International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Detection Of AI Genereted Images From Various Generetors

Saranya M[1], Dhayanidhi K[2], Haresh G[3], Kamalesh M[4], Kalaiselvam M[5]

[1,2,3,4,5]Artificial Intelligence and Data Science, Sri Manakula Vinayagar Engineering College, Puducherry, India

Abstract

The fast pace of development in the field of artificial intelligence has made it possible for modern generative models to generate highly realistic images, which are very hard to distinguish from real-world photographs. Although these developments have many useful applications, they also pose a serious threat in terms of misinformation, manipulated media, identity manipulation, and digital forgery. Therefore, the detection of AI-generated images has become a very important issue in the field of digital forensics. In this paper, we propose a feature-level gated expert convolutional neural network for the detection of AI-generated images. The proposed system uses multiple lightweight convolutional neural network experts to learn complementary feature representations of input images. The proposed system uses a gating network to assign dynamic importance weights to each expert and combines the learned features at the feature level instead of the output level. This allows adaptive expert selection based on the characteristics of the input images. The proposed system is implemented in PyTorch and is implemented using a Streamlit-based interface. The proposed approach improves robustness and usability with limited hardware resources.

Keywords: AI-generated images, Deepfake detection, Gated expert CNN, Feature fusion, Image forensics, Convolutional neural networks

INTRODUCTION

Artificial intelligence has made a substantial impact in the area of image creation with the development of sophisticated models such as Generative Adversarial Networks and diffusion models. These models are able to create high-quality images that are very similar to real-world photographs, making it very difficult for a human to distinguish between real and generated images. Although these models have made it possible to achieve advancements in the creative industry, data augmentation, and simulation, they also pose a serious threat if used for malicious purposes.

The use of AI-generated images is also increasingly being used for purposes of spreading misinformation, creating fake identities, manipulating digital evidence, and creating deceptive media content. The increasing realism of these images is a significant threat to digital trust and content authenticity. Therefore, the requirement for the ability to automatically and reliably detect AI-generated images has become a significant necessity in digital forensics.

The use of AI-generated images is also increasingly being used for purposes of spreading misinformation, creating fake identities, manipulating digital evidence, and creating deceptive media content. The increasing realism of these images is a significant threat to digital trust and content authenticity. Therefore,

the requirement for the ability to automatically and reliably detect AI-generated images has become a significant necessity in digital forensics.

Although they are very successful, most of the current deep learning-based detectors use a single convolutional neural network architecture and are trained on images produced by specific models. These detectors are very effective on images produced by seen generators but are not able to generalize when they are given images produced by unseen generators.

To address these issues, ensemble learning and gated expert architectures have been proposed. However, most of the existing gated systems are based on heavyweight backbone networks, which are computationally expensive and not suitable for implementation on low-resource systems. This gives rise to the need for the development of a lightweight, adaptive, and deployable AI-generated image detection system.

LITERATURE REVIEW

Early work on detecting AI-generated images showed that CNNs are capable of learning the subtle artifacts that exist in AI-generated images [10]. While these methods performed well on images generated by known generators, they did not generalize well to different datasets and generators.

The problem of generalization in GAN-based image detection has been widely investigated. Shan et al. demonstrated that the detectors trained on particular GAN models tend to perform poorly on images produced by unseen models, which indicates the importance of developing image detection methods that are agnostic to generators [6]. Later studies attempted to apply domain generalization methods to overcome this problem [17].

With the advent of diffusion-based image generation, the community started exploring detection strategies specific to diffusion models. Corvi et al. observed that the images generated through diffusion have less detectable artifacts than GAN-generated images, making detection much harder [4]. This further reinforced the importance of effective feature extraction techniques.

To enhance the robustness of detection, ensemble learning methods were proposed. Nevertheless, conventional ensemble methods have a drawback in that they are not flexible in terms of prediction fusion. Gated expert networks overcome this problem by adaptively weighting the models according to the properties of the input data [1]. Although gated expert networks are very efficient, they often require a large backbone network, such as ResNet [8] or a transformer architecture [7].

