IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DEEPFAKE DETECTION USING TRANSFER LEARNING AND MULTI-TASK CASCADED CONVOLUTIONAL NEURAL NETWORK

Sindhu V
Department of Computer Science and Engineering.
Rajalakshmi Institute of Technology Chennai, India

Pandithurai O
Assistant Professor
Department of Computer Science and Engineering.
Rajalakshmi Institute of Technology Chennai, India

Sasi Devi S
Department of Computer Science and Engineering.
Rajalakshmi Institute of Technology Chennai, India

Abstract— The project focuses on detecting deepfake images using a dataset containing real and fake images. This employs transfer learning, a technique in which a pre-trained model (EfficientNetB0) is used as a starting point and fine-tuned on a new dataset (comprising real and deepfake images). This approach leverages the knowledge learned by the pre-trained model on a large dataset (ImageNet) and adapts it to the specific task of deepfake detection. For face detection and alignment, the MTCNN (Multi-task Cascaded Convolutional Neural Network) algorithm is utilized. MTCNN is a state-of-the-art deep learning model known for its accuracy in detecting faces and facial landmarks. It is used to locate and extract faces from frames of a video, which are then processed and fed into the deepfake detection model. Data augmentation is applied to the dataset using techniques such as rotation, shifting, shearing, zooming, and flipping. This process helps increase the diversity of the training data, improving the model's ability to generalize to unseen images. The Deepfake detection model consists of a base EfficientNetB0 model followed by a Global Average Pooling 2D layer, a Dropout layer for regularization, and a Dense layer with a sigmoid activation function for binary classification (real or fake). The model is trained using the Adam optimizer with a varied range of learning rates and binary cross-entropy loss. During training, the model's performance is monitored using various metrics such as accuracy, precision, recall, and F1-score. Early stopping is employed to prevent overfitting, and the best model is saved using Model Checkpoint to ensure that the model with the lowest validation loss is retained. In conclusion, the project integrates cutting-edge technologies such as transfer learning, deep learning models like EfficientNetB0, and advanced face detection algorithms like MTCNN to develop a robust deepfake detection system.

List Terms— Pre-Trained Model, MobileNetV3, MTCNN, Pooling, Dropout, Dense Layer, Sigmoid activation, regularization, F1-score, Model Checkpoint, cross-entropy loss, Adam optimizer.

I. INTRODUCTION

The advent of deepfake technology has sparked widespread concern due to its potential for misuse, particularly in creating convincingly realistic fake videos for malicious purposes. As deepfake algorithms become more sophisticated and accessible, the threat of misinformation and manipulation in various sectors, including politics, entertainment, and social media, has escalated. Detecting deepfakes has emerged as a critical countermeasure to safeguard against the spread of misinformation and protect the integrity of digital content. In this project, our primary objective is to develop a deepfake detection system using machine learning techniques. By leveraging advanced algorithms and deep learning models, we aim to create a system capable of accurately identifying manipulated or synthetic media content. Through this endeavor, we seek to contribute to the ongoing efforts to combat the proliferation of deepfake technology and its potential impact on society. Our approach involves collecting a diverse dataset

comprising both real and fake media content. We will preprocess and augment the dataset to enhance its diversity and robustness. The core of our detection system will be a convolutional neural network (CNN), a class of deep learning models known for their effectiveness in image classification tasks. We will train CNN using the augmented dataset, with a focus on optimizing its performance in distinguishing between real and fake media. The developed deepfake detection system has the potential to significantly impact various sectors, including journalism, cybersecurity, and digital forensics. By providing a reliable means of detecting deepfakes, our system aims to empower individuals and organizations to identify and mitigate the risks associated with manipulated media content.

