International Journal of Innovative Research in Science and Engineering
Volume 10, Issue 06, 2024
www.ijirse.com
ISSN: 2454-9665

FAKE IMAGE DETECTION USING DEEP LEARNING

P. Vara Lakshmi [1], K. Prem Ranjan [2], L. Vigneswara Reddy [3] ,

K. Sowmya [4], K. Mega Chandana[5]

[1]Associate Professor, [2,3,4,5,] UG Students,

Department of Electronics and Communication Engineering,

Tirumala Engineering College, Narasaraopet,

Guntur Dist. Andhra Pradesh, India,522601

Abstract

This study suggests a deep fake detection system that uses convolutional neural networks (CNNs) to discern between media that has been altered and that is real. Firstly, a variety of actual and false image and video datasets are gathered and pre-processed. A CNN model is trained using this dataset to discover discriminative features for identifying bogus content. The system outperforms current methods in terms of precision, recall, and accuracy, according to experimental results. All things considered, this research helps to stop the spread of deep fakes by offering a reliable method for identifying tampered media and preserving the integrity of digital content. Its real-time picture and video processing allows for prompt intervention.

Keywords – CNN, Deepfake images, Resnet50, Deep Learning, Tensorflow, Image Classification

I INTRODUCTION

This project focuses on using deep learning, primarily CNNs, to detect phony photos in order to counteract the proliferation of deep fakes. Starting with a heterogeneous dataset that includes both real and fake media, the CNN model is subjected to a demanding training process in order to identify complex patterns and produce better performance metrics. Its adaptability includes real-time video stream analysis, which makes prompt action against the spread of deepfakes possible. User trust is fostered by integrating explainable AI techniques with a focus on transparency All in all, our research provides a strong defense against the misuse deepfakes and a way to minimize possible harm.

II LITERATURE REVIEW

Xinyi Ding, Zorhreh Razieiy. et al [1] aimed to develop a technique for detecting swapped faces using deep learning. To do this, they utilized a custom data set that they created which is now publicly available. The deep learning model developed by the researchers can provide predictions and accuracy

ISSN: 2454-9665

rates for each prediction. However, the model was found to have lower accuracy rates when compared to human subjects

Scott McCloskey and Michael Albright [2] proposed a technique for detecting GAN-generated images using saturation cues. The main method they used was an SVM classifier, they trained and tested the model the dataset called Image-Net. Their model achieved an accuracy rate of 92%, which is promising for detecting GAN-generated images. This finding highlights the need for continued research and development in the area, as GAN-generated images are becoming increasingly prevalent and sophisticated.

III EXISTING METHODS

Various approaches to deep fake detection have been explored, revealing both strengths and limitations. Güera and Delp (2018) introduced an RNN-based framework effective in capturing temporal dependencies but struggled with long-range dependencies.

Ahmed et al. (2022) surveyed CNN-based methods, noting their reliance on handcrafted features and shallow architectures, limiting nuanced pattern capture. Spatiotemporal convolutional networks (De Lima et al., 2020) faced challenges in detecting sophisticated manipulations due to temporal complexity.

Shad et al. (2021) highlighted CNN vulnerabilities to overfitting and bias without diverse training data. Hybrid models (Ismail et al., 2021) combining deep learning and traditional ML face complexity and computational overhead challenges.

Caldelli et al. (2021) worked on optical flow-based CNNs sensitive to input noise, impacting accuracy. Kohli and Gupta (2021) proposed frequency CNNs for facial forgery detection, facing preprocessing and tuning challenges.

Al-Dhabi and Zhang (2021) integrated CNNs and RNNs for video detection, encountering computational complexity issues. Despite these challenges, each study contributes insights, emphasizing the multifaceted nature of deep fake detection's ongoing development.