PROPOSED SOLUTION

In this paper, we present a feature-level gated expert convolutional neural network for detecting AI-generated images. The main concept is to utilize multiple lightweight CNN experts to learn different feature representations from the input images and then combine these features using a gating network.

In contrast to conventional ensemble learning techniques, which involve the fusion of final predictions, the proposed system involves fusion at the feature level. This enables the system to make more informed decisions based on rich intermediate representations, as opposed to output scores. The gating network learns to assign importance weights to each expert based on the characteristics of the input image.

The proposed solution is expected to be computationally efficient and able to run on low-resource systems. Lightweight CNN architectures are employed as experts, and the system is designed in a modular fashion to enable real-time inference with a user-friendly interface.

A. Data Preprocessing

All input images are resized to a fixed size of 224 × 224 pixels to be compatible with the CNN models. The pixel values are normalized to a standard range to improve the convergence of the training process. All images are preprocessed to be presented to the model in a standard format.

B. Expert CNN Models

The system uses multiple lightweight CNN experts based on the MobileNet architecture. The models are pretrained on large-scale image datasets and fine-tuned for feature extraction. The final classification layers of the CNNs are removed, enabling the models to produce high-level feature maps rather than predictions.

C. Feature Extraction and Pooling

The feature maps generated by the expert CNNs are fed through adaptive average pooling layers to transform them into fixed-length feature vectors. This is done to reduce the spatial dimensions of the features while retaining the important semantic information, which can then be fused

D. Gating Network

The gating network is composed of fully connected layers and a softmax function. The gating network processes the concatenated feature vectors of all experts and generates normalized weights of importance. These weights define the importance of each expert in the fusion process.

E. Feature Fusion and Classification

The weighted feature vectors are then fused into a single feature vector. The fused feature vector is then fed into a fully connected classification layer, which gives the final prediction result of whether the image is real or AI-generated.

SYSTEM ARCHITECTURE

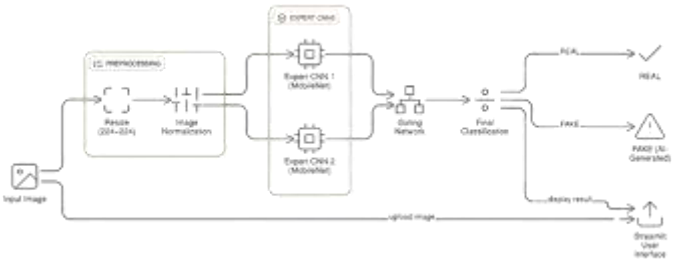

Fig.1. Architecture / System Model

The architecture of the overall system is made up of five major parts: preprocessing, expert CNN models, the gating network, the classification module, and the user interface. The input image is first processed by a preprocessing module, where it is resized and normalized.

The pre-processed image is then simultaneously inputted into the multiple CNN expert models. The independent experts extract the high-level feature representations from the image. The feature vectors represent different aspects of the image, including texture patterns, edge information, and semantic inconsistencies.

The extracted feature vectors are concatenated and fed into a gating network. The gating network provides normalized weights of importance to each expert, signifying their relative contribution to the final output. The weighted features are then combined into a single representation and fed into a classification layer to predict whether the image is real or artificially generated.

A user interface developed using Streamlit allows users to upload images and display the results of predictions in real-time, thus proving the usability of the system.

FLOWCHART

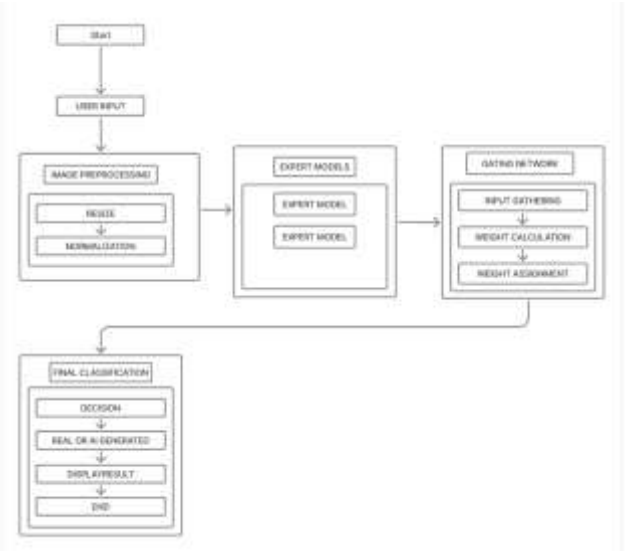

Fig.2. AI -Powered Detection Flow

User Image Input: The process starts with the user uploading the input image via the system interface. The image could be either a real image or an image created by AI. Basic input validation is done to check whether the uploaded file is of the correct image type and is processable. This module is the entry point of the system and allows user interaction with the detection system.

