JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DEEP FAKE DETECTION USING MACHINE LEARNING

1. **Thippareddy Karthik Reddy** – B. Tech Student
2. **Kollabathula Rakshan Dev** - B. Tech Student
3. **Gonugunta Chamikar Venkata Srikar** - B. Tech Student
4. **Keesara Poojitha** - B. Tech Student
5. **Prof. E. Srinivasa Reddy** – Professor

Department of Computer Science and Engineering
ANU College of Engineering and Technology, Acharya Nagarjuna University, Guntur, AP.

Abstract:

Social media has seen widespread adoption across diverse domains such as education, entertainment, science, and advertising. However, its misuse has spawned the creation of fake images and videos through deep learning technologies, presenting a formidable challenge in detection. These fabricated videos, commonly known as deepfakes, are digitally manipulated footage, including altered clips of celebrities, crafted with advanced image editing tools, rendering them indistinguishable to the human eye. Traditional software and technical methods struggle to effectively identify such deceptive content, as creators continuously refine their methods. The proliferation of varied fake videos poses a societal threat.

In response, we've devised a technique leveraging the potent XceptionNet architecture, renowned for its excellence in image classification tasks, to spot deepfakes. Our model targets altered faces within realistic, intricate videos by tapping into the hierarchical features learned by convolutional neural networks (CNN). Trained on a diverse dataset of 400 videos featuring an array of facial expressions, lighting conditions, and settings, our approach ensures resilience against sophisticated deepfake methods. By extracting frames from the videos and categorizing them as "real" or "fake," we bolster the model's adaptability and tackle overfitting through a suite of data augmentation and regularization techniques, resulting in an impressive 93.5% accuracy. Furthermore, the integration of adversarial training empowers our model to discern between genuine and synthetic videos more adeptly, thus amplifying its efficacy in deepfake detection.

Keywords: XceptionNet, CNN, Data Augmentation, Regularization Techniques, Adversarial Training

1.Introduction:

Deepfake videos are manipulated media created using advanced deep learning algorithms. They involve replacing faces in existing videos with realistic synthetic ones, making it difficult to distinguish between genuine and fake content. The creation process requires a large amount of video data to train the algorithm, which can then be used to fabricate convincing videos. The potential misuse of deepfakes is alarming, as they can be used to spread disinformation, defame individuals, and even influence political elections. To combat this, researchers are developing detection algorithms, forensic techniques, and policy measures to mitigate the spread of deepfakes.

Identifying manipulated media, such as images, videos, and audio recordings, is the goal of deepfake detection. To uncover these subtle manipulations, experts employ various techniques, including:

- Analyzing visual inconsistencies and artifacts

- Examining facial and body movements for anomalies

- Verifying biometric data, like facial landmarks and voice characteristics

- Utilizing deep learning models to detect patterns and manipulations

- Leveraging blockchain and cryptography to ensure content authenticity and integrity

These methods help expose and combat deepfakes, ensuring the trustworthiness of digital media.

Convolutional Neural Networks (CNNs) are a type of deep learning model that excels in processing structured data like images. They've revolutionized computer vision tasks, enabling accurate image classification, object detection, and segmentation. In deepfake detection, CNNs play a crucial role in identifying manipulated images and videos.

A CNN typically consists of:

1. Convolutional Layers: These layers use adaptable filters to extract features like edges, textures, and patterns from input images.

2. Pooling Layers: These layers reduce spatial dimensions while preserving essential information.

3. Activation Functions: Non-linear functions like ReLU introduce non-linearity, enabling the network to grasp complex relationships.

4. Fully Connected Layers: These layers perform classification tasks, leveraging high-level features from convolutional layers.

During training, CNNs use backpropagation and gradient descent algorithms to minimize a loss function, adjusting weights to predict labels accurately.

In deepfake detection, CNNs can:

- Extract discriminative features from authentic and manipulated images

- Analyze facial features, textures, and artifacts introduced by deepfake generation algorithms

- Classify images as authentic or manipulated

By leveraging CNNs, deepfake detection models can effectively identify and flag manipulated media.

2. Literature Review:

Various Researchers across the world have proposed deepfake detection techniques to identify fake visual content. Here we concisely review some of the techniques as mentioned in the literature.