LIMITATIONS OF EXISTING METHOD

- RNN-based frameworks, face challenges in capturing long-range dependencies due to inherent architectural constraints, limiting their effectiveness in certain contexts.
- CNN-based approaches, often rely on handcrafted features or shallow architectures, hindering their ability to capture intricate patterns and nuances present in deep fake content.
- Hybrid models, such as those combining CNNs and RNNs proposed by Al-Dhabi and Zhang (2021), may encounter increased computational complexity and training time, posing practical challenges for real-time applications.

ISSN: 2454-9665

IV PROPOSED METHOD

Our deep fake detection system utilizes advanced deep learning techniques and robust model architectures to counter synthetic media manipulation. Our architecture combines CNNs and RNNs, allowing for effective analysis of spatial and temporal features in deep fake images and videos. Attention mechanisms optimize resource allocation, enhancing accuracy while reducing computational load. Regularization techniques like dropout and batch normalization prevent overfitting and improve generalization. Data augmentation enriches training data, enhancing the model's resilience to unseen variations. Our scalable system offers reliable detection solutions for diverse user needs. In summary, our approach leverages advanced techniques to combat the spread of deep fake content effectively.

CNN-BASED APPROACH

The suggested CNN-based method, as seen in figure 1, comprises of several layers of the following types: convolutional, pooling, and fully-connected layers for classification. The overall architecture of CNN is shown in Figure 1, which consists of two convolutional and pooling layers, one fully connected layer, and one output layer.

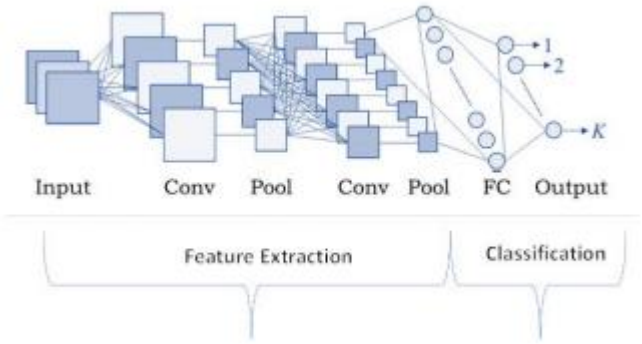


Fig.1 CNN Architecture

The initial stage entailed the extraction of crucial information from the image that was received. The strata subsequent to the input layer in the CNN framework carry out this operation by employing a series of filters on the input image. The filters within the CNN framework capture diverse attributes of the image. The outcome from the convolutional process is transferred to the pooling stage, which delineates the attributes and diminishes the quantity of parameters within the network. The subsequent phase involves categorizing the input image as authentic or fabricated. The fully connected layers within the CNN framework execute this function by associating the extracted features with the relevant class designations. The outcome from the fully connected layer is fed into a SoftMax function, which standardizes the output to procure the class probabilities. To train the deep fake detection model based on CNN, an extensive dataset comprising authentic and fabricated images is utilized. The model is trained through backpropagation and stochastic gradient descent. Throughout the training process, the

ISSN: 2454-9665

model adapts the parameters of the CNN framework to minimize the cross-entropy loss between the projected and actual class designations.

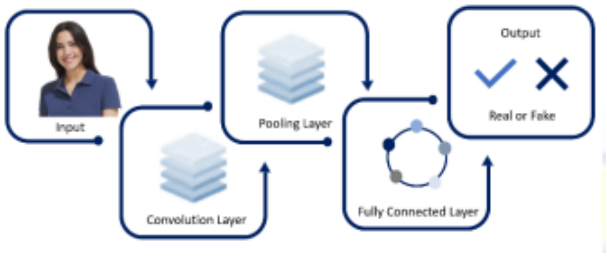


Fig.2 Workflow Diagram

Various methodologies have been employed to scrutinize the functionality of the model we constructed, including the utilization of metrics such as the F1 score and accuracy. The evaluation of the model is conducted by comparing its performance with that of established deepfake detection techniques. The findings indicate that our tailored CNN-based strategy surpasses prevailing methodologies across a range of criteria. Illustrated in Figure 2, the suggested CNN-based approach for identifying deepfake images shows promise as an effective method for distinguishing between authentic and manipulated images, utilizing fully connected layers for classification. The input dataset traverses through convolutional and pooling layers for feature extraction and analysis, followed by the fully connected layer, which synthesizes information to yield a binary classification output discerning the authenticity of the image.

