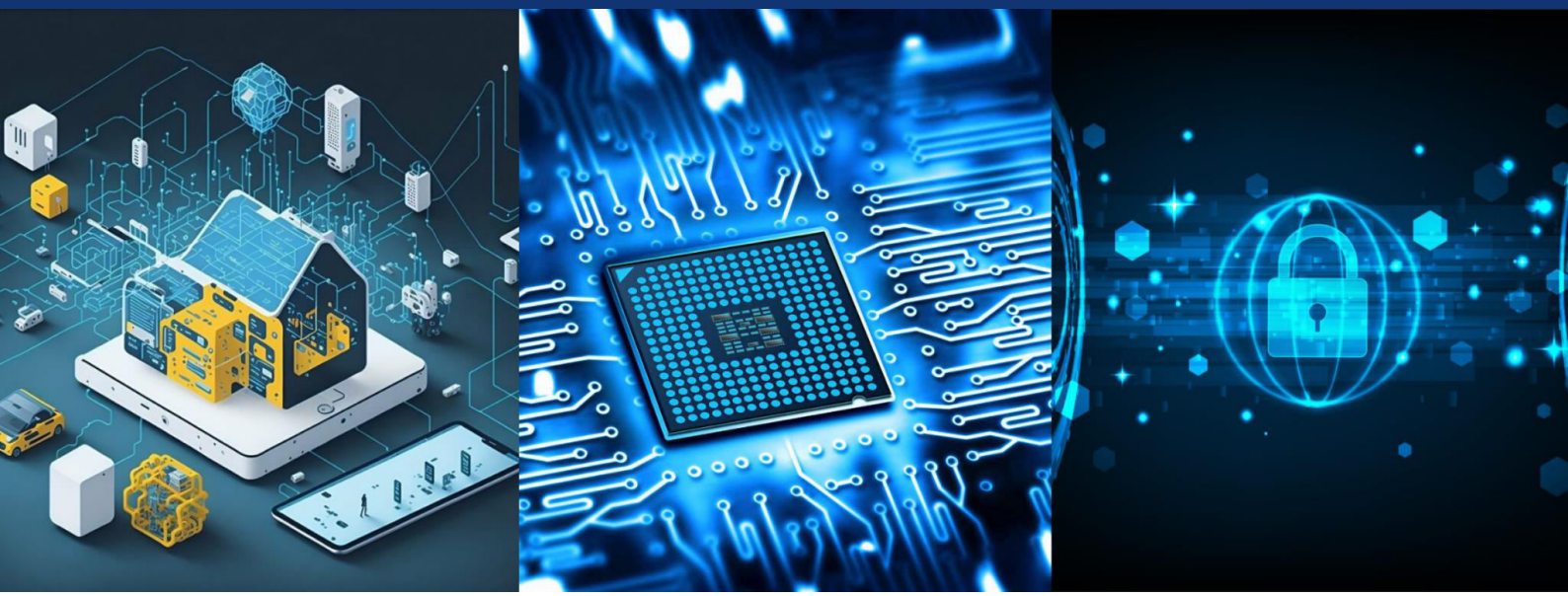
IJIRCCE

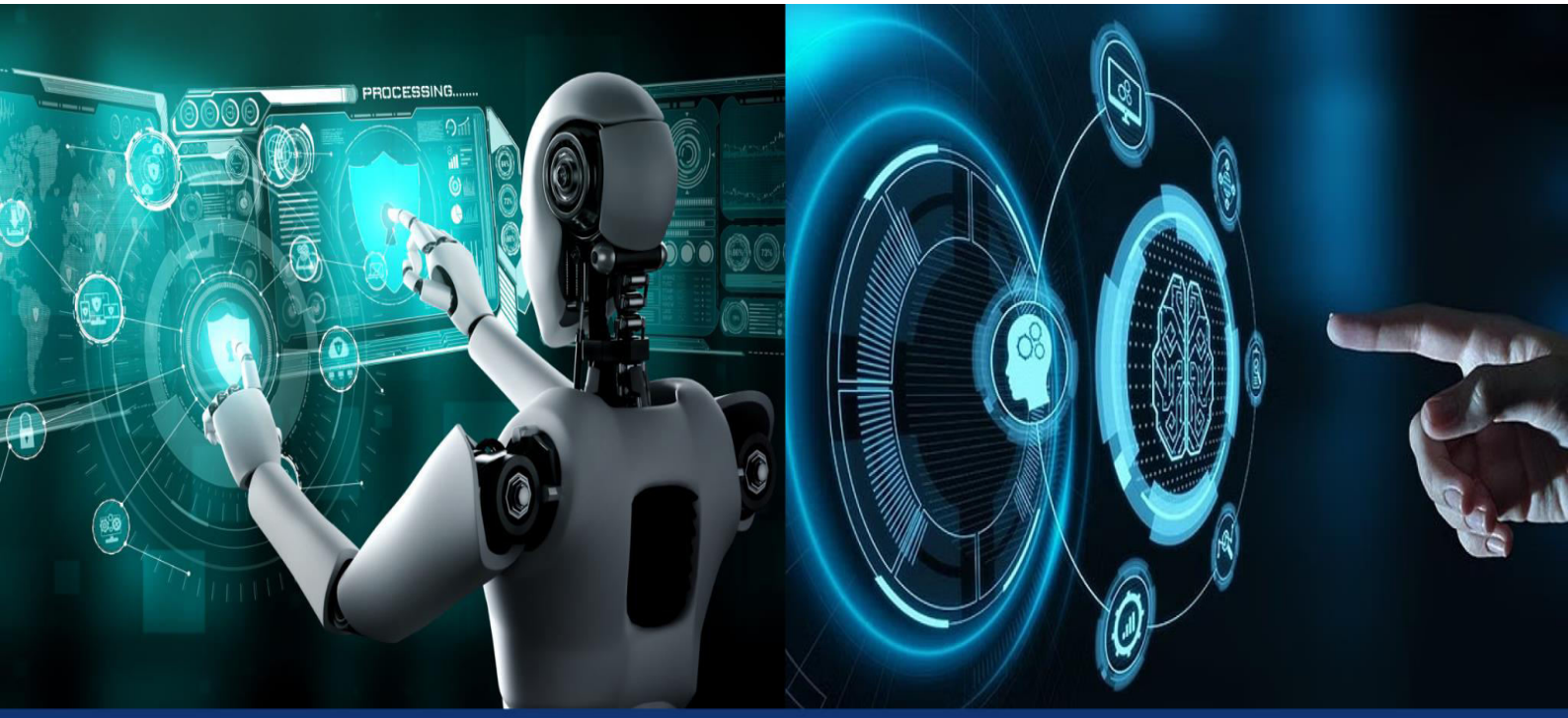
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 5, May 2025

© 2025 IJIRCCE | Volume 13, Issue 5, May 2025| DOI:10.15680/IJIRCCE.2025.1305170

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

AI Generated and Real Image Classification with Transfer Learning and Self-Attention

Dr. S. Lakshmana Pandian

Professor, Department of Computer Science and Engineering, Puducherry Technological University, Puducherry, India

Jones Nivedita. S, Devika Byjou, Gajjala Anusha

Final Year Students, Department of Computer Science and Engineering, Puducherry Technological University, Puducherry, India

ABSTRACT: The rapid growth of generative AI technologies has escalated the possibility of distinguishing between AI-generated images and authentic images, which raises issues like the spread of misinformation. In this research, an efficient deep learning model was developed utilizing EfficientNetV2B0 and incorporated a self-attention mechanism to overcome the lack of long-range spatial dependency capture in traditional CNNs. The model is trained using the ''AI vs. Human-Generated Images'' dataset on Kaggle to ensure thorough evaluation of the model. The model performs exceptionally well on distinguishing real images from fakes due to self-attention's ability to capture subtle global discrepancies like unnaturally aligned object edges, abnormal shadows, strange lighting, and irregular textures that blend with the background, elements that CNNs fail to capture. Furthermore, Attention Rollout enhances the visual transparency of the model's reasoning, thus making the system more interpretable. The system offers strong accuracy - 94.55%, generalization, and reliable solutions for detecting AI-generated pictures and verifying the authenticity of media documents.