II. LITERATURE SURVEY

In [1] The rapid advancement in generative deep learning technologies, particularly Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), has led to the proliferation of deepfakes - highly realistic and manipulated multimedia content. While these forgeries can range from benign entertainment to harmful misinformation, their potential for misuse in spreading falsehoods and harassing individuals is undeniable. This paper explores the complexities of Deepfake detection amidst the evolving landscape of digital media manipulation. With a focus on the inadequacies of current state-of-the-art image classification models, such as Efficient Net, ResNeXt, and XceptionNet, in accurately identifying Deepfake content due to issues like high model overfitting, this study delves into the challenges posed by the sophistication of Deepfake generators. Highlighting the significance of diverse and comprehensive datasets for training, such as UADFV, Face Forensics++, Celeb-DF, Google DFD, and the DFDC dataset, the paper underscores the problem of neural network overfitting when faced with limited variation in manipulated faces. To address these challenges, we propose innovative data augmentation techniques, namely improved Random Erasing, and the Face Cutout method, which intelligently targets and modifies facial features within the training data to enhance model robustness against unseen Deepfakes. Our experimental results demonstrate the efficacy of these approaches in improving Deepfake detection capabilities. This paper not only sheds light on the current state of Deepfake detection but also proposes practical solutions to enhance the accuracy and reliability of these systems, paving the way for future research in the field.

In [2] The surge in digitally altered media, especially deepfake videos, presents a formidable challenge in identifying authentic versus doctored content. This study explores deepfake detection through cutting-edge Convolutional Neural Networks (CNN), specifically EfficientNet-B4 and XceptionNet, utilizing the FF++ and Celeb-DF (v2) datasets for analysis. Our approach includes preprocessing the Celeb-DF dataset to extract and focus on facial frames, training the models and assessing their effectiveness through metrics like log loss and the Area Under the Curve (AUC). Results indicate the high efficacy of both architectures in distinguishing between genuine and manipulated videos, underscoring the necessity for continuous updates in detection methodologies to keep pace with the advancing deepfake generation technologies.

In [3] This study introduces a cutting-edge detection model, termed ReLU-Swish Efficient Net (RSE-Net), designed to identify deepfake content with high accuracy across various generation methods. RSE-Net innovates by integrating an ensemble of Efficient Net architectures, enhanced through a fusion strategy for superior deepfake detection capabilities. A key modification involves substituting the standard Swish activation with ReLU within the conv2D layers of the initial EfficientNetB0 model, optimizing computational efficiency and minimizing overfitting risks. Performance evaluation on prominent deepfake datasets, Face Forensics++ and CelebDF, demonstrates RSE-Net's exceptional accuracy (99.7% on Face Forensics++, 96.09% on CelebDF) and its robustness in real-world applications, making it an effective solution for discerning manipulated media content.

In [4] This study introduces a novel and efficient approach for detecting fake facial videos, addressing the challenges of accuracy, robustness, and high parameter counts faced by current models. It features a refined EfficientNetV2-S architecture for core feature extraction, coupled with the MTCNN algorithm for precise facial detection and alignment. The model also incorporates advanced data augmentation to boost its robustness. By streamlining the backbone network and optimizing its structure, the model significantly reduces complexity and parameter dependency. The addition of the CBAM attention module enhances the model's ability to discern detailed facial features. Testing on Face Forensics++, Celeb-DF-v1, and DFDC datasets show a notable reduction in parameters by 50% and accuracy gains of 1.1%, 2.8%, and 3.1%, respectively, outperforming existing models like Xception, CapsuleNet, and DefakeHop to various extents.

In [5] The rapid advancement in deepfake technology has presented significant security risks. While numerous detection techniques have been developed, they often struggle to identify deepfakes created with unfamiliar methods. To address this, some researchers have created their own datasets, though these tend to capture artifacts specific to certain face-blending techniques, limiting their applicability. This study introduces the Cluster Decision Network (CDNet), a novel approach aimed at enhancing the general detection capabilities of deepfake identification systems. CDNet incorporates a selective attention mechanism focusing on key facial features such as the eyes, nose, and mouth, significantly reducing the model's size. Additionally, it employs a cluster classifier inspired by contrastive learning to make full use of the feature representation. Our comprehensive testing demonstrates that CDNet surpasses current leading solutions in detecting a broad range of deepfakes while maintaining the smallest model size.

III. METHODOLOGIES

The proposed system includes several methods as follows:

3.1. DATA COLLECTION:

The dataset used in this project is a critical component of our deepfake detection system. It consists of a carefully curated collection of images, including authentic real images and various types of fake images generated using deepfake and face2face techniques. This diverse dataset is essential for training our model to distinguish between real and manipulated images effectively.

