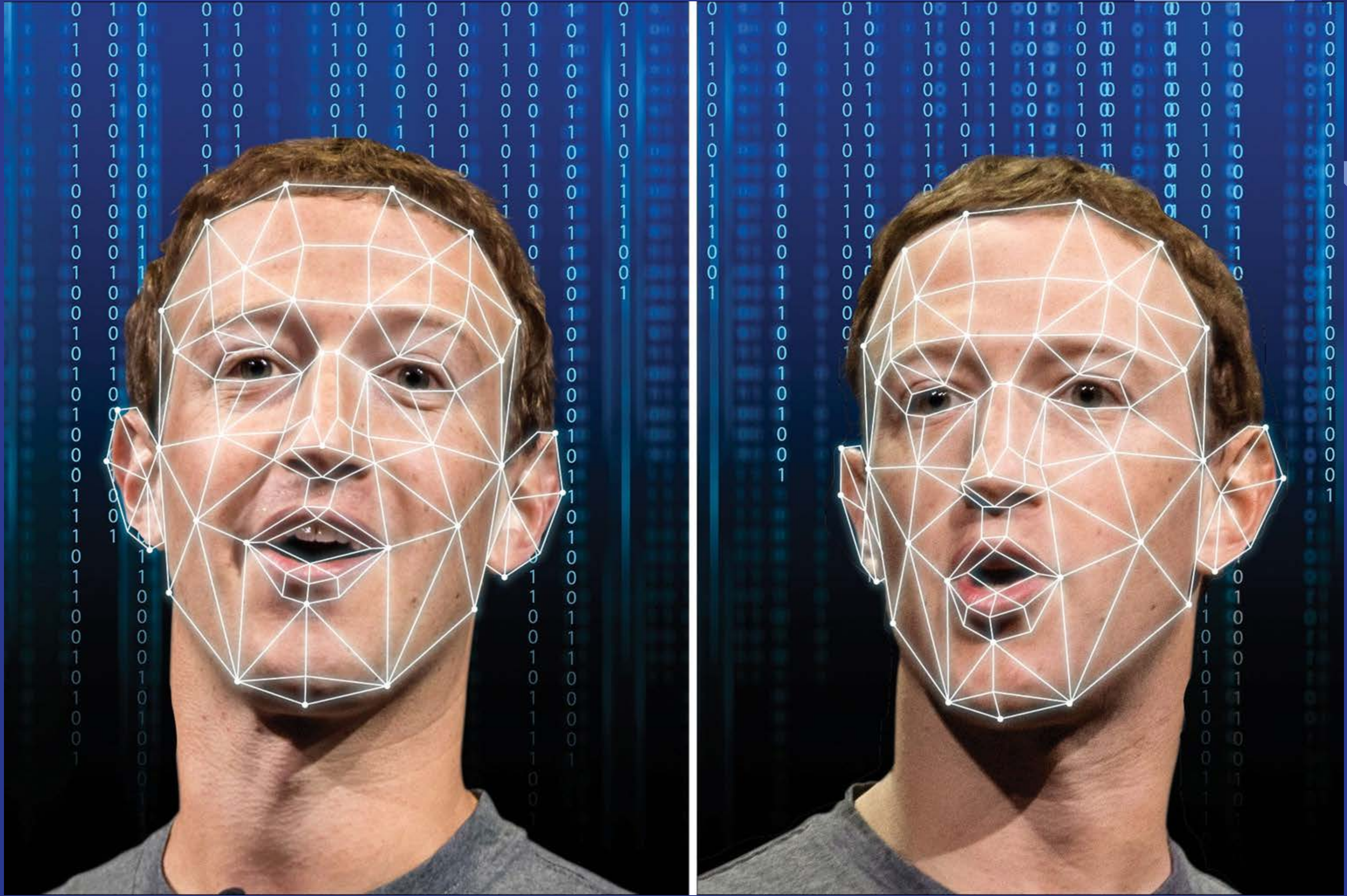
Deepfake Detection Using Deep Learning Techniques

Authors: Sayantan Bhattacharyya, Milind Chakraborty, Nitin Sharma

Guide: Prof. Dharmendra Singh Rajput



Introduction

Research Topic: Deepfake Detection Using Deep Learning Techniques

Problem:

- Deepfake videos threaten individual and national security.
- They manipulate public opinion, spread misinformation, and endanger individuals.
- Current detection methods are limited in accuracy and effectiveness.

Objective:

- Investigate and develop robust techniques for detecting deepfake videos.
- Utilize advancements in vision transformers and inception net technology for accurate detection.

Importance:

- Critical need for innovative solutions in combating deepfake threats.
- Developing a reliable detection method is crucial for safeguarding individuals and strengthening national security.

Literature Review

Overview:

- Summarizes key findings from relevant research papers.

Methods and Results:

- Various approaches for deepfake detection explored by researchers.
- CNN, LSTM, VGG network, optical flow, and dense units utilized for frame feature extraction, image augmentation, and residual conversion.

Accuracy and Performance:

- Different models achieved varying levels of accuracy.
- Ranging from 75.46% to 97.1% depending on the methodology and dataset used.

Significance:

- Literature highlights the ongoing efforts to develop effective deepfake detection methods.
- Provides valuable insights for informing our own research approach and methodology.