KEYWORDS: AI-generated image detection, self-attention mechanism, transfer learning, explainable AI.

I. INTRODUCTION

The sooner availability and sophistication of generative AI technologies have resulted in the rapid generation of synthetic images in almost all fields. As generative models like GANs become more prevalent, the distinction between authentic and machine-created images blurs even more. Although these images are not always photorealistic, their imitation of actual images is often enough to deceive or mislead people, especially when consumed digitally and in a glance. This, in turn, raises concerns about the integrity of media, the risk of misinformation, and the utmost requirement of efficient means to tell apart authentic images from AI autonomously produced ones.

In order to mitigate these issues, automated image classification based on deep learning techniques, particularly with Convolutional Neural Networks (CNNs), have gained popularity. Beyond ImageNet, CNNs have proven to be very fruitful even in general purpose image classification due to their exceptional capabilities in local feature extraction. In prior research, a baseline model built using EfficientNetV2B0 was introduced. It utilized transfer learning to enhance versatility for the binary classification problem of distinguishing AI generated images from real ones. This model's feature tuning enabled it to capture pertinent data—more so, ensemble comparisons provided rivalry analysis beyond other CNN-based models to hone in on design pros and cons.

Regardless of these improvements, approaches based on CNNs are still fundamentally limited. Focus on local features and progressively increasing receptive fields means that broader spatial relationships which may show signs of superficial inconsistencies indicative of generated content can be overlooked. These limitations affect the model's ability to detect more subtle globally distributed cues—subtle gradients, covert object presence positioned in frames, or inconsistent patterns—non-structured photorealistic AI images may portray. In addition, different distributions or data quality might pose a problem for generalizing, which weakens the robustness and scalability of the detection systems within actual scenarios.

To overcome these challenges, our research introduces a modified deep learning framework that enhances the EfficientNetV2B0 architecture with a self-attention mechanism. This integration allows the model to focus on both local and global contextual patterns within an image, improving its ability to capture long-range dependencies and detect spatial anomalies that are difficult to observe using convolution alone. The self-attention mechanism evaluates relationships between different regions of the image, thereby strengthening the model's capacity to distinguish between authentic and synthetic characteristics. In addition, we implement Attention Rollout, a method that offers visual interpretations of the model's decisions by highlighting regions that influenced the output, thus enhancing the transparency and explainability of the detection process.

Our method was assessed using the balanced "AI vs. Human-Generated Images" dataset from Kaggle, as it provides a good benchmark for assessing classification performance on real versus AI-generated images. Each AI-generated image was paired with a human-created image, facilitating comparative learning, which is why the dataset was structured in pairs. The model performed well in terms of accuracy, generalization, and interpretability, surpassing CNN baseline models.

In conclusion, the work described in this paper illustrates a clear deep learning framework based on self-attention CNNs designed to detect AI-generated images, integrating self-attention mechanisms with CNNs, and employing visual explanations to maintain interpretability, thus allowing application to practical media verification issues.

II. RELATED WORK

The widespread emergence of AI-produced images has raised serious challenges to distinguishing synthetic images from actual photographs, and new models that can be strong and explainable must be developed for classification. It has always made use of Convolutional Neural Networks (CNNs) in recent studies as a building block for image classification because of the success of local feature extraction in CNNs. Bird and Lotfi [1] presented the CIFAKE dataset and an interpretable CNN architecture to classify fake images, stressing model transparency. For the purpose of addressing generator diversity, Fattah Saskoro et al. [2] introduced a Gated Expert CNN that utilizes several subnetworks, which improves robustness with different sources of images. Hossain et al. [3] have shown that the integration of CNNs with Vision Transformers (ViTs) can greatly enhance performance by learning both local and global dependencies, complemented by Grad-CAM-based interpretability. Hybrid methods have also been investigated, including the combination of classical characteristics with deep learning models by Taspinar and Cinar [4], and frequency domain analysis by Poredi et al. [5], which emphasized subtle spectral inconsistencies frequently overlooked in the spatial domain.