3.2. DATA PREPROCESSING:

Before training our model, we preprocess the dataset to ensure consistency and enhance the model's ability to generalize. This includes resizing all images to a standard size, such as 224x224 pixels, to maintain uniformity. Additionally, we apply data augmentation techniques such as random rotations, shifts, flips, and zooms. These techniques help expose the model to a broader range of image variations, improving its robustness.

3.3.MODEL ARCHITECTURE:

For our deepfake detection system, we chose the EfficientNetB0 architecture due to its proven performance and efficiency. Efficient Net models are known for achieving state-of-the-art results with fewer parameters, making them ideal for our resource-constrained environment. We leverage transfer learning by using a pre-trained EfficientNetB0 model originally trained on the ImageNet dataset. By fine-tuning only, the top layers of the model, we can adapt it to our specific deepfake detection task while retaining the valuable features learned from ImageNet.

3.4.TRAINING:

During the training phase, we feed batches of augmented images into the model and adjust its weights to minimize the binary cross-entropy loss function. We use the Adam optimizer for efficient weight updates, and we monitor the model's performance using a validation set. Early stopping is employed to prevent overfitting and ensure that the model generalizes well to unseen data.

3.5.TESTING:

The face extraction module uses the MTCNN algorithm to detect and extract faces from images or video frames. MTCNN is a state-of-the-art deep learning model known for its accuracy in detecting faces and facial landmarks. It identifies the location of faces in an image, extracts the face region, and processes it further for analysis. By isolating the face region, the module reduces noise and focuses on the essential features necessary for deepfake detection, improving the overall accuracy of the system.

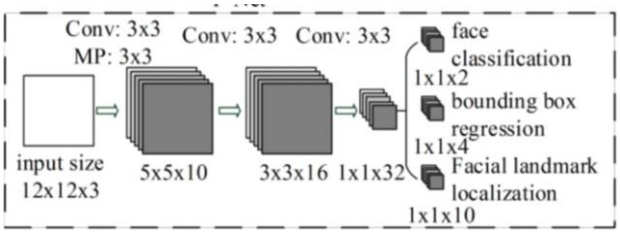
3.6.EVALUATION:

To evaluate the performance of our trained model, we use a separate test set that was not seen during training. We calculate various evaluation metrics, including accuracy, precision, recall, and F1-score, to assess the model's ability to correctly classify real and fake images. These metrics provide valuable insights into the model's performance and help us identify areas for improvement.

IV. ARCHITECTURE

4.1.INPUT:

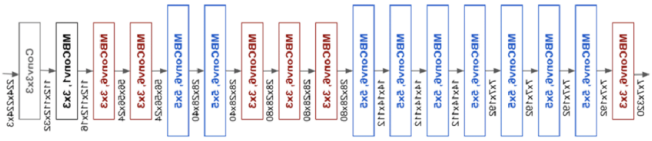
The input to the model consists of images resized to 224x224 pixels. This standard size ensures compatibility with the EfficientNetB0 architecture and allows for efficient processing of image data.



P-NET MTCNN

4.2.BASE MODEL:

The EfficientNetB0 serves as the base model for our deepfake detection system. This model was chosen for its efficiency and effectiveness in image classification tasks. By leveraging the pre-trained weights from the ImageNet dataset, the model already possesses knowledge about various features present in images, which can be beneficial for our task.



4.3.LAYERS:

After the base model, we add a GlobalAveragePooling2D layer to reduce the spatial dimensions of the feature maps and extract global features from them. This is followed by a Dropout layer with a dropout rate of 0.2 to prevent overfitting by randomly setting a fraction of input units to zero during training. Finally, we add a Dense layer with a sigmoid activation function for binary classification. This layer outputs a single value between 0 and 1, representing the probability that the input image is a deepfake.

4.4.OPTIMIZER:

We use the Adam optimizer for training our model. Adam is well-suited for tasks like deep learning due to its adaptive learning rate and momentum properties, which help converge faster and more efficiently.

4.5.LOSS FUNCTION:

The binary cross-entropy loss function is used for this binary classification task. It measures the difference between the predicted probabilities and the actual labels, penalizing incorrect predictions more heavily.