D. Pan, L. Sun, R. Wang, X. Zhang, and R. Sinott[evaluated the effectiveness of Xception and MobileNet, two deepfake detection methods, in identifying forged videos. Using FaceForensics++ datasets, they achieved high accuracy ranging from 81% to 90% across various deepfake techniques. They also proposed a novel voting mechanism that combines the results from all four methods, enhancing detection accuracy.

A. Mitra, Saraju P. M Peter Corcoran, and E. Kougianos proposed an efficient deepfake detection framework for social media videos, achieving high accuracy while reducing computational cost by up to 80%. By leveraging key frame extraction to analyze only representative frames, they significantly minimize computational overhead. Then, they employ a carefully chosen Xception-based CNN to extract robust features from these keyframes. Finally, a classifier network trained on extensive datasets differentiates real and deepfake videos based on the extracted features, enabling accurate detection with reduced computational cost.

M. Alshaihli, O. Elharous, S. Al- Maded, and A. Boridane proposed a technique called FaceFakeNet, which uses ResNet50 as a local feature extractor for RGB images. The ResNet50 backbone is followed by parallel attention architectures that capture richer contextual feature details. The output from the spatial and channel attention modules is concatenated and fed into a convolution layer, resulting in a one-channel output that indicates whether the input is real or fake.

Y. Wang, V. Zarghami, and S. Cui introduced LBP-Net, a novel approach to detecting fake faces using binary image representation texture. Their experiments show that LBP-Net is more robust against image tampering and augmentations than existing methods. Additionally, ensemble models demonstrate higher accuracy in identifying manipulated images compared to individual models.

Giuliani, M., Narducci, F., Placitelli, A., Guaragni, L., Coletti, A. developed a method that exploits the artifacts introduced by GANs in reconstructed images to detect deep fakes. They trained a GAN-based autoencoder to reconstruct authentic faces, which learns to accurately reproduce real faces but struggles with deep fakes. The detection of deep fakes is based on identifying the discrepancies in reconstruction quality, allowing for effective identification of manipulated images.

3. Proposed Methodology:

To develop an effective machine learning model for detecting fake images, the following detailed approach should be implemented:

a) Data Collection:

1. Collecting Real Images:

- Use public datasets (e.g., ImageNet, COCO) to source real images.

- Ensure the images are diverse, covering various subjects, backgrounds, and lighting conditions.

2. Collecting Fake Images:

- Utilize existing datasets with manipulated or deepfaked images (e.g., DeepFake Detection challenge, FaceForensics++).

- Generate fake images using various techniques (e.g., GANs, deepfake software) for diversity.

3. Balancing the Dataset:

- Ensure an equal number of real and fake images to avoid bias.

- Apply data augmentation to underrepresented categories if necessary.

b) Data Preparation:

1. Standardization:

- Resize all images to a uniform size (e.g., 224x224 pixels).

- Convert images to a standard colour space, such as RGB.

2. Feature Extraction:

- Extract edges using methods like the Canny edge detector.

- Compute texture features using techniques like Local Binary Patterns (LBP).

- Generate colour histograms for each image.

3. Normalization:

- Normalize pixel values to a common range (e.g., 0 to 1 or-1 to 1).

c) Model Selection:

1. Selecting a Model:

- Begin with Convolutional Neural Networks (CNNs) due to their effectiveness in image classification tasks.

- Consider architectures like VGG16, ResNet50, or EfficientNet.

2. Experimentation:

- Test different architectures and hyperparameters (e.g., learning rate, number of layers, batch size).

- Use techniques such as grid search or random search for hyperparameter tuning.

d) Model Training:

1. Training:

- Use binary cross-entropy as the loss function for binary classification.

- Implement data augmentation (e.g., rotations, flips, zooms) to increase dataset size and robustness.

2. Optimization:

- Use optimizers like Adam or SGD with momentum.

- Apply regularization techniques (e.g., dropout, weight decay) to prevent overfitting.

e) Model Evaluation:

1. Validation:

- Split the dataset into training, validation, and test sets.

- Evaluate the model on the validation set using metrics such as accuracy, precision, recall, and F1 score.

2. Fine-Tuning:

- Adjust the model architecture or hyperparameters if performance is unsatisfactory.

- Use cross-validation to ensure robust evaluation.

f) Model Deployment:

1. Real-Time Classification:

- Deploy the model using a suitable framework for production (e.g., TensorFlow Serving, ONNX, FastAPI).