Other research provided novel contributions, including Kumar et al. [6], who suggested a baseline CNN for comparison, and Yoo Jeong Ha et al. [7], who investigated the gap between human and machine understanding of AI-generated images. Xi et al. [8] proposed a cross-attention-based dual-stream architecture that handles spatial and frequency features in parallel for better accuracy. Epstein et al. [9] concentrated on real-time detection using lightweight CNNs that are appropriate for use on large-scale web platforms.

Together, these works highlight the need for model explainability, generalization across various generators, and computational efficiency. Finally, the literature under review concurs with applying sophisticated techniques such as transfer learning, attention mechanisms, and explainable AI to enhance classification accuracy and explainability. This project then continues in that line by utilizing EfficientNetV2B0 for transfer learning, incorporating self-attention layers to learn global patterns, utilizing data augmentation for improved generalization, and utilizing Attention Rollout for interpretability. Together, these methods tackle important challenges in classifying real vs. AI-generated images using a scalable and interpretable deep learning system.

© 2025 IJIRCCE | Volume 13, Issue 5, May 2025| DOI:10.15680/IJIRCCE.2025.1305170

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

III. PROPOSED ALGORITHM

A. IMAGE CLASSIFICATION

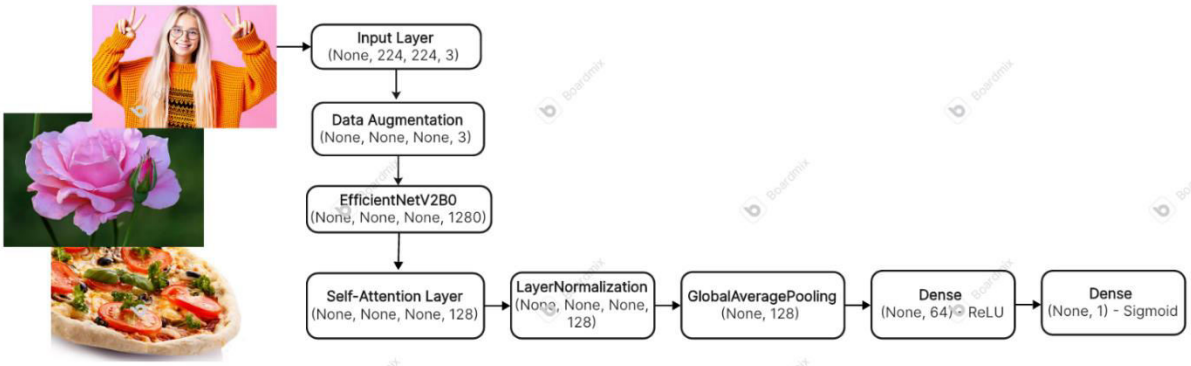


Figure 1: Model Architecture

The proposed architecture makes use of a deep learning model based on the EfficientNetV2B0 backbone, supported by data augmentation, self-attention mechanism, and dense layers for final classification. All aspects of the architecture support better learning, generalization, and explainability of AI-generated image detection.

First, input images are run through a series of data augmentation operations, consisting of horizontal flip, rotation, zoom, and changes in height and width. These augmentations are done during training to enhance the capability of the model to generalize across images having different distributions and to prevent overfitting. Input image sizes are normalized to 224 × 224 × 3 prior to going through the model.

The core of the network is EfficientNetV2B0, a convolutional neural network that is pre-trained on ImageNet and is famous for its effective scaling of depth, width, and resolution with compound coefficients. The model is an effective feature extractor, extracting local spatial patterns from the images. In the early training stage, the pre-learned features are frozen in the base model so that only the top layers are trained to adjust to the particular task of classification. In the subsequent stage, fine-tuning is done by unfreezing the last several layers to enable the model to learn task-specific patterns.

