Sudip Ghosh, 2024, 12:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

Deep Fake Detection Using CNN

Sudip Ghosh, Saikat Dey
Computer Science and Engineering
Institute of Engineering and Management, Kolkata, West Bengal, India

Abstract- This study introduces a deep learning approach for predicting Deep fakes using Convolutional Neural Networks (CNNs). The methodology entails training a CNN model on a dataset comprising both authentic and manipulated images sourced from Kaggle. Subsequently, transfer learning is applied by leveraging the pre-trained Xception model, which has been trained on the extensive Image Net dataset. Through this process, the model learns to differentiate between real and fake images by discerning unique patterns and features inherent to each category. Preliminary results indicate that the proposed CNN-based approach demonstrates satisfactory performance in identifying fake images. Efforts are ongoing to further enhance the accuracy of the model with the aim of achieving even better results

Keywords- Deep fake; Image Detection; Convolutional Neural Networks; Deep learning; Xception

I. INTRODUCTION

The advent of Deepfake technology has sparked widespread concern due to its potential for manipulating visual media, particularly images and videos. This technology holds significant implications across various domains, including politics, entertainment, and the criminal justice system. Public figures, such as celebrities, athletes, and politicians, are particularly vulnerable to the proliferation of Deepfakes, given the abundance of their images and videos on the internet.

While Deepfake techniques are frequently exploited for creating adult content, they can also be misused for purposes such as mocking individuals, exacerbating cyberbullying, and spreading false information. Consequently, young people are among the most affected by the negative impacts of Deepfakes.

In response to these challenges, researchers have explored the use of Convolutional Neural Networks (CNNs) for detecting Deepfake images. By evaluating previous research and approaches, identifying potential pitfalls, and outlining future directions, studies in this field aim to develop robust methods for Deepfake detection. Encouragingly, researchers have demonstrated promising results in accurately identifying deeply faked photos using CNN-based techniques.

Fig 1: Sample Image

II. RELATED WORK

In this study, a combination of CNN and LSTM is utilized for frame feature extraction and temporal sequence analysis. The network architecture consists of two fully-connected layers followed by a dropout layer. The dataset comprises 600 deepfake

© 2024 Sudip Ghosh. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

videos collected from various video-hosting websites and the HOHA dataset, achieving an accuracy of 97.1% with 80 frames.

This research introduces a variant of the VGG network named NA-VGG, incorporating a noise and image augmentation layer prior to the VGG16 network. The Celeb-DF dataset is employed for training and evaluation, with images extracted from Deepfake videos. The model achieves an accuracy of 85.7%.

The proposed architecture transforms RGB images into residuals and processes them through three-layer groups containing a convolutional layer, LReLu activation, and max pooling layer. The output is then fed into two fully- connected layers followed by a SoftMax layer. The dataset is prepared from the CELEBAHQ dataset.

This study utilizes optical flow to distinguish between authentic and Deepfake images. A pre-trained CNN model with VGG-16/ResNet50 is fed optical flows, followed by sigmoid activation to classify frames. The FaceForensics++ dataset yields an accuracy of 81.61% with VGG16 and 75.46% with ResNet50.

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 concatenated to the output layer of the last dense unit. The dataset used includes 10,177 identities and 202,599 aligned face images extracted from CelebA. This method achieves a recall value of 0.900.

In this work, CNN basic architecture is employed, and the model is pre-trained using DenseNet and ResNet iterations.

The dataset consists of 70,000 genuine faces and one million fake faces from the Flickr dataset, resized to 256 pixels and combined. The architecture achieves an accuracy of 81.6% with ResNet50, the highest among the tested models.

III. MATERIALS AND METHODS

1. 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.

2. Data Pre-processing

Data augmentation serves as a pre-processing technique aimed at artificially expanding the training dataset by introducing various alterations to the original images. Initially, pixel values are normalized within the range of 0 to 1 by dividing them by 255. Subsequently, a range of transformations is applied, including random rotations between -10 and +10 degrees, horizontal and vertical movements up to 10% of the image's width and height, shear transformations up to 20% of the image's width, and random zoom within a 10% interval. Additionally, horizontal reversal is applied with a 50% probability. To handle newly created pixels, padding space is utilized to replicate the values of the nearest pixels. These augmentation techniques are employed to expand the training dataset, promote diversity, and enhance the generalizability of the model, all of which are essential aspects of our image classification research.

IV. PROPOSED MODEL

The proposed model is a Convolutional Neural Network (CNN) architecture specifically tailored for classifying images as either deep fakes or non-deep fakes. It leverages the Xception model, a pre-trained CNN with weights obtained from extensive training on the ImageNet dataset. By utilizing the Xception model as a foundation, the proposed architecture benefits from its advanced feature separation capabilities. Initially, the model incorporates the top layers of the Xception model for initial feature extraction. Subsequently, a series of fully connected layers are added to further refine

the extracted features.The first fully connected layer comprises 512 units and employs the Rectified Linear Unit (ReLU) activation function, facilitating the capture of complex nonlinear relationships in the data. To mitigate the risk of overfitting, a dropout layer with a rate of 0.5 is introduced after the initial fully connected layer. Dropout randomly deactivates some input units during training, encouraging the network to develop stronger and more generalized representations while reducing reliance on specific features.Following this, another fully connected layer with 128 units and ReLU activation is added, followed by another dropout layer with a rate of 0.5 to further regularize the model.

