A CNN Based Approach to Detect Deepfake

Prof. Nivedhitha M, Pritam Mondal, Arya Panja
School of Information Technology & Engineering
Department of Computer Application
Vellore Institute of Technology, Vellore – 632014, Tamil Nadu, India.
E-mail: nivedhitha.m@vit.ac.in

Abstract

This study proposes a deep learning-based approach for Deepfake prediction using CNN. The preferred method involves training a CNN architecture on a dataset of real and fake images obtained in Kaggle, followed by transfer learning using Xception, which has already been trained on the ImageNet dataset. The model learns to distinguish between real and fake images by identifying patterns and features unique to each class. The results show that the proposed CNN-based approach performs decently in predicting fake images, and we are aiming forward to achieve better results.

Key Words
Deepfake; Image Detection; Convolutional Neural Networks; Deep learning; Xception

I. Introduction

Deepfake technology has caused a stir regarding the exploitation of altered information, particularly images and movies. This has significant implications for a number of fields, including politics, the entertainment industry, and the criminal justice system. Public personalities, such as celebrities, sports, and politicians, are the most severely affected by Deepfakes due to the abundance of videos and images that are readily available online. Deepfake technologies are mostly used to produce adultery material, despite the fact that they are occasionally used to make fun of other people. Cyberbullying is affecting a lot of young people. This study investigates the use of CNNs for deepfake image recognition, evaluating previous studies and approaches, identifying difficulties, and outlining potential future paths. Researchers have shown encouraging results in recognizing deepfake photos with high accuracy.

II. Related Work

[1] This paper uses CNN and LSTM for frame feature extraction and temporal sequence analysis, with a shallow network comprising two fully-connected layers and one dropout layer. The dataset used contains 600 deepfake videos from multiple video-hosting websites and HOHA dataset, with an accuracy of 97.1% with 80 frames. **[2]** This paper proposes a type of VGG network based on noise and image augmentation (NA-VGG) by adding an SRM filter layer and an image augmentation layer in front of the VGG16 network. Celeb-DF dataset is used for training evaluation. The image is extracted from the Deepfake video. The model achieved an accuracy of 85.7 %. **[3]** The proposed architecture transforms RGB pictures into residuals and feeds them into three-layer groups containing a convolutional layer, LReLu, and a max pooling layer. The output is then fed into two fully-connected layers and the SoftMax layer is used to produce the output. Dataset is prepared from CELEBAHQ dataset. **[4]** This study uses optical flow to differentiate between a deepfake and a genuine picture. A CNN pre-trained with VGG-16/ResNet50 is fed optical flows, followed by a sigmoid activation to determine if a frame is false or real. The FaceForensics++ dataset gives an accuracy of 81.61% with VGG16 and 75.46% with ResNet50. **[5]** The proposed CFFN consists of three dense units with a transition layer of 0.5 & a growth rate of 24. A convolution layer with 128 channels and 3x3 kernel size is concatenated to the output layer of the last dense unit. To obtain the discriminative feature representation, a dense layer is inserted last. The dataset used in the experiments was extracted from CelebA, with 10,177 of identities and 202,599 aligned face images. This method has a recall value of 0.900. **[6]** The authors of this work employed the CNN basic architecture and pre-trained the model using several DenseNet and ResNet iterations. Data was supplied into the model and the matching output was produced. The Flickr dataset had 70,000 genuine faces and one million phony faces, and the images were downsized to 256 pixels and combined. The architecture achieved an accuracy of 81.6% with ResNet50, which is the highest.

III. Materials and Methods

- **Dataset**

To improve model generalization, a comprehensive dataset of 140,000 Kaggle images (70,000 real, 70,000 fake) was randomly sampled. 20,000 images were selected for training, ensuring diversity and balanced representation. This approach enabled robustness and accurate classification of real and fake images, forming a foundational step in our research.

- **Data Pre-processing**

By performing various alterations on the original photos, data augmentation is a pre-processing technique used to fictitiously expand the amount of the training dataset. Here firstly, we applied rescaling by dividing the pixel values by 255, thereby normalizing the input pixel range to 0-1. Subsequently, we introduced random rotations within the range of -10 to +10 degrees, horizontal and vertical shifts of up to 10% of the image's width and height, respectively, and shearing transformations up to 20% of the image's width. Furthermore, random zooming within 10% range and horizontal flipping with a 50% probability were employed. To address newly created pixels, the "nearest" fill mode was used, copying the values of the closest pixel. By leveraging these augmentation techniques, we aimed to augment the training dataset, fostering diversity and improving generalizability, which are crucial aspects of our research on image classification.

IV. Proposed Model

The proposed model is a Convolutional Neural Network (CNN) architecture specifically designed for the task of classifying images as either deepfake or non-deepfake. The model leverages the Xception model, which is a pre-trained CNN with weights obtained from training on the extensive ImageNet dataset. By utilizing the Xception model as the base, the proposed architecture benefits from its advanced feature extraction capabilities. The model begins with the inclusion of the Xception model's top layers for initial feature extraction. Following this, a sequence of fully connected layers is added to further refine the extracted features. The first fully connected layer comprises 512 units and employs the Rectified Linear Unit (ReLU) activation function, enabling the network to capture complex nonlinear relationships within the data. To mitigate the risk of overfitting, a dropout layer with a rate of 0.5 is inserted after the initial fully connected layer. Dropout randomly deactivates a portion of the input units during training, encouraging the network to develop more robust and generalized representations by reducing reliance on specific features. Subsequently, another fully connected layer with 128 units and ReLU activation is introduced, followed by another dropout layer with a rate of 0.5. This additional dropout layer aids in further regularizing the model. Finally, a fully connected layer with 64 units and ReLU activation is appended to the architecture. The output layer consists of a single unit with sigmoid activation, producing a probability score that signifies the likelihood of an image being classified as a deepfake. The sigmoid activation function constrains the output between 0 and 1, enabling an interpretable probability interpretation.