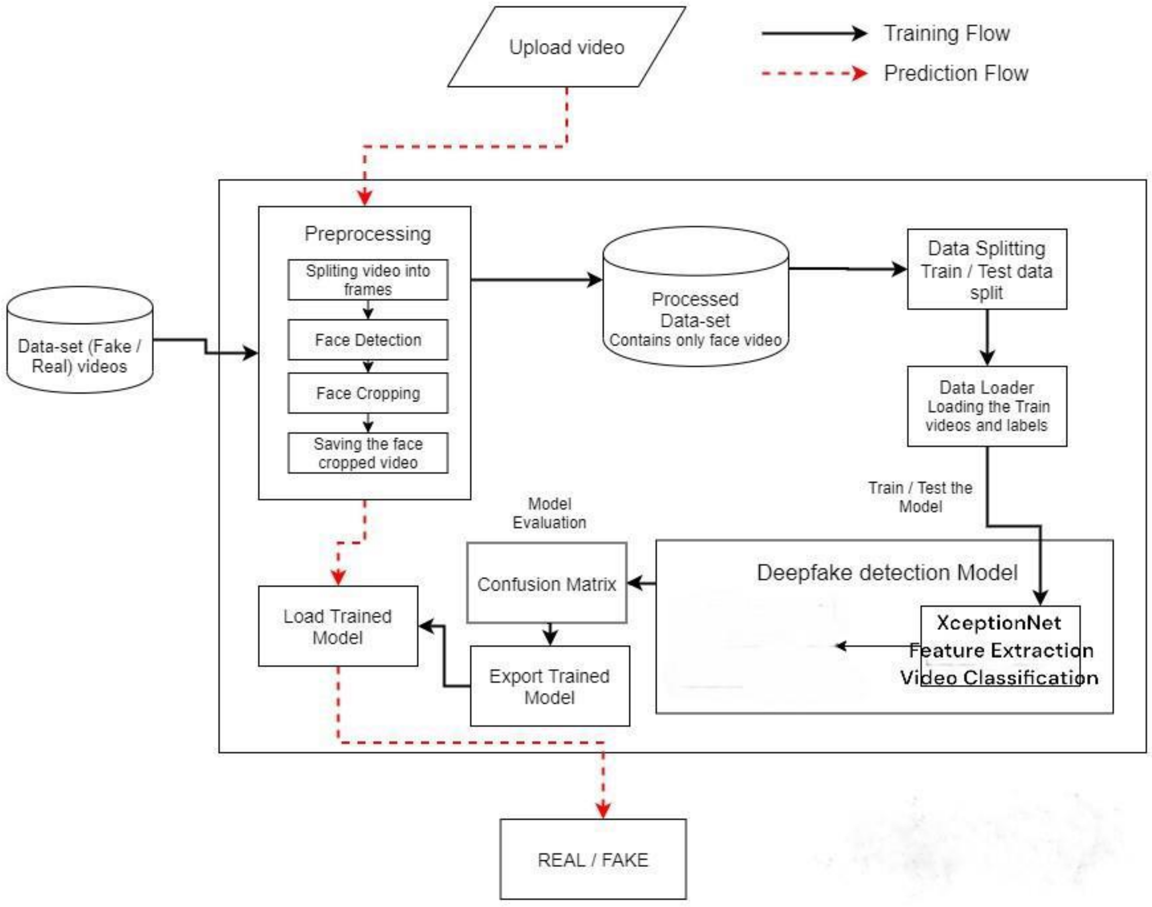
- Ensure the system can handle real-time classification with low latency.

2. Monitoring and Maintenance:

- Continuously monitor the model's performance in production.

- Retrain the model periodically with new data to maintain accuracy and adapt to new types of fake images.

Following these steps will help build a robust and effective machine learning model for detecting fake images. Each step ensures the model is accurate, efficient, and capable of performing well in real-world applications.



4. Implementation:

Data Preprocessing:

Video Preprocessing:

1. Frame Extraction:

   - Use libraries like OpenCV to extract frames from video files.

2. Face Detection:

   - Apply face detection algorithms (e.g., Dlib, OpenCV) to locate faces within each frame.

3. Cropping and Resizing:

   - Crop and resize face regions to a standardized size for consistency.

Image Preprocessing:

1. Resizing:

   - Resize images to a uniform resolution, typically square dimensions.

2. Normalization:

   - Normalize pixel values to a common scale for better model training convergence.

3. Data Augmentation:

   - Apply transformations such as translation, scaling, and flipping to increase dataset diversity and improve model generalization.

