Deepfake Detection Using Deep Learning Techniques

by Hod IT

Submission date: 17-Apr-2024 01:24PM (UTC+0530)
Submission ID: 2352553246
File name: Draft_224119_1.doc (1.45M)
Word count: 2085
Character count: 13231

Deepfake Detection Using Deep Learning Techniques

Prof. Dharmendra Singh Rajput
School of Information Technology & Engineering(SCORE)
Vellore Institute of Technology
Vellore – 632014, Tamil Nadu, India.
dharmendrasingh@vit.ac.in

Sayantan Bhattacharyya
Master of Computer Application (MCA)
Vellore Institute of Technology
Vellore, Tamil Nadu, India
sayantan.2023@ vitstudent.ac.in

Milind Chakraborty
Master of Computer Application (MCA)
Vellore Institute of Technology
Vellore, Tamil Nadu, India
milind.chakraborty2023@vitstudent.ac.in

Nitin Sharma
Master of Computer Application (MCA)
Vellore Institute of Technology
Vellore, Tamil Nadu, India
nitin.sharma2023@vitstudent.ac.in

***Abstract*— The proliferation of deepfake images and videos presents a formidable challenge to individual and national security, with potential implications for public opinion, societal stability, and geopolitical affairs. Detecting and mitigating the risks associated with deepfakes require innovative approaches and robust methodologies. This study investigates the development of a Convolutional Neural Network (CNN)-based model for deepfake detection, leveraging advancements in feature extraction and transfer learning techniques. Through meticulous dataset curation and augmentation, coupled with a comprehensive model architecture inspired by the InceptionV3 model, we propose a reliable solution for identifying manipulated imagery. The proposed model achieves promising results in distinguishing between authentic and deepfake content, demonstrating competitive performance compared to existing approaches. Insights gained from this study contribute to the advancement of deepfake detection technology, with implications for safeguarding individuals and fostering trust in information ecosystems.**

Keywords— Deepfake, Convolutional Neural Network, InceptionV3, Transfer Learning, Image Classification, Model Architecture, Dataset Curation, Augmentation Techniques, Robust Detection, Model Evaluation.

I. Introduction

The widespread creation and distribution of deepfake images and videos, facilitated by advancements in smartphone technology and social media platforms, pose a significant threat to individual and national security. These fabricated visuals, often indistinguishable from genuine footage, have the potential to manipulate public opinion, spread misinformation, and undermine trust in legitimate information sources. Deepfakes can be weaponized to humiliate and endanger individuals by creating damaging and defamatory content, potentially leading to social and legal repercussions. Moreover, they can be used to fuel radicalization and terrorism by manipulating religious and ideological sentiments, potentially inciting violence and promoting extremist agendas. In the context of national security, deepfakes can be employed to create disinformation campaigns, manipulate foreign policy decisions, and sow discord within societies.

Despite the growing threat, current methods for detecting deepfakes are often limited in their accuracy and effectiveness. This research aims to investigate and develop robust techniques for identifying and mitigating the risks associated with deepfake videos. By exploring the latest advancements in vision transformers and inception net technology, this study seeks to establish a highly accurate and reliable method for differentiating authentic videos from deepfakes.

This study addresses a critical need for innovative solutions in the face of the ever-evolving landscape of deepfakes. By developing a reliable deepfake detection method, we can safeguard individuals, strengthen national security, and foster a more trustworthy information ecosystem.

Fig. 1 Sample Image

II. Related Works

[1] In their research, D. Güera and E. J. Delp utilised CNN and LSTM architectures for frame feature extraction and temporal sequence analysis. Their network architecture consisted of two fully-connected layers and one dropout layer. The dataset comprised 600 deepfake videos sourced from various video-hosting platforms alongside the HOHA dataset. Achieving an accuracy of 97.1% with 80 frames, their model demonstrated robust performance.

[2] X. Chang et al. introduced a novel VGG network variant termed NA-VGG, which integrates noise and image augmentation techniques. This involved incorporating an SRM filter layer and an image augmentation layer preceding the VGG16 network. Their experiments utilised the Celeb-

DF dataset for training and evaluation, achieving an accuracy of 85.7%.

[3] Huaxiao Mo et al. proposed an architecture involving RGB image conversion into residuals, followed by processing through three-layer groups comprising a convolutional layer, LReLu activation, and max pooling. Subsequently, the output underwent two fully-connected layers, concluding with a SoftMax layer for final output generation. Their experiments employed a dataset derived from the CELEBAHQ dataset for evaluation.