To take care of the drawback of CNNs in capturing long-range dependencies and global information, a proprietary self-attention mechanism is added subsequent to the convolutional feature extraction. The self-attention layer is formulated to improve the model to pick up on fine-grained global inconsistencies frequently discovered in AI-produced images, like ill-aligned object boundaries, artificial textures, or erratic lighting. In this process, input features are initially normalized and mapped into three vectors: queries (Q), keys (K), and values (V) via learned linear transformations. Q (Query) represents the information that each pixel or feature seeks from other parts of the image. K (Key) represents the information available from each pixel or feature to be matched with the query. V (Value) contains the actual content that will be passed along after attention is applied based on the query-key match.
The attention scores are calculated using the scaled dot-product attention formula:

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d^k}}\right)V$$

where d^k represents the dimensionality of the key vectors. Such scores are what indicate the relevance of every feature relative to others in the image, allowing the network to pay attention to informative areas across spatial dimensions. Global contextual features are highlighted by attention-weighted output, which is essential in distinguishing between real and AI-generated content.

© 2025 IJIRCCE | Volume 13, Issue 5, May 2025| DOI:10.15680/IJIRCCE.2025.1305170

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

After the attention module, a Global Average Pooling layer is utilized to lower the spatial size of the feature maps and pool information globally. It is then accompanied by a fully connected layer with 64 units and ReLU activation, which injects non-linearity and maps the high-level features into a low-dimensional space. A dropout rate of 0.6 is used to avoid overfitting by randomly disabling neurons during training. The final output layer has one neuron with a sigmoid activation function that gives a probability score as to whether the image is real or generated by AI. The model is trained using the binary cross-entropy loss function and optimized using the Adam optimizer with varying learning rates in the training and fine-tuning phases.

Through the integration of convolutional feature extraction, global context modeling through self-attention, and efficient training methods, the introduced architecture attains competitive classification performance. In experimental testing, the model reports a classification accuracy of 94.55% on Kaggle's "AI vs. Human-Generated Images" dataset. In addition, Attention Rollout is employed to visualize attention weights, making the model's predictions more interpretable and imparting explanation transparency in decision-making.
This architecture, shown in Figure 1, therefore is a strong, interpretable, and precise solution to the problem of AI-generated image detection. The model performance is evaluated using several key metrics, including accuracy, precision, recall, F1-score, and the confusion matrix. These metrics offer a comprehensive understanding of the model's ability to distinguish between AI-generated and human-generated images.
Accuracy measures the proportion of correctly predicted images out of the total predictions.

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

Precision indicates the proportion of true AI-generated images among those predicted as AI-generated.

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

Recall measures the proportion of correctly identified AI-generated images out of all actual AI-generated images.

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

F1-Score represents the harmonic mean of precision and recall, providing a balanced metric for imbalanced datasets.

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

Confusion Matrix summarizes prediction results with true positives, false positives, true negatives, and false negatives.

$$\begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix}$$

B. EXPLAINABLE AI

The Attention Rollout process is a method created to understand the internal reasoning of attention-based models by seeing how information passes through attention layers. It is especially helpful in deep learning models with self-attention, as it identifies the relative weightage of various input regions that add up to the final prediction. In image classification, this can be used to understand which regions of the image the model depends on the most while making the decision.

In the case of single self-attention layer models, such as the one applied in the proposed model, this is a more simplified process. The self-attention process produces an attention matrix A from the scaled dot-product operation:

$$A = Softmax\left(\frac{QK^T}{\sqrt{d^k}}\right)V$$

© 2025 IJIRCCE | Volume 13, Issue 5, May 2025| DOI:10.15680/IJIRCCE.2025.1305170

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

where Q and K are the query and key matrices, and d^k is the attention space dimensionality. The matrix $A \epsilon \mathbb{R}^{n \times n}$ shows how strong the attention of one token (image patch) gives to each other token in the image. As there is just a single layer of attention, the rollout attention R is the same as the base attention matrix

$$\mathbb{R} = A$$

This matrix is then utilized to calculate attention-based visual explanations by aggregating or averaging attention scores over image regions. Each entry in the output rollout map represents the amount of influence one region exerts over another during the model's prediction process.