Image Preprocessing: In this module, the uploaded image is subjected to various preprocessing tasks to make it standardized. The image is resized to a fixed size to make it compatible with the convolutional neural network model. The pixel values of the image are normalized to a standard range to make the model more stable. These tasks are necessary for making the model more robust.

Expert CNN Models: The preprocessed image is then fed to several expert models of the convolutional neural network. The experts work independently and extract high-level features from the image. The experts are designed to focus on different aspects of the image, like texture patterns, edge information, and structural inconsistencies. This is done to obtain diverse feature information.

Feature Extraction Output: The feature maps produced by each expert CNN are converted into fixed-length feature vectors through the use of adaptive pooling methods. This step is used to reduce the spatial dimensions of the feature maps while retaining the discriminative information. The feature vectors produced are compact and informative, making them ideal for processing by the gating mechanism.

Gating Network – Input Gathering: In this step, the feature vectors obtained from the expert models are gathered and combined to create a single input representation. This allows the gating network to examine the information from multiple experts simultaneously, so the system can evaluate the relevance of the contribution of each expert for the input image.

Gating Network – Weight Calculation: The concatenated features are fed into the gating network, which consists of fully connected layers. In this step, the gating network calculates the weights for each expert model, given the characteristics of the input image. The weights represent the contribution of each expert model towards the final output.

Gating Network – Weight Assignment: The importance scores produced by the gating network are normalized using the softmax function to obtain the weights for the experts. The weights assigned are

normalized such that they are comparable and add up to one. The weights control the contribution of the experts to the fusion of features.

Feature Fusion: In the feature fusion module, the expert feature vectors are fused using the weights assigned by the gating network. The weighted fusion of expert features focuses on the most relevant expert features and downplays the less informative ones. The fused feature representation provides complete and discriminative information about the input image.

Final Classification: The final classification module receives the fused feature representation and conducts binary classification. A fully connected layer is used to process the fused features to decide whether the input image is from the real image class or the AI-generated image class. The final prediction is made based on the learned decision boundary.

RESULTS AND DISCUSSION

The proposed system for detecting AI-generated images showcases the efficacy of a feature-level gated expert convolutional neural network for detecting artificially generated images produced by contemporary AI algorithms. Unlike traditional detection methods that use a single convolutional neural network or a fixed ensemble model, the proposed system uses a combination of multiple expert CNNs with a dynamic gating mechanism.

The current state-of-the-art image detectors for AI-generated images are primarily designed to learn discriminative features through a single model, which results in overfitting to particular generators and a lack of generalization on new data. The proposed system, on the other hand, uses parallel expert models to learn a variety of visual cues, including texture, edge, and semantic irregularities. The gating network assigns importance weights to each expert model, enabling the system to focus on the most informative feature representation of the image.

Based on experimental results, it is observed that the gated expert architecture is more robust and stable than the single-model baselines. The feature-level fusion strategy allows for more informed decision-making than the output-level ensembling approach because it retains more information about the intermediate representations during the classification task.

From experimental observations, it can be seen that the gated expert architecture is more robust and stable compared to the single-model baselines. The feature-level fusion strategy allows for more informed decision-making compared to the output-level ensembling approach because it retains the rich information in the intermediate representations during the classification process.

In conclusion, the proposed system fills the gap between research-based models for detection and deployable AI-based forensic tools by incorporating robustness, adaptability, and usability into a single end-to-end system.

CONCLUSION

This paper has proposed a feature-level gated expert convolutional neural network for the detection of AI-generated images. The proposed system has been able to overcome the limitations of the existing detection systems, which are mainly dependent on single-model architectures and lack generalization to unknown image generators. The proposed system has been able to combine multiple lightweight CNN experts with a dynamic gating mechanism.