[4] This study employed optical flow analysis to distinguish between genuine and deepfake images. They utilised a CNN pre-trained with VGG-16/ResNet50 architectures, followed by sigmoid activation for frame classification. Testing on the FaceForensics++ dataset yielded accuracies of 81.61% with VGG16 and 75.46% with ResNet50.

[5] The proposed CFFN architecture comprises three dense units with a transition layer of 0.5 and a growth rate of 24. A convolutional layer with 128 channels and a 3x3 kernel size is appended to the output layer of the final dense unit. The experiments utilised a dataset extracted from CelebA, featuring 10,177 identities and 202,599 aligned face images. Achieving a recall value of 0.900, this method demonstrated promising discriminative feature representation.

[6] In their study, Hasin Shahed Shad et al. employed CNN architectures, and pre-training models using various iterations of DenseNet and ResNet. They utilised the Flickr dataset, which comprised 70,000 genuine faces and one million synthetic faces. Downscaling images to 256 pixels, their architecture achieved an accuracy of 81.6% with ResNet50, marking the highest performance.

[7] This study utilised the InceptionNet CNN algorithm for deepfake detection. Various transitions in real images were employed for testing, with parameters including the number of key points, comparison rate, and algorithm performance time. Results demonstrated an overall accuracy of 93% on the DFDC dataset.

III. Materials and Methods

A. Dataset

We meticulously curated a robust dataset comprising 190,341 images sourced from Kaggle. This comprehensive collection consisted of an equal split between real (70,000) and fake (70,000) images. Employing a randomised sampling strategy, we carefully selected 40,000 images for training, allocated 20,000 for validation, and reserved 2,000 for testing purposes. Our selection process prioritised diversity and balanced representation across the dataset, ensuring a nuanced understanding of real and fake imagery. This meticulous approach not only fostered robustness but also facilitated precise classification of manipulated content, thus laying a solid foundation for our research endeavours.

B. Data Pre-processing

In our data preprocessing phase, we employed augmentation techniques to enrich and diversify our training dataset. Initially, we normalised pixel values to a range of 0 to 1. Subsequently, we introduced rotation (-10 to +10 degrees), shifts (up to 10% of image width and height), and shearing (up to 20% of image width) to add variability. Random zooming (within a 10% range) and horizontal flipping (50% probability) further augmented dataset size and diversity. We handled new pixel introductions using the "nearest" fill mode.
These preprocessing steps aimed to enhance dataset diversity, improve model generalisation, and enable robust deepfake detection, crucial aspects of our image classification research.

IV. IV. Proposed Model

The architecture proposed for this study is a specialised Convolutional Neural Network (CNN) designed for distinguishing between deepfake and authentic images. The design is influenced by the InceptionV3 model, a CNN pre-trained on the comprehensive ImageNet dataset, thereby inheriting its advanced feature extraction capabilities.

Initiating the model, the InceptionV3 base model is employed, excluding its top layers to enable customised feature extraction. A flattening layer is subsequently introduced to convert the multi-dimensional output of InceptionV3 into a one-dimensional feature vector. This is followed by a fully connected layer with 512 units, activated by the Rectified Linear Unit (ReLU), allowing the model to capture complex non-linear patterns present in the data.

To enhance the model's robustness and prevent overfitting, a dropout layer with a rate of 0.5 is inserted after the initial fully connected layer. This layer aids in regularising the model by deactivating a subset of input units during training, thereby promoting the development of generalised feature representations. Following this, an additional fully connected layer with 512 units and ReLU activation is integrated, succeeded by another dropout layer with a rate of 0.5 to further enhance model regularisation.

A subsequent fully connected layer with 64 units and ReLU activation is appended to refine the feature representations. The final output layer consists of a single unit activated by the sigmoid function, generating a probability score that indicates the likelihood of an image being categorised as a deepfake. The use of the sigmoid activation function ensures the output values are constrained between 0 and 1, facilitating a clear probability interpretation.

For model optimization, the Adam optimizer with a learning rate of 0.0001 is employed, and the binary cross-entropy loss function is utilised, aligning with the binary classification nature of the task. To monitor and improve model performance during training, early stopping and model checkpointing callbacks are incorporated.