Data Splitting:

1. Splitting the Dataset:

   - Split the dataset into training and test sets, ensuring a balanced distribution of real and fake samples.

2. Maintaining Proportional Representation:

   - Ensure real and fake samples are proportionally represented in each split to prevent bias.

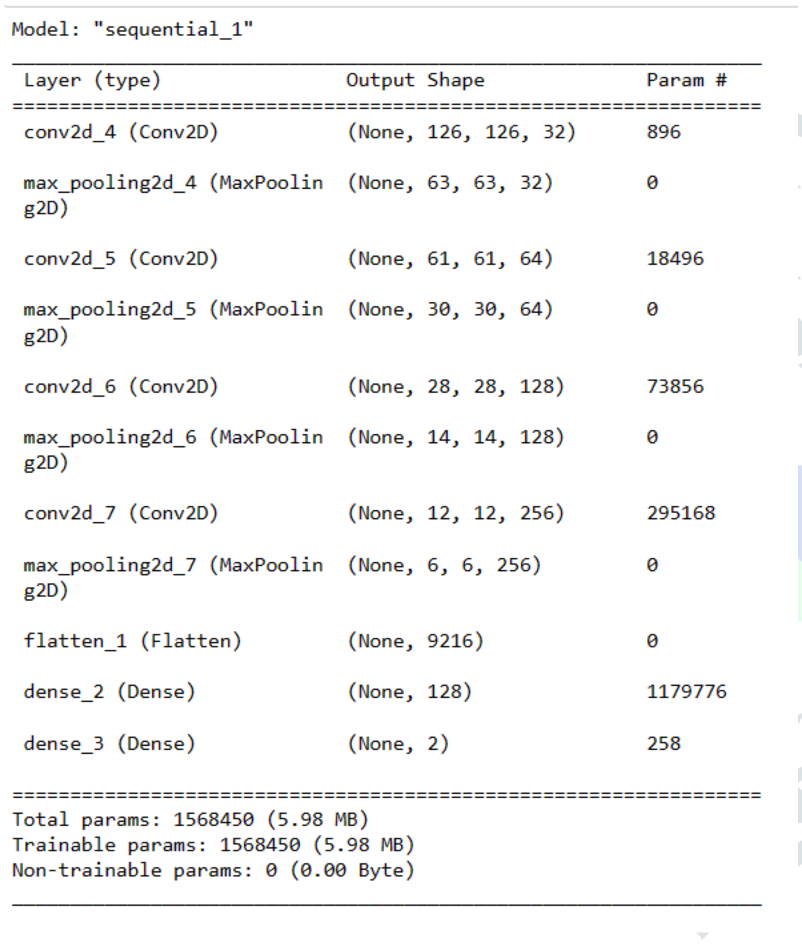
CNN:

XceptionNet, developed by Google researchers, is a neural network architecture that uses Depthwise Separable Convolutions instead of traditional Inception modules. This innovative approach involves a depthwise convolution followed by a pointwise convolution, effectively acting as an Inception module with a maximally large number of towers. The architecture includes Depthwise Separable Convolution blocks, Maxpooling layers for downsampling, and shortcut connections similar to ResNet. These elements together form a highly efficient and scalable deep convolutional neural network, significantly enhancing performance in image classification tasks.