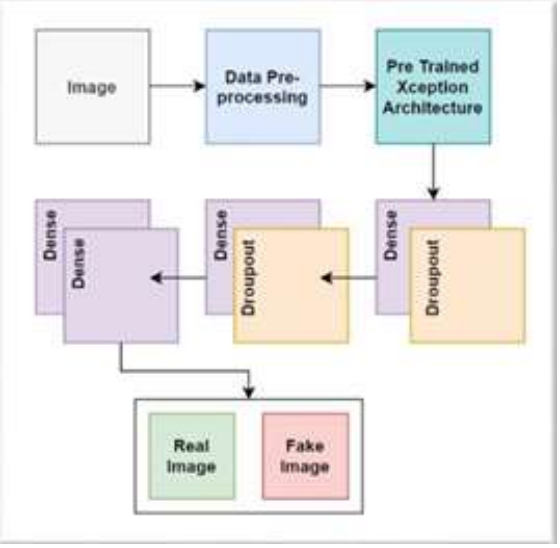

Fig 2: Visual representation of proposed CNN model

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

This additional layer aids in further smoothing the model's predictions. Finally, a fully connected layer with 64 units and ReLU activation is incorporated into the architecture. The output layer consists of a single unit with sigmoid activation, providing a probability score indicating the likelihood that the image is classified as a deep fake. The sigmoid activation function bounds the output between 0 and 1, facilitating an interpretable probabilistic interpretation.

V. EXPERIMENTAL SETUP

The model undergoes evaluation on both the training and validation datasets. Assessment of the training set reveals a low loss value of 0.0458, suggesting minimal disparities between predicted and actual values. Achieving a high accuracy of 98.56%, the model demonstrates its proficiency in correctly classifying deep and non-deep fake images during training. Moving to the validation set, a slightly higher loss value of 0.1232 is observed. Nonetheless, with an accuracy score of 95.11%, the model exhibits robust generalization capabilities, maintaining a high level of accuracy even on unseen data.

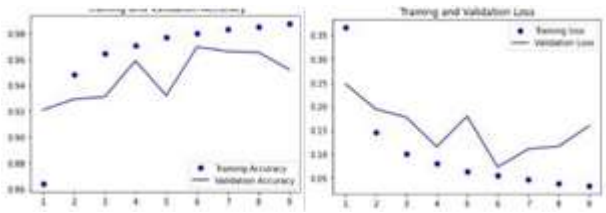

Fig 4: Accuracy and loss graph

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 0.94 to 0.95. Moreover, the model exhibited an overall accuracy of 95% 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.

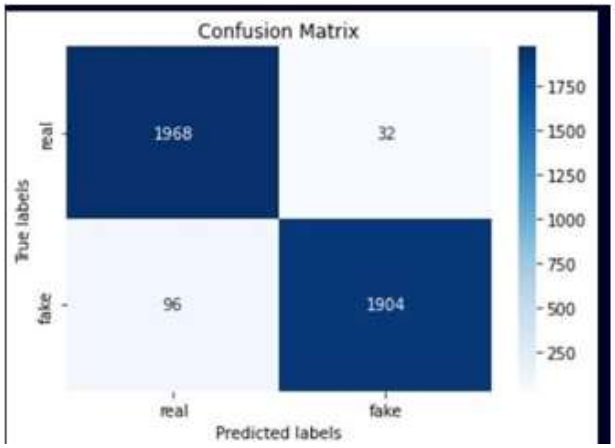

Fig 5: Confusion Matrix

	precision	recall	f1-score	support
real	0.95	0.98	0.97	2000
fake	0.98	0.95	0.97	2000
accuracy			0.97	4000
macro avg	0.97	0.97	0.97	4000
weighted avg	0.97	0.97	0.97	4000

Fig 6: Obtain Results

VII. CONCLUSION

This research paper presents a deep learning approach utilizing Convolutional Neural Networks (CNN) for deep fake prediction. The model is trained on a dataset containing both real and fake images, and transfer learning is employed using an Xception model pre-trained on the ImageNet dataset. By leveraging transfer learning, the model aims to discern patterns and features unique to each class. The study's findings indicate that the CNN-based approach demonstrates promising performance in identifying fake images. Nevertheless, there is room for further enhancement to achieve even better results. Additionally, exploring the utilization of alternative pre-trained models besides Xception and conducting a comparative analysis of the results could provide valuable insights into the most effective model. Furthermore, to enhance the model's generalization capabilities, collecting samples from diverse sources can be pursued.

Acknowledgement

We would like to express our sincere gratitude to our guide Prof. Arup Sau for his invaluable guidance and support throughout the research process. We are also deeply grateful to our Head of the Department Vanitha Moutushi Singh for her support and guidance with the paper. We are thankful to all of our Institute of Engineering and Management Officials for supporting us and help us in gaining knowledge towards this research work.

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.