CIFAKE: Image Classification and Explainable Identification of AI - Generated Synthetic Images

S. Annapurna (21A31A4427)
P. Samyuktha (21A31A4418)
B. Srikanth (21A31A4443)
G. Sneha Ratna (21A31A4410)
G. L. Shiva Teja (21A31A4445)

ABSTRACT

A synthetic dataset was generated using latent diffusion to create high-quality AI images that mimic the CIFAR-10 dataset, enabling comparison with real photographs. The study addressed a binary classification problem, distinguishing between real and AI-generated images, using Convolutional Neural Networks (CNNs) for classification. After training 36 CNN architectures and optimizing hyperparameters, the model achieved an accuracy of 92.98% in correctly classifying images. Gradient Class Activation Mapping (Grad-CAM) was applied to interpret the model's decision-making process, revealing that the model relied on small background imperfections rather than the main entity in the image. The CIFAKE dataset, designed for this study, has been made publicly available to the research community to foster further work on AI image recognition and authenticity detection.

EXISTING SYSTEM OVERVIEW

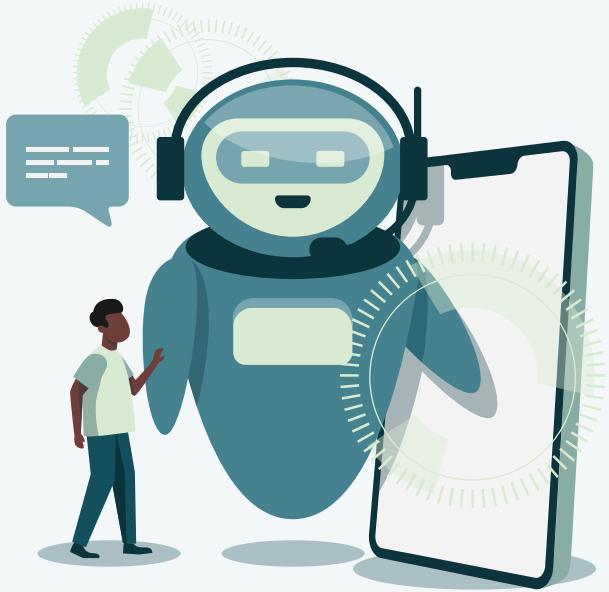
The existing systems for detecting AI-generated images use a variety of advanced techniques, including Latent Diffusion Models (LDMs) like Stable Diffusion, DALL-E, and Imagen, which generate high-quality synthetic images. These models produce images with complex visual features, making detection challenging. Methods like DE-FAKE, EfficientNet, and Vision Transformers have shown good performance in detecting synthetic images, while Optical Flow Techniques and hybrid CNN-LSTM models are used for detecting synthetic human faces and deepfake videos. However, these systems face several limitations, including struggles with high-quality images, generalizability across different datasets, and issues with interpretability.

LIMITATIONS OF EXISTING SYSTEM

- **Reliance on Visual Glitches**: Methods depend on visible imperfections, which are rare in advanced models.
- **Limited Generalizability:** Detection systems are often tailored to specific datasets, limiting their boarded application.
- **Lack of Interpretability**: Most models are black-boxes, making them unreliable for sensitive use cases.
- **Missed Subtle Features**: Current methods often overlook small imperfections in high-quality images.
- **Accuracy Challenges**: Existing methods struggle with low accuracy, especially for high-fidelity images.

PROPOSED SYSTEM OVERVIEW

The proposed system uses a robust Convolutional Neural Network (CNN) to classify images as real or synthetic, achieving an impressive accuracy of 92.98%. The CNN model is fine-tuned with hyperparameter optimization and enhanced by Gradient Class Activation Mapping (Grad-CAM), an explainable AI technique that visualizes which parts of an image are most influential in the model's decision-making process. The system introduces a new dataset, CIFAKE, comprising 120,000 images (60,000 real and 60,000 synthetic). A dynamic retraining mechanism ensures that the model adapts to emerging synthetic image features, making it effective over time. This approach focuses on detecting subtle visual anomalies in high-fidelity synthetic images, offering a more comprehensive solution than existing methods.



ADVANTAGES OF PROPOSED SYSTEM

- **Higher Accuracy**: The CNN achieves 92.98% accuracy, outperforming existing methods for high-quality synthetic images.
- **Improved Explainability**: Grad-CAM visualizes key features, enhancing model transparency over black-box methods.
- **Dynamic Adaptability**: The system updates itself through retraining to handle new synthetic image challenges.
- **Comprehensive Dataset**: CIFAKE provides a diverse and high-quality dataset for better training and scalability.
- **Focus on Subtle Imperfections**: The system detects small glitches and anomalies, improving detection of high-fidelity synthetic images.

SOFTWARE REQUIREMENTS

Operating System	Windows 10/11
Development Software	Python 3.10
Programming Language	Python
Integrated Development Environment (IDE)	Visual Studio Code
Front End Technologies	HTML5, CSS3, Java Script
Back End Technologies or Framework	Django
Database Language	SQL
Database (RDBMS)	MySQL
Database Software	WAMP or XAMPP Server
Web Server or Deployment Server	Django Application Development Server
Design/Modelling	Rational Rose

ALGORITHMS

- **Convolutional Neural Network (CNN):** A deep learning model that learns features from images for tasks like classification.
- **Gradient-weighted Class Activation Mapping (Grad-CAM):** A technique that visualizes image areas influencing a model's prediction.

CONCLUSION

- The study achieved 92.98% accuracy in recognizing AI-generated images and created the CIFAKE dataset for future research.
- It addresses the challenge of ensuring the authenticity of visual data.

FUTURE IMPLEMENTATION

- Future work could explore attention-based methods to improve accuracy.
- Updating the CIFAKE dataset and expanding to other domains could enhance the approach’s applicability.
- Exploring cross-domain applications, such as in healthcare or facial recognition, could broaden the method's impact.

THANK YOU