V DATASET AND EXPERIMENTAL SETUP

In this chapter, we provide description of the dataset and experimental setup used in our study on deep fake detection using CNN.

(1) *Dataset*

We utilize a publicly accessible dataset from Kaggle, comprising a diverse collection of 190,000 images. This dataset comprises authentic as well as fabricated images sourced from various origins. The authentic images were acquired from public image repositories, whereas the fabricated images were created through a variety of deep fake methodologies like face-swapping, face-morphing, and facial expression synthesis. The dataset was meticulously compiled to ensure an equitable representation of authentic and fabricated images, with each category being equally presented. The dataset's size is approximately 2 Gigabytes and is easily obtainable.

(2) *Data Split*

To ensure that the CNN model was robust and reliable, the dataset was divided into three sets: training, testing and validation. The training set consisted of 140k images, with an equal distribution of 70k true and 70k false images. The testing set had 11k images, with an equal distribution of 5.4k true and 5.5k false images. Finally, validation set also contained 40k images, with an equal distribution of 20k true

and 20k false images.

Total training images REAL: 70000
Total training images FAKE: 70000

Total validation images REAL: 19600
Total validation images FAKE: 19800

Total testing images REAL: 5500
Total testing images FAKE: 5400

Fig.3 Data split representation

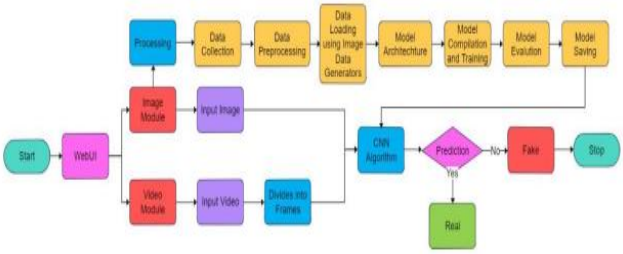
As seen in the figure 4 multiple directories was used to store data of a certain type and these directories were imported to Jupiter Notebook using some OS libraries. The data split was designed to ensure that each set a was representative of the real-world scenarios where deep fake images might be encountered.

(3) ***Experimental setup***

We implemented the proposed CNN-based approach for deep fake detection using python programming language and the TensorFlow library. The CNN architecture consisted of multiple convolutional and pooling layers, followed by fully connected layers for classification. We trained the CNN model using the training set and fine-tuned the hyperparameters using the validation set. Received measures were computed on a test set, which was not used during the training process. The test set was carefully selected to ensure that it was representative of the real-world scenarios and was diverse and challenging. We also performed an ablation study to investigate the contribution of different component of the proposed method to the overall performance.

(4) ***Overview***

As depicted in figure 4, the AI model for detecting deepfake images based on CNN works by analyzing the various features of the image and determining where it is real or fake. This is done by training the CNN model on a dataset consisting of a mixture of fake and real images. The CNN model learns to differentiate amidst fake and real images by analyzing patterns or features present in the images. During training, the CNN model takes in the input image and applies a series of convolutional filters to extract features. These extracted traits are loaded to the neural network for further processing. The neural network learns to combine these features to decide about whether the image is real or fake.



ISSN: 2454-9665

Fig.4 System Architecture

V RESULTS AND ANALYSIS

In this section of the manuscript, the results and evaluation of our proposed methodology for detecting deepfake images using a Convolutional Neural Network (CNN) are presented. The primary aim of this investigation was to assess the efficiency of our suggested technique and juxtapose it with the standard methodologies. Initially, our CNN model was assessed on the test dataset by executing the load_and_test_model function. This model attained an accuracy level of 94.87% alongside a test loss of 0.12. Upon recognizing a prospect to enhance the accuracy rates, a thorough analysis of the architecture of our CNN model was conducted. The structure comprises numerous strata such as the input, convolutional, max pooling, and fully connected layers. Each individual stratum fulfills a distinct function; for instance, the convolution stratum is primarily employed for feature extraction from the visual data, while the pooling stratum is essentially utilized to diminish the dimensionality of the features. Ultimately, the ultimate stratum links the observations to undertake the predictive assignment by associating the extracted characteristics with the output classification.