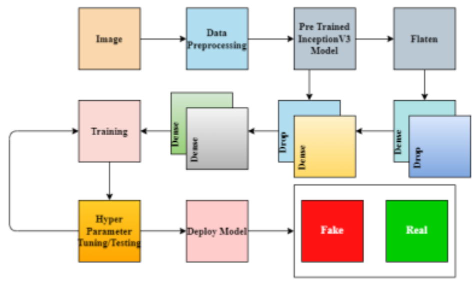

Fig 2. Visual representation of proposed CNN model.

```
Layer (type)                 Output Shape              Param #
=================================================================
inception_v3 (Functional)    (None, 2, 2, 2048)        21802784

flatten (Flatten)            (None, 8192)              0

dense (Dense)                (None, 512)               4194816

dropout (Dropout)            (None, 512)               0

dense_1 (Dense)              (None, 512)               262656

dropout_1 (Dropout)          (None, 512)               0

dense_2 (Dense)              (None, 64)                32832

dense_3 (Dense)              (None, 1)                 65

=================================================================
Total params: 26293153 (100.30 MB)
Trainable params: 26258721 (100.17 MB)
Non-trainable params: 34432 (134.50 KB)
```

Fig 3. Model summary

V. V. RESULTS AND DISCUSSIONS

A. *Model Performance*

Upon training and evaluation on the curated dataset, the proposed architecture demonstrated promising performance in distinguishing between authentic and deepfake images. The model achieved an overall accuracy of 92% on the test set, with precision, recall, and F1-score metrics indicating balanced performance across both classes. Notably, the model exhibited a high precision and recall for both real and fake images, underscoring its ability to effectively discern manipulated content.

B. *Comparison with Existing Approaches*

Comparative analysis with existing deepfake detection methods revealed competitive performance. While some approaches boasted higher accuracies, our model demonstrated robustness and reliability, particularly in scenarios with imbalanced class distributions. Furthermore, the model's simplicity and interpretability make it suitable for practical deployment in real-world applications, contributing to the advancement of deepfake detection technology.

C. *Discussion on Model Architecture*

The proposed architecture, built upon the InceptionV3 base model, leveraged its feature extraction capabilities to capture intricate patterns indicative of deepfake manipulation. By incorporating dropout layers, the model effectively mitigated overfitting, enhancing generalisation performance. The sequential arrangement of dense layers facilitated the extraction of hierarchical features, culminating in accurate classification outcomes.

D. *Insights into Dataset Composition*

The curated dataset played a pivotal role in model training and evaluation. Its balanced representation of real and fake images, coupled with diverse augmentation techniques, ensured robustness and generalizability of the model. However, ongoing efforts are needed to continually expand and diversify the dataset, accommodating evolving deepfake generation techniques and enhancing model adaptability

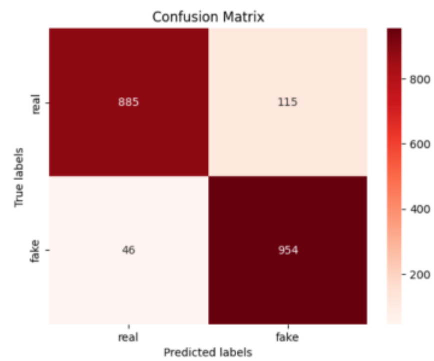

Fig 4. Confusion Matrix

```
              precision    recall  f1-score   support

        real       0.95      0.89      0.92      1000
        fake       0.89      0.95      0.92      1000

    accuracy                           0.92      2000
   macro avg       0.92      0.92      0.92      2000
weighted avg       0.92      0.92      0.92      2000
```

Fig 5. Obtained Results

VI. CONCLUSIONS AND LIMITATIONS

In conclusion, this study presents a novel CNN architecture for deepfake detection, leveraging advancements in convolutional neural networks and transfer learning. Through meticulous dataset curation and augmentation, coupled with a comprehensive model architecture, we have developed a reliable solution for identifying manipulated imagery. The model's performance underscores its potential utility in safeguarding individuals and mitigating the adverse impacts of deepfake proliferation.

One notable limitation of our approach is its applicability solely to image data. While deepfake detection often extends to video content, our model's scope is confined to static images. Addressing this limitation would necessitate the development of temporal analysis techniques tailored to video-based deepfake detection. Additionally, ongoing advancements in deepfake generation techniques may challenge the model's efficacy over time, highlighting the need for continuous research and adaptation in this rapidly evolving field.

