Enhancing Data Authenticity: A Convolutional Neural Network Approach for Distinguishing Real and AI-Generated Images

1st Sajid Hasan
dept. of CSE
Ahsanullah University Of Science And Technology
Dhaka, Bangladesh
sajid.cse.200104148@aust.edu

2nd Tabassum Tara Lamia
dept. of CSE
Ahsanullah University Of Science And Technology
Dhaka, Bangladesh
tabassum.cse.200104128@aust.edu

3rd Parvez Ahammed
dept. of CSE
Ahsanullah University Of Science And Technology
Dhaka, Bangladesh
parvez.cse.200104129@aust.edu

4th Ashraf Uz Zaman
dept. of CSE
Ahsanullah University Of Science And Technology
Dhaka, Bangladesh
ashraf.cse.200104144@aust.edu

***Abstract*—Artificial intelligence (AI) generated images are now so superior to photographs taken in real life that people are unable to distinguish between the two. This is due to recent developments in synthetic data.Artificial GAN-based synthesized images have been widely spread over the Internet with the advancement in generating naturalistic and photo-realistic images.In light of the vital importance of data authenticity and reliability, this article suggests to improve our computer vision-based recognition of AI-generated images.Initially, a synthetic dataset is generated that mirrors the ten classes of the already available CIFAR-10 dataset with latent diffusion, providing a contrasting set of images for comparison to real photographs.The study then suggests classifying the photos into two groups: Real and Fake, using a Convolutional Neural Network (CNN).A fine-tuning approach was employed using three pre-trained convolutional neural network architectures, namely ResNet50V2, EfficientNetV2B0, and EfficientNetB7.The best method could accurately categorize the pictures with 97.509% accuracy. Our best detector was a pre-trained EfficientNetB7 and EfficientNetV2B0 fine-tuned on our dataset with a batch size of 128 and an initial learning rate of 0.001 for 20 epochs.**

***Index Terms*—Image, AI-generated Images, Deep Learning, Artificial Intelligence, Image Classification**

I. Introduction

Image synthesis is the process of generating artificial images from different input modalities, i.e., text, sketch, audio, or image.It is used in many applications, such as art generation, photo editing, photo inpainting, and computer-aided design.The field of synthetic image generation by Artificial Intelligence (AI) has developed rapidly in recent years,and the ability to detect AI-generated photos has also become a critical necessity to ensure the authenticity of image data.

Image synthesis has received intense research, especially after Generative Adversarial Networks (GAN).It wasn't too long ago for generative technology to produce images with significant visual flaws that the human eye could detect, but these days, AI models could create high-fidelity, photorealistic images in a matter of seconds.This has led to a situation where there is consumer-level technology available that could quite easily be used for the violation of privacy and to commit fraud.Therefore, many researchers have incorporated GANs into image synthesis, which has led to a significant enhancement in generated images.Therefore, many researchers have incorporated GANs into image synthesis, which has led to a significant enhancement in generated images.

Generative imagery that is indistinguishable from photographic data raises questions both ontological, those which concern the nature of being, and epistemological, surrounding the theories of methods, validity, and scope. Ontologically, given that humans cannot tell the difference between images from cameras and those generated by AI models such as an Artificial Neural Network, in terms of digital information, what is real and what is not?This study explores the potential of using computer vision to enhance our newfound inability to recognise the difference between real photographs and those that are AI generated.Latent Diffusion Models (LDMs), a type of generative model, have emerged as a powerful tool to generate synthetic imagery[1].

A dataset, called CIFAKE, is generated with latent diffusion.The CIFAKE dataset provides a contrasting set of real and fake photographs and contains 120,000 images (60,000 images from the existing CIFAR-10 dataset) and 60,000 images generated for this study.Second, by using the CIFAKE dataset for classification, this study suggests a way to enhance our shrinking capability as humans to recognize AI-generated images through computer vision.These scientific contributions provide important steps forward in addressing the modern challenges posed by rapid developments of modern

technology, and have important implications for ensuring the authenticity and trustworthiness of data.