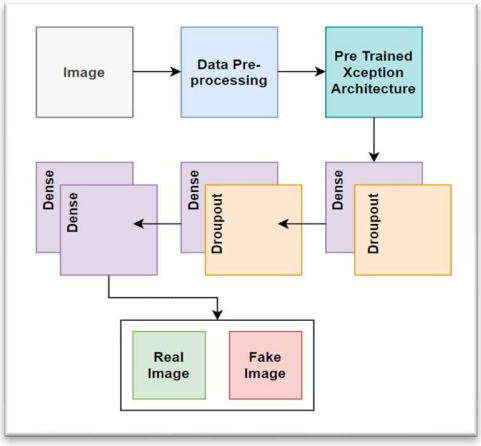
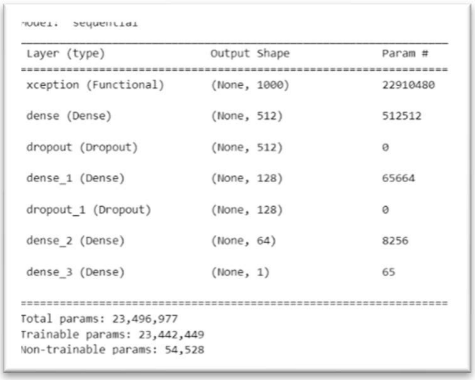


Fig 2. Visual representation of proposed CNN model.



Layer (type)	Output Shape	Param #
xception (Functional)	(None, 1000)	22910480
dense (Dense)	(None, 512)	512512
dropout (Dropout)	(None, 512)	0
dense_1 (Dense)	(None, 128)	65664
dropout_1 (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 64)	8256
dense_3 (Dense)	(None, 1)	65

Total params: 23,496,977
Trainable params: 23,442,449
Non-trainable params: 54,528

Fig 3. Model summary

V. Experimental Setup

The model is evaluated on both the training and validation datasets. The training set evaluation demonstrates a low loss value of 0.1296, indicating minimized discrepancies between predicted and actual values. The accuracy score of 95.13% reflects the model's proficiency in correctly classifying deepfake and non-deepfake images within the training set. The validation set evaluation exhibits a slightly higher loss value of 0.1784. However, the accuracy score of

93.29% indicates that the model maintains a good level of generalization.

VI. Results and Discussions

The evaluation results validate the proposed model's effectiveness in accurately classifying deepfake images. The model achieved notable precision, recall, and F1-scores for both the "real" and "fake" classes, with values ranging from 0.92 to 0.93. Moreover, the model exhibited an overall accuracy of 93% on the test dataset, affirming its robustness in distinguishing between deepfake and non-deepfake images. These findings substantiate the model's potential for practical implementation in real-world scenarios, contributing to the advancement of deepfake detection techniques.

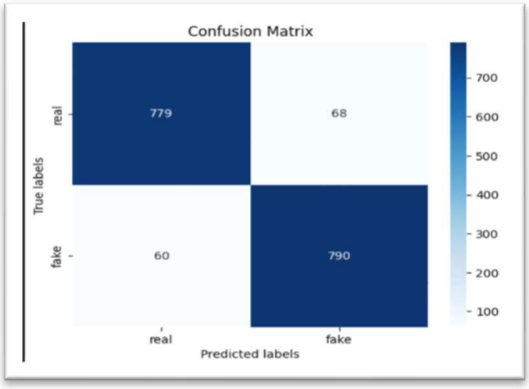


Fig 5. Confusion Matrix

	precision	recall	f1-score	support
real	0.93	0.92	0.92	847
fake	0.92	0.93	0.93	850
accuracy			0.92	1697
macro avg	0.92	0.92	0.92	1697
weighted avg	0.92	0.92	0.92	1697

Fig 6. Obtain Results

VII. Conclusion

This research paper proposed a deep learning-based approach using CNN for deepfake prediction. By training a CNN architecture on a dataset of real and fake images and leveraging transfer learning with the Xception model pre-trained on the ImageNet dataset, the model aimed to discern patterns and features specific to each class. The results of the study indicated that the CNN-based approach showed promising performance in predicting fake images. However, further improvements are desired to achieve even better results. We can further test our dataset with available pre-trained models other than Xception and perform a comparative study of the results and conclude which model works the best. We can also work further on generalization of the model by obtaining samples from various sources.

VIII. Acknowledgement

We would like to express our sincere gratitude to our guide Prof. Nivedhitha M for his invaluable guidance and support throughout the research process. We are also deeply grateful to our Head of the Department Dr. Vanitha M for her support and guidance with the paper. We are thankful to all of our Vellore Institute of Technology (Vellore) Officials for supporting us and help us in gaining knowledge towards this research work.

IX. 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. Deepfake Video Detection through Optical Flow Based CNN, Irene Amerini, Leonardo Galteri, Roberto Caldelli, Alberto Del Bimbo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 0-0.
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.