5. Results:

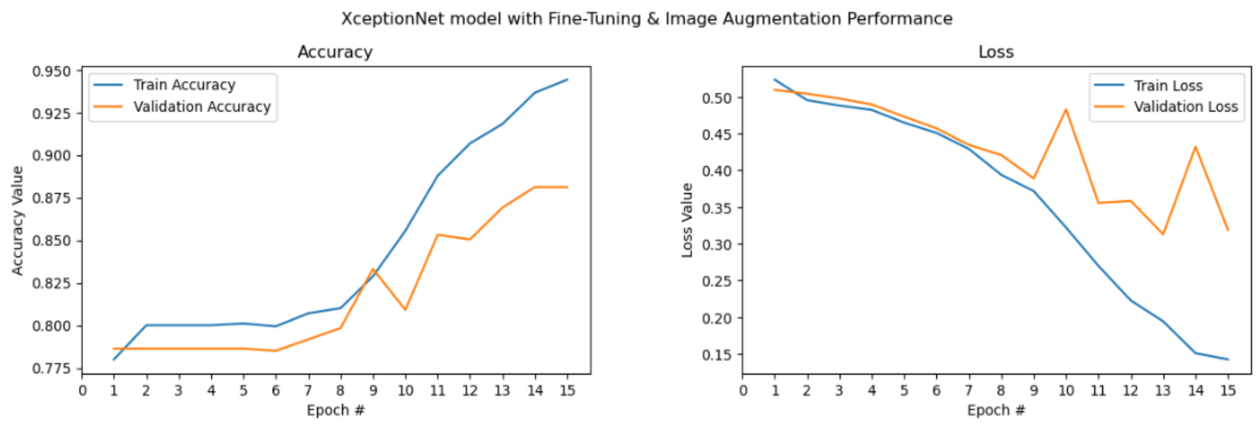
Deep fake detection using machine learning involves training a model with a dataset of both real and fake images and then using this model to classify new images. The effectiveness of this process depends on factors like the quality of the training data, the model's complexity, and the techniques used for detection. Machine learning algorithms are quite adept at identifying simple manipulations, such as resizing or cropping. However, detecting more advanced fakes, such as deepfakes, is more challenging due to their complexity and realism. While machine learning is a powerful tool for detecting fake images, it's important to understand that no detection method is perfect. Human judgment and expertise are still essential to confirm an image's authenticity, as machine learning alone may not catch all types of fakes accurately.

Proposed CNN Model Summary:



```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_4 (Conv2D)           (None, 126, 126, 32)      896

 max_pooling2d_4 (MaxPoolin  (None, 63, 63, 32)        0
 g2D)

 conv2d_5 (Conv2D)           (None, 61, 61, 64)        18496

 max_pooling2d_5 (MaxPoolin  (None, 30, 30, 64)        0
 g2D)

 conv2d_6 (Conv2D)           (None, 28, 28, 128)       73856

 max_pooling2d_6 (MaxPoolin  (None, 14, 14, 128)       0
 g2D)

 conv2d_7 (Conv2D)           (None, 12, 12, 256)       295168

 max_pooling2d_7 (MaxPoolin  (None, 6, 6, 256)         0
 g2D)

 flatten_1 (Flatten)         (None, 9216)              0

 dense_2 (Dense)             (None, 128)               1179776

 dense_3 (Dense)             (None, 2)                 258

=================================================================
Total params: 1568450 (5.98 MB)
Trainable params: 1568450 (5.98 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Epoch Analysis:



6. CONCLUSION:

We introduced a methodology for discerning between authentic and deep fake videos, accompanied by the confidence level of our model's predictions. Our approach adeptly analyzes videos at a rate of one frame per second, yielding high accuracy. Employing the XceptionNet CNN model, we extract frame-level features and conduct classification to determine the authenticity of the video. Our model seamlessly processes the video into a sequence of frames, facilitating accurate predictions.

FUTURE SCOPE:

-Further, this dataset and model will be used to combine detection of deepfake audio, video, or lip-synced videos.

-For lip-synced video, FakeAVCeleb dataset can be used

-Artificially generated audio signals can be detected using this algorithmic model.

REFERENCES:

[1] ARossler, DCozzolino, L.Verdoliva, "FaceForensics++: Learning to Detect Manipulated Facial Images" in arXiv:1901.08971.

[2] Dataset: https://drive.google.com/drive/folders/1AwQ8adyGquf8uBMeuHJk-HJsKfnZ3D6V?usp=sharing

[3] YLi , XYang , Pu Sun "Celeb-DF: A Large-scale Challenging Dataset for Deep Fake Forensics" in arXiv:1909.12962

[4] Deepfake examples that terrified and amused the internet: https://www.creativebloq.com/features/deepfake-examples

[5] G. Antipov, M. Baccouche, and J.-L. Dugelay. Face aging with conditional generative adversarial networks. arXiv:1702.01983, Feb. 2017

[6] J. Thies et al. Face2Face: Real-time face capture and reenactment of rgb videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2387–2395, June 2016. Las Vegas, NV.

[7] Deepfakes, Revenge Porn, And The Impact On Women: https://www.forbes.com/sites/chenxiwang/2019/11/01/deepfakes-revenge-porn-and the-impact-on-women/

[8] The rise of the deep fake and the threat to democracy https://www.theguardian.com/technology/ng-interactive/2019/jun/22/the-rise-of-the-deepfake-and-the-threat-to-democracy

[9] Deep Fake using GAN Tianiang shen Ruian liu Ju Bai Zheng Li Report16.pdf (ucsd.edu)