The remainder of this paper is organized as follows.We briefly introduce the literature review in Section 2. The background study are presented in Section 3. We conduct the experiments and discussed about the dataset in Section 4,5 and observe the results in section 6.The limitations of our approach in Section 7. Finally, the paper is concluded, and the future work is introduced in Section 8.

II. Literature Review

Misinformation and fake news is a significant modern problem, and machine-generated images could be used to manipulate public opinion.Situations where synthetic imagery is used in fake news can promote its false credibility and have serious consequences. it was observed that synthetically generated signatures could overcome signature verification systems with ease.Image classification is a fundamental computer vision task in which images get assigned specified labels or categories based on their visual content. Through this process, objects, scenes, or patterns within images can be recognized and categorized by machines. Image classification has widespread applications, ranging from facial recognition and autonomous vehicles to medical image analysis and content filtering.Convolutional Neural Networks (CNNs) are a class of deep learning models designed specifically for image-related tasks.Data augmentation is a pivotal technique for improving model generalization. By applying random transformations to training data, models become more robust to variations, particularly in diverse datasets that encompass real and AI-generated images.

III. Background Study

A. Distinguishing Real and Synthetic Imagery

Recent technological developments have made it possible to produce images of such high quality that people are unable to distinguish between photographs taken with real cameras and images that are simply the weights and biases of an artificial neural network hallucinating them.Research indicates that artificially generated human faces have the potential to be used in false acceptance attacks and to obtain unauthorised access to digital systems, making cybersecurity another major concern.

B. Convolutional Neural Network

Deep Learning has proved to be a very powerful tool because of its ability to handle large amounts of data.The interest to use hidden layers has surpassed traditional techniques, especially in pattern recognition. One of the most popular deep neural networks is Convolutional Neural Networks (also known as CNN or ConvNet) in deep learning, especially when it comes to Computer Vision applications.

C. ResNet50V2

The process of fine-tuning a pre-trained neural network model—usually trained on a large dataset for a general task—into a smaller, domain-specific dataset is known as CNN (Convolutional Neural Network) fine-tuning. The goal is to apply the pre-trained model's knowledge and feature representations to a larger dataset and modify it for a particular, frequently related task.Residual Network 50 Version 2, is a specific variant of the ResNet architecture, which is a type of Convolutional Neural Network (CNN).

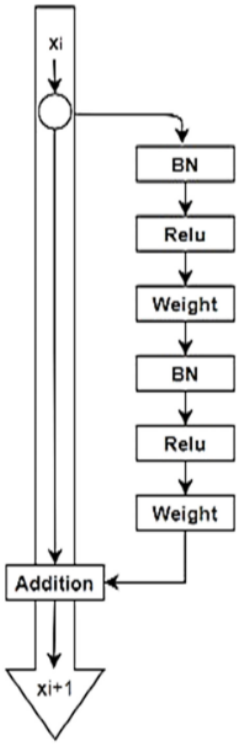

Fig. 1. ResNet50v2 architecture

D. EfficientNetV2BO

EfficientNetV2BO is a variant of the EfficientNet family, a series of convolutional neural network (CNN) architectures designed to achieve high performance with a relatively low number of parameters. Developed as an extension of the original EfficientNet models, EfficientNetV2BO specifically focuses on efficiency in terms of computational resources while maintaining competitive accuracy in various computer vision tasks. For tasks like object detection, image segmentation, and image classification, EfficientNetV2BO strikes a compromise between model size and performance. This is especially useful when memory and computational resource limitations are an issue.

E. EfficientNetB7

EfficientNetB7 is a specific variant within the EfficientNet family, which comprises a series of convolutional neural

network (CNN) architectures designed for optimal performance and efficiency.EfficientNet models, including B7, are characterized by a unique compound scaling method that uniformly scales the network's dimensions (width, depth, and resolution) to achieve an optimal balance between model size and accuracy.