Figure 2: Attention Rollout

The Attention Rollout mechanism's output is a heatmap that indicates the areas of the input image the model most actively attended to while generating its prediction. The brighter the area in the heatmap, the greater the attention weights, and they indicate where in the image the most contribution came for the final choice.

This visualization enables researchers and practitioners to understand which parts of space within the image contributed most toward the classification result. This not only enhances transparency but also confers a dimension of trustworthiness and explainability to the AI-produced image detection model.

C. EXPERIMENTAL HARDWARE AND SOFTWARE

The proposed work is developed using Python due to its rich ecosystem of libraries for deep learning and data science. TensorFlow will serve as the core deep learning framework, offering extensive support for building and deploying neural networks. Keras, integrated within TensorFlow, will be used as the high-level API to simplify model construction and training processes. All experiments will be conducted on Kaggle Notebooks, which provide a cloud-based environment with integrated hardware acceleration.

For hardware, the project leverages the NVIDIA Tesla T4 GPU provided by Kaggle for enhanced computational performance. The Tesla T4 is built on the NVIDIA Turing architecture and is optimized for both training and inference workloads. It features 2,560 CUDA cores, 16 GB of GDDR6 memory, and supports mixed-precision computing using Tensor Cores, significantly accelerating deep learning operations. This GPU enables efficient handling of large datasets and complex models, substantially reducing training time and improving model throughput.

IV. SIMULATION RESULTS

The model proposed, which integrates transfer learning and a self-attention mechanism, was trained in stages: a first stage with the base layers' weights frozen and early stopping strategy, and a second stage where all layers apart from the last 10 of the base model were unfrozen and trained with a decrease in the learning rate. This method allowed the model to maintain the acquired low-level features while fine-tuning high-level representations tailored for the task.The incorporation of the self-attention mechanism greatly improved the capability of the model to learn global contextual relationships and subtle visual cues contained within the images, which were commonly neglected by conventional convolutional networks. This enhancement was of utmost importance in improving the overall classification performance of the model. Following comprehensive assessment on the test set, the model performed a remarkable 94.55% accuracy and 0.2114 loss, registering strong predictive power and good generalization across unseen data.

© 2025 IJIRCCE | Volume 13, Issue 5, May 2025| DOI:10.15680/IJIRCCE.2025.1305170

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The model was also characterized by robust and well-balanced class-wise performance across evaluation metrics. A precision of 94.59% reflects the model's reliability in making correct predictions of AI-generated as well as real images without high false positive rates. A 94.51% recall emphasizes its ability to identify the majority of the applicable instances without failing to recognize a large number of actual positives. The F1 score, which is the harmonic mean of precision and recall, was at 94.55%, indicating the overall efficiency and accuracy of the model in correctly classifying both categories with little bias. All these metrics combined confirm the model's appropriateness for real-world use cases encompassing AI-generated image recognition.

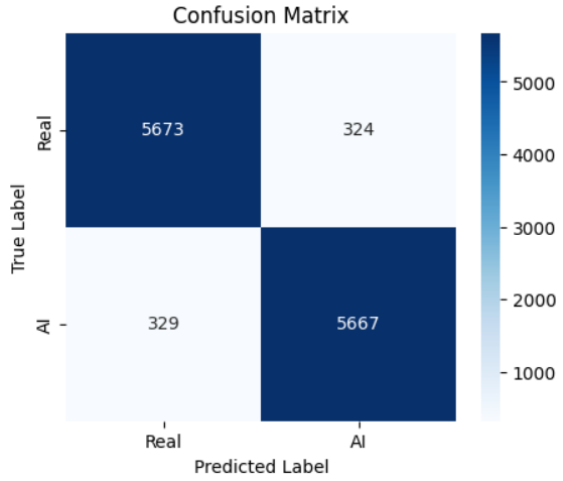


Figure 4: Confusion Matrix

V. CONCLUSION AND FUTURE WORK

In conclusion, this research presents an effective deep learning approach for detecting AI-generated images by integrating a self-attention mechanism with a pre-trained EfficientNetV2B0 model. The model achieves high accuracy and interpretability, making it suitable for real-world media verification tasks.