4.6.METRICS:

The primary metric to evaluate the model's performance is accuracy, which measures the proportion of correctly classified images out of the total number of images. This metric provides a straightforward indication of how well the model distinguishes between real and fake images.

V. FUTURE SCOPE

5.1.ENHANCED MODEL:

In the future, we can potentially explore using larger Efficient Net models, such as EfficientNetB1, B2, or even higher, to improve the model's accuracy. These larger models can capture more intricate patterns and features in images, which could be beneficial for detecting subtle differences between real and fake images.

5.2.VIDEO ANALYSIS:

Extending the model to analyze videos for deepfake detection is a promising direction. This involves processing frames of a video to detect deepfakes, considering the temporal information between frames. Techniques like 3D convolutions or recurrent neural networks (RNNs) could be employed to capture the temporal dependencies in videos, enhancing the model's ability to detect deepfakes in moving content.

5.3.REAL-TIME DETECTION:

Developing a real-time deepfake detection system would be valuable for the immediate identification of fake content. This would require optimizing the model and its deployment to achieve low latency. Techniques like model quantization, which reduces the precision of the model's weights, and model pruning, which removes unnecessary weights, can be explored to make the model more lightweight and suitable for real-time applications. Additionally, deploying the model on edge devices or using cloud-based solutions with low latency can further improve real-time detection capabilities.

VI. CONCLUSION

In conclusion, our deepfake detection system has shown promising results in identifying fake images, thanks to the strategic use of the EfficientNetB0 model and data augmentation techniques. By leveraging transfer learning, we were able to harness the power of a pre-trained model and fine-tune it to detect deepfake images with high accuracy.

However, there is still room for improvement and further research in this field. One area for enhancement is the exploration of more advanced model architectures and techniques. For example, experimenting with larger Efficient Net models or incorporating attention mechanisms could potentially boost the model's performance.

Additionally, considering the evolving nature of deepfake technology, continuous research and development are crucial. Future efforts could focus on developing more robust detection methods that can adapt to emerging deepfake techniques. Collaborative efforts between researchers, industry experts, and policymakers will be essential in addressing the challenges posed by deepfake technology and safeguarding against its potential misuse.

VII. REFERENCES

(1) S. A. Minhas, S. Mushtaq, and A. Javed, “EfficientNetB0 Ensemble Model for Unified Deepfakes Detection," 2023 17th International Conference on Open-Source Systems and Technologies (ICOSST), Lahore, Pakistan, 2023.

(2) O. Ahmed et al., "Deepfake Detection System using Deep Learning," 2023 Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, Egypt, 2023.

(3) B. Yasser et al., "Deepfake Detection Using Efficient Net and XceptionNet," 2023 Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, Egypt, 2023.

(4) Z. Hou, Z. Hua, K. Zhang, and Y. Zhang, "CD Net: Cluster Decision for Deepfake Detection Generalization," 2023 IEEE International Conference on Image Processing (ICIP), Kuala Lumpur, Malaysia, 2023.

(5) R. Yang, D. Xu, and Y. Cheng, "Lightweight detection method for deepfake face video," 2023 8th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), Okinawa, Japan, 2023.

(6) H. Agarwal, A. Singh, and R. D, "Deepfake Detection Using SVM," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2021.

(7])J. K. Lewis et al., "Deepfake Video Detection Based on Spatial, Spectral, and Temporal Inconsistencies Using Multimodal Deep Learning," 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Washington DC, DC, USA, 2020.

(8) R. Tolosana, R. Vera-Rodriguez, J. Fierer, A. Morales, and J. Ortega-Garcia, ‘‘Deepfakes and beyond A survey of face manipulation and fake detection,’ Inf. Fusion, vol. 64, pp. 131–148, Dec. 2020.

(9)L. Guarnera, O. Giudice, and S. Battiato, "Deepfake Detection by Analyzing Convolutional Traces," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 2020.

(10) Web Result Deepfake Face Image Detection Based On Improved VGG Convolutional Neural Network, “2020 39th Chinese Control Conference (CCC), Shenyang, China, 2020.

(11) K. Zhu, B. Wu, and B. Wang, "Deepfake Detection with Clustering-based Embedding Regularization," 2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC), Hong Kong, China, 2020.