IV. Datasets

The CIFAR-10 dataset consists of 60, 000, 32×32 RGB images of real subjects.The classes within the dataset are airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. There are 50,0 images per class. For each class 100,000 images are used for training and 20000 for testing, divided into two classes.

V. Methodology

CNN model is used for binary image classification.The Sequential model is used to linearly stack the layers.Conv2D applies 32 filters of size 3x3 using the ReLU activation function.MaxPooling2D Performs max pooling to reduce spatial dimensions.Fully Connected Layer Consists of 64 neurons with the ReLU activation function.The Adamax optimizer with a learning rate of 0.001 is used to compile the model. For problems with binary classification, the loss function is binary crossentropy.

The ResNet50V2 model is loaded with pre-trained weights from ImageNet data.A Sequential model is defined with the pre-trained base model as the first layer.Additional layers are added on top, including a dense layer with 256 neurons and ReLU activation, a dropout layer with a dropout rate of 0.4 to reduce overfitting, and a final dense layer with a single neuron and sigmoid activation for binary classification.

EfficientNetV2B0 Utilizes the EfficientNetV2B0 architecture, known for its efficiency and performance.It uses global max pooling.

A Sequential model is defined with the pre-trained base model b7 as the first layer of EfficientNetB7. Additional layers are added on top, including a dense layer with 256 neurons and ReLU activation, a dropout layer with a dropout rate of 0.4 to reduce overfitting, and a final dense layer with a single neuron and sigmoid activation for binary classification.

VI. Results and Observation

Convolutional neural networks (CNNs) were utilized in our image classification experiment, and the model performed admirably on the test set. The test set's accuracy of 92.04% showed that the model could correctly classify images. With a precision of 88.29%, the model demonstrated a low false-positive rate, which is a measure of its accuracy in predicting positive instances. Additionally, the recall—a metric that measures the model's capacity to record every positive instance—reached a high of 96.93%, demonstrating the model's efficacy in recognizing pertinent patterns. Additionally, the model's ability to distinguish between positive and negative instances was measured by computing the Area Under the Curve (AUC), which came out to be 0.51.

Our model performed exceptionally well on the test data when it came to fine-tuning ResNet50V2 for image classification. 96.25% accuracy was attained, demonstrating a high percentage of accurate classifications. Precision, a measure of how well positive predictions worked out, was 95.50%, which is a low false-positive rate. Moreover, the recall, which indicates how well the model captures positive examples, reached a noteworthy 97.09%.Despite these achievements, the Area Under the Curve (AUC) value was calculated to be 0.49. For EfficientNetV2B0 classifier,the accuracy came to 97.51%.Recall and precision both achieved remarkable results of 97.90% and 97.14%, respectively. In spite of these achievements, the AUC (Area Under the Curve) value was 0.50.On the otherhand,for EfficientNetB7 the accuracy is nearly same to EfficientNetV2B0 which is 97.50%,precision 97.14% and recall 97.89%.

VII. Future Work

Future work could involve exploring other technique for classification of the dataset provided.. For example, the implementation of attention-based approaches are a promising new field that could provide increased ability and an alternative method of explainable AI.Furthermore, with even further improvements to synthetic imagery in the future, it is important to consider updating the dataset with images generated by these approaches.

VIII. Conclusion

This study has proposed four different method to improve our deteriorating recognition skills for AI-generated images.The results of this study demonstrated that binary classification could be accomplished with an accuracy of about 97.509% and that the synthetic images were of high quality with complex visual attributes.The CIFAKE dataset's release represents a significant contribution.This study laid out ways to enhance human perception by battling fire with fire because the reality of AI producing images that are indistinguishable from real-life photographic images raises fundamental questions about the boundaries of human perception.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695, 2022.