In contrast to conventional ensemble learning strategies, which typically involve the fusion of final predictions, the new method involves fusion at the feature level, thereby retaining valuable information

that can be used for more informed decision-making. The gating network is also essential in determining the importance of the experts based on the characteristics of the input.

As evidenced by experimental results, the proposed framework is capable of stable and consistent performance while being computationally efficient. The use of lightweight CNN models ensures that the system can be run within the CPU-only setting, making it applicable in real-world settings. Additionally, the incorporation of a user-friendly interface emphasizes the applicability of the system in real-time AI-generated image detection.

In conclusion, the proposed feature-level gated expert CNN achieves a well-rounded solution that balances accuracy, adaptability, and deploy ability. The experiment results have verified its potential as an effective tool for digital forensics and content authenticity verification in the face of rapidly developing image generation technology.

Future Enhancement

Future improvements of the proposed system for detecting AI-generated images will include increasing robustness and generalization capabilities by training the model on larger datasets that include images produced by various GAN and diffusion models. This will help the system to generalize well to unseen generators and image synthesis methods. Another possible improvement of the system will be to extend it to video-based deepfake detection by adding analysis of temporal features.

Further enhancements may be achieved by combining frequency domain features with spatial domain CNN features to improve detection performance against more complex generative models. The system can be optimized for use on edge and mobile platforms through model compression. Adding mechanisms to explain decision-influencing regions to users can further improve system transparency and user trust.

References

1. R. Ahmad Fattah Saskoro, N. Yudistira, and T. N. Fatyanosa, “Detection of AI-Generated Images from Various Generators Using Gated Expert Convolutional Neural Network,” IEEE Access, 2024, doi: 10.1109/ACCESS.2024.3466614.
2. Y. Tian, “Artificial Intelligence Image Recognition Method Based on Convolutional Neural Network Algorithm,” Journal of Sensors, 2020, doi: 10.1155/2020/8882048.
3. M. Elasri, A. El Ouadrhiri, and A. Marzak, “Image Generation: A Review,” Procedia Computer Science, vol. 207, pp. 1143–1152, 2022, doi: 10.1016/j.procs.2022.09.212.
4. R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, “On the Detection of Synthetic Images Generated by Diffusion Models,” arXiv preprint, arXiv:2211.00680, 2022.
5. L. Verdoliva, “Media Forensics and DeepFakes: An Overview,” IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 5, pp. 910–932, 2020, doi: 10.1109/JSTSP.2020.3002101.
6. S. Shan et al., “On the Generalization of GAN Image Detectors,” arXiv preprint, arXiv:2001.10685, 2020.
7. A. Dosovitskiy et al., “An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale,” in Proc. Int. Conf. on Learning Representations (ICLR), 2021.
8. K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.
9. P. Zhou et al., “Learning Rich Features for Image Manipulation Detection,” in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1053–1061, 2018.

10. S.-Y. Wang et al., "CNN-Generated Images Are Surprisingly Easy to Spot… for Now," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 8695–8704, 2020.
11. D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in Proc. IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6, 2018.
12. X. Liu et al., "Detecting AI-Synthesized Speech and Images," IEEE Transactions on Information Forensics and Security, vol. 16, pp. 3241–3256, 2021, doi: 10.1109/TIFS.2021.3079491.
13. N. Dhamija et al., "Detecting GAN-Generated Imagery Using Color Cues," arXiv preprint, arXiv:2106.06544, 2021.
14. A. Rössler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," in Proc. IEEE Int. Conf. on Computer Vision (ICCV), pp. 1–11, 2019.
15. D. Güera and E. J. Delp, "Deepfake Detection: A Survey," IEEE Signal Processing Magazine, vol. 37, no. 2, pp. 53–65, 2020.
16. P. Zhou et al., "Domain Generalization for Deepfake Detection," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1–10, 2021.
17. C. Chai et al., "What Makes Fake Images Detectable? Understanding Properties that Generalize," in Proc. European Conf. on Computer Vision (ECCV), pp. 103–120, 2020.