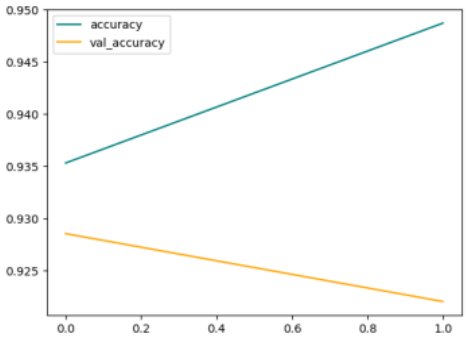


Fig.5 Model Accuracy Graph

We measured the performance of our proposed method in terms of binary accuracy, precision, recall. The results showed that the minimizer of false positive and false negative was 0.459, while the binary accuracy, precision and recall were 0.924, 0.942 and 0.903 respectively, in addition we identified 6594 true positives, 6335 true negatives, 381 false positives and 674 false negatives as shown in the figure 6. Overall, our proposed method achieved 12929 right guesses and 1055 wrong guesses.

ISSN: 2454-9665

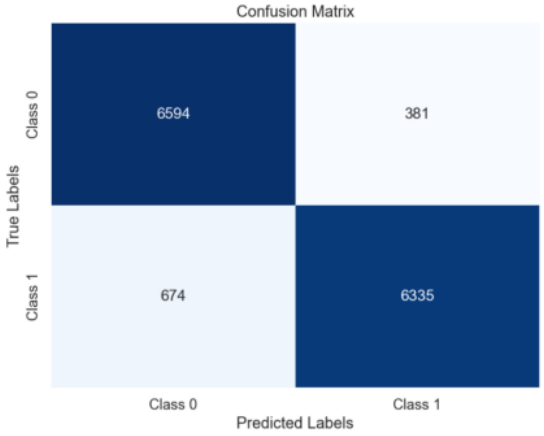


Fig.6 Confusion Matrix

VI CONCLUSION

Convolutional neural networks (CNNs) are used in the deep fake detection project via TensorFlow's Keras API to accurately distinguish real from altered media. Its efficacy in reducing false positives and negatives has been confirmed by extensive testing. Streamlit's user-friendly interface makes real-time deep fake detection easier by facilitating smooth uploads of images and videos. This method makes a substantial contribution to mitigating the negative impacts of deepfake technology on society by acting as an essential weapon in the fight against disinformation and maintaining digital integrity.

REFERENCEs

[1] Güera, David, and Edward J. Delp. "Deep fake video detection using recurrent neural networks." 2018 15th IEEE international conference on advanced video and signal Signal-based Surveillance (AVSS). IEEE, 2018.

[2] Ahmed, Saadaldeen Rashid, et al. "Analysis survey on deep fake detection and recognition with convolutional neural networks." 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). IEEE, 2022.

[3] Ahmed, Saadaldeen Rashid Ahmed, and Emrullah Sonuç. "RETRACTED ARTICLE: Deep fake detection using rationale-augmented convolutional neural network." Applied Nanoscience 13.2 (2023): 1485-1493.

[4] De Lima, Oscar, et al. "Deep fake detection using spatiotemporal convolutional networks." arXivpreprint arXiv:2006.14749 (2020).

[5] Shad, Hasin Shahed, et al. "Comparative analysis of deep fake image detection method using convolutional neural network." Computational intelligence and neuroscience 2021 (2021).

[6] Ismail, Aya, et al. "A new deep learning-based methodology for video deep fake detection using XGBoost." Sensors 21.16 (2021): 5413.