared to the Custom CNN.

G. F1 Score:

The F1 score curves for both the Custom CNN and ResNet-18 models across 10 epochs were analyzed.



ResNet-18 showed a consistent increase in F1 score, stabilizing around 97% for validation, indicating a strong balance between precision and recall and better generalization. In contrast, Custom CNN's F1 score fluctuated around 95%, suggesting less stability. Overall, ResNet-18 outperformed Custom CNN, providing more reliable performance in distinguishing real from AI-generated images.

VI. Conclusion:

Based on the validation metrics, it can be concluded that the ResNet-18 model outperformed the Custom CNN in terms of accuracy, precision, recall, and F1 score. The ResNet-18 model demonstrated a more consistent and reliable performance, particularly in precision, where it showed less variability and higher stability. This higher precision and better generalization are critical for applications where minimizing false positives is essential. Overall, ResNet-18 is more suitable for distinguishing between real and AI-generated images due to its robust and balanced performance across multiple metrics.

REFERENCES:

[1] T. Guo, J. Dong, H. Li and Y. Gao, "Simple convolutional neural network on image classification," 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA), Beijing, China, 2017, pp. 721-724, doi: 10.1109/ICBDA.2017.8078730.

[2] J. J. Bird and A. Lotfi, "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images," Nottingham Trent University, Nottingham, UK.

[3] Y. Wang, Y. Hao, and A. X. Cong, "Harnessing Machine Learning for Discerning AI-Generated Synthetic Images," arXiv preprint arXiv:2401.07358, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2401.07358.

[4] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-Generated Images Are Surprisingly Easy to Spot... for Now," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8695-8704.

[5] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[6] H. A. Khalil and S. A. Maged, "Deepfakes Creation and Detection Using Deep Learning," 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), Cairo, Egypt, 2021, pp. 1-4, doi: 10.1109/MIUCC52538.2021.9447642.

[7] A. V. S. Abhishek, V. R. Gurrala, and L. Sahoo, "Resnet18 Model With Sequential Layer For Computing Accuracy On Image Classification Dataset," Department of Computer Science and Engineering, GITAM University, Visakhapatnam, Andhra Pradesh, India.

[8] F. Ramzan, M. U. G. Khan, A. Rehmat, S. Iqbal, et al., "A Deep Learning Approach for Automated Diagnosis and Multi-Class Classification of Alzheimer's Disease Stages Using Resting-State fMRI and Residual Neural Networks," Journal of Medical Systems, vol. 44, no. 2, Dec. 2019, doi: 10.1007/s10916-019-1475-2.

[9] A. S. Gupta, K. P. Shreneter and S. Sehgal, "Visual Veracity: Advancing AI-Generated Image Detection with Convolutional Neural Networks," 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2024, pp. 1-6, doi: 10.1109/ICRITO61523.2024.10522113.

[10] A. Roberts, "Binary Cross Entropy: Where To Use Log Loss In Model Monitoring," Arize, Jan. 01, 2023.

[11] G. E. Bartos and S. Akyol, "Deep Learning for Image Authentication: A Comparative Study on Real and AI-Generated Image Classification," Alba Regia Technical Faculty, Obuda University, Szekesfehervar, Hungary, and Faculty of Engineering and Natural Sciences, Kütahya Health Sciences University, Kütahya, Turkey.

[12] M. Z. Hossain, F. Uz Zaman and M. R. Islam, "Advancing AI-Generated Image Detection: Enhanced Accuracy through CNN and Vision Transformer Models with Explainable AI Insights," 2023 26th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 2023, pp. 1-6, doi: 10.1109/ICCIT60459.2023.10440990.

[13] H. V, K. P and M. A, "Art of Detection: Custom CNN and VGG19 for Accurate Real Vs Fake Image Identification," 2023 6th International Conference on Recent Trends in Advance Computing (ICRTAC), Chennai, India, 2023, pp. 306-312, doi: 10.1109/ICRTAC59277.2023.10480775.

[i]

One Drive Link: https://stummuac-my.sharepoint.com/:u:/g/personal/23623012_stu_mmu_ac_uk/Eb5MYemby6VCo7hGGK92pgQB3ocZzJbmex_t5jIaje1MzQ?e=QjM7Tc