Future enhancements may include refining the self-attention mechanism through the use of multi-head attention to better capture various features across the image. Incorporating more diverse datasets and unsupervised learning strategies could further improve generalization and robustness. Additionally, optimizing the model for edge deployment by reducing computational complexity without compromising accuracy would enable practical applications in resource-constrained environments like mobile or embedded systems.

REFERENCES

1. J. J. Bird and A. Lotfi, "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images," in IEEE Access, vol. 12, pp. 15642-15650, 2024, doi: 10.1109/ACCESS.2024.3356122.
2. R. Ahmad Fattah Saskoro, N. Yudistira and T. Noor Fatyanosa, "Detection of AI-Generated Images From Various Generators Using Gated Expert Convolutional Neural Network," in IEEE Access, vol. 12, pp. 147772-147783, 2024, doi: 10.1109/ACCESS.2024.3466614.
3. 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.
4. Y. S. Taspinar and I. Cinar, "Distinguishing Between AI Images and Real Images with Hybrid Image Classification Methods," 2024 13th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 2024, pp. 1-4, doi: 10.1109/MECO62516.2024.10577770.

5. N. Poredi, D. Nagothu and Y. Chen, "Authenticating AI-Generated Social Media Images Using Frequency Domain Analysis," 2024 IEEE 21st Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 2024, pp. 534-539, doi: 10.1109/CCNC51664.2024.10454640.
6. U. E. B. Kumar, C. Vivekreddy and K. Sujatha, "A Convolution Neural Network Based Classifier for Discerning AI Generated Images and Real Images," 2024 4th Asian Conference on Innovation in Technology (ASIANCON), Pimari Chinchwad, India, 2024, pp. 1-7, doi: 10.1109/ASIANCON62057.2024.10837986.
7. Yoo Jeong Ha, A., Passananti, J., Bhaskar, R., Shan, S., Southen, R., Zheng, H., & Zhao, B. Y. (2024). Organic or diffused: Can we distinguish human art from AI-generated images? *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*, 4822-4836. https://doi.org/10.1145/3658644.3670306
8. Z. Xi, W. Huang, K. Wei, W. Luo and P. Zheng, "AI-Generated Image Detection using a Cross-Attention Enhanced Dual-Stream Network," 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Taipei, Taiwan, 2023, pp. 1463-1470, doi: 10.1109/APSIPAASC58517.2023.10317126.
9. D. C. Epstein, I. Jain, O. Wang and R. Zhang, "Online Detection of AI-Generated Images," 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Paris, France, 2023, pp. 382-392, doi: 10.1109/ICCVW60793.2023.00045.
10. 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.
11. B. Singh and D. K. Sharma, ‘‘Predicting image credibility in fake news over social media using multi-modal approach,’’ Neural Comput. Appl., vol. 34, no. 24, pp. 21503–21517, Dec. 2022.
12. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, ‘‘Grad-CAM: Visual explanations from deep networks via gradient-based localization,’’ in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 618–626.
13. Sharnish, G., P. Mishra, and A. Kohli. Image Denoising for an Efficient Fake Image Identification. in 2022 International Conference on Edge Computing and Applications (ICECAA). 2022. IEEE.
14. Rana, M.S., et al., Deepfake detection: A systematic literature review. IEEE access, 2022. 10: p. 25494-25513.
15. Bo Liu, Fan Yang, Xiuli Bi, Bin Xiao, Weisheng Li, and Xinbo Gao. Detecting generated images by real images. In European Conference on Computer Vision, pages 95–110. Springer, 2022.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details