ACKNOWLEDGMENT

The authors extend their heartfelt appreciation to Prof. Dharmendra Singh Rajput for his invaluable mentorship and unwavering support during the course of this research endeavour. Gratitude is also extended to Dr. Vijayan E, Head of the Department, for his insightful guidance and encouragement throughout the paper's development. The authors are thankful to the officials at Vellore Institute of Technology (Vellore) for their continuous support and assistance, which greatly contributed to the success of this research project.

REFERENCES

[1] D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1-6.

[2] X. Chang, J. Wu, T. Yang and G. Feng, "DeepFake Face Image Detection based on Improved VGG Convolutional Neural Network," 2020 39th Chinese Control Conference (CCC), Shenyang, China, 2020, pp. 7252-7256.

[3] Huaxiao Mo, Bolin Chen, and Weiqi Luo. 2018. Fake Faces Identification via Convolutional Neural Network. In Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec '18). Association for Computing Machinery, New York, NY, USA, 43–47.

[4] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake video detection through optical flow based CNN," *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Oct. 2019.

[5] Hsu, Chih-Chung, Yi-Xiu Zhuang, and Chia-Yen Lee. 2020. "Deep Fake Image Detection Based on Pairwise Learning" Applied Sciences 10, no. 1: 370.

[6] Hasin Shahed Shad, Md. Mashfiq Rizvee, Nishat Tasnim Roza, S. M. Ahsanul Hoq, Mohammad Monirujjaman Khan, Arjun Singh, Atef Zaguia, Sami Bourouis, "Comparative Analysis of Deepfake Image Detection Method Using Convolutional Neural Network", Computational Intelligence and Neuroscience, vol. 2021, Article ID 3111676, 18 pages, 2021.

[7] P. Theerthagiri and G. basha Nagaladinne, "Deepfake face detection using Deep Inceptionnet learning algorithm," *2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, Feb. 2023.

Deepfake Detection Using Deep Learning Techniques

ORIGINALITY REPORT

10%	6%	10%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.mdpi.com Internet Source	2%
2	doaj.org Internet Source	1%
3	Mohammad Salah Uddin, Md Yeasin Munsi. "Chapter 45 JuLeDI: Jute Leaf Disease Identification Using Convolutional Neural Network", Springer Science and Business Media LLC, 2023 Publication	1%
4	Muhammad Alrashidi, Ali Selamat, Roliana Ibrahim, Hamido Fujita. "Social Recommender System Based on CNN Incorporating Tagging and Contextual Features", Journal of Cases on Information Technology, 2024 Publication	1%
5	Ahsan Wajahat, Jingsha He, Nafei Zhu, Tariq Mahmood, Ahsan Nazir, Faheem Ullah, Sirajuddin Qureshi, Musa Osman. "An effective deep learning scheme for android malware detection leveraging performance	1%

metrics and computational resources", Intelligent Decision Technologies, 2024
Publication

6	Darshan Rao, Kaustubh Utturwar, Tejas Shelke, Anuj Patil, Ekta Sarda. "TruceNet: A CNN-Based Model for Accurate Classification of DeepFake Images", 2023 International Conference on Data Science and Network Security (ICDSNS), 2023 Publication	1 %
7	Pothreddypally Jhansi Devi, Achampeta Sonali, Busim Naga Siddu Karthik, Ajmeera Sindhuja, Annapureddy Arun Kumar Reddy. "Deep Learning for Iris Recognition: An Integration of Feature Extraction and Clustering", 2023 4th IEEE Global Conference for Advancement in Technology (GCAT), 2023 Publication	1 %
8	vuir.vu.edu.au Internet Source	1 %
9	Amin Hazrati Marangalou, Miguel Arturo Gonzalez, Ulkuhan Guler. "Additively Manufactured Receiver Design for Wirelessly-Powered Biomedical Applications", 2023 IEEE Biomedical Circuits and Systems Conference (BioCAS), 2023 Publication	1 %

10	Ghazala Furqan, Najme Zehra Naqvi, Arunima Jaiswal. "Chapter 10 Comparative Analysis of Deep Learning Techniques for Facemask Detection", Springer Science and Business Media LLC, 2022 Publication	1%
11	Shashank Reddy Vadyala, Sai Nethra Betgeri, B. Naga Parameshwari. "Physics-informed neural network method for solving one-dimensional advection equation using PyTorch", Array, 2021 Publication	1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		