References:

- Citations of relevant research papers for further reading and validation of findings.

Literature Review - Citation 1

Authors: D. Güera and E. J. Delp
Methodology:
- Used CNN and LSTM for frame feature extraction and temporal sequence analysis.
- Shallow network with two fully-connected layers and one dropout layer.

Dataset:
- Contains 600 deepfake videos from multiple sources and the HOHA dataset.

Accuracy:
- Achieved 97.1% accuracy with 80 frames.

Significance:
- Demonstrates effectiveness of CNN and LSTM in deepfake detection.
- Provides a strong baseline for comparison with other methods.

Literature Review - Citation 2

Authors: X. Chang et al.
Methodology:
- Proposed a VGG network based on noise and image augmentation.
- Utilized an SRM filter layer and image augmentation layer.

Dataset:
- Trained and evaluated on the Celeb-DF dataset.

Accuracy:
- Achieved an accuracy of 85.7%.

Significance:
- Introduces innovative approach using noise and augmentation for detection.
- Shows promising results on a widely used dataset.

Literature Review - Citation 3

Authors: Huaxiao Mo et al.
Methodology:
- Converted RGB images into residuals and passed through convolutional layers.
- Used three-layer groups with convolutional layers, LReLu activation, and max pooling.

Dataset:
- Prepared from the CELEBA HQ dataset.

Accuracy:
- Actual accuracy not mentioned in provided information.

Significance:
- Highlights a unique approach of converting images into residuals for detection.
- Provides insights into leveraging architectural designs for deepfake detection.

Literature Review - Citation 4

Authors: Irene Amerini, Leonardo Galteri, Roberto Caldelli, Alberto Del Bimbo
Methodology:

- Used optical flow and CNN pre-trained with VGG-16/ResNet50.
- Utilized sigmoid activation to determine frame authenticity.

Dataset:

- Utilized the FaceForensics++ dataset.

Accuracy:

- Achieved 81.61% accuracy with VGG16 and 75.46% with ResNet50.

Significance:

- Demonstrates the use of optical flow for deepfake detection.
- Provides insights into the effectiveness of different CNN architectures.

Literature Review - Citation 5

Authors: Hsu, Chih-Chung, Yi-Xiu Zhuang, Chia-Yen Lee
Methodology:
- Proposed a CFFN consisting of dense units with transition layers and a growth rate.
- Utilized a convolution layer with 128 channels and 3x3 kernel size.

Dataset:
- Utilized a dataset extracted from CelebA.

Accuracy:
- Achieved a recall value of 0.900.

Significance:
- Introduces a novel architecture for deepfake detection.
- Shows promising recall values for identifying manipulated images.

Literature Review - Citation 6

Authors: Hasin Shahed Shad et al.
Methodology:

- Employed basic CNN architecture and pre-trained models using DenseNet and ResNet iterations.
- Dataset consisted of 70,000 genuine faces and one million fake faces.

Accuracy:

- Achieved an accuracy of 81.6% with ResNet50.

Significance:

- Demonstrates the effectiveness of pre-trained models for deepfake detection.
- Provides insights into handling large-scale datasets for training.

Literature Review - Citation 7

Authors: Theerthagiri P, Basha Nagaladinne
Methodology:
- Utilized the InceptionNet Convolutional Neural Network (CNN) algorithm for deepfake detection.
- Different types of transitions in real images were used for testing.

Dataset:
- Utilized the DFDC dataset.

Accuracy:
- Achieved an overall accuracy of 93%.

Significance:
- Highlights the effectiveness of InceptionNet for deepfake detection.
- Provides insights into performance metrics on real-world datasets.

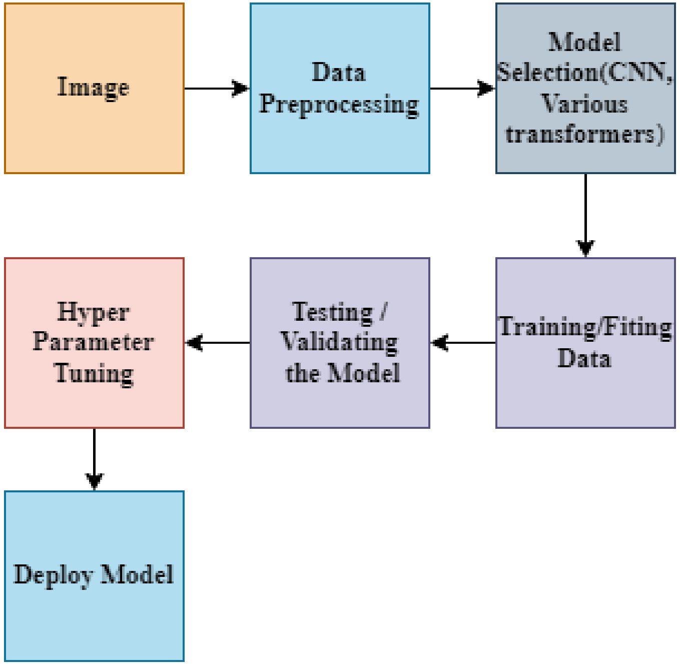
Proposed Framework

Overview:

- Presents the proposed deepfake detection framework.
- Highlights the sequential steps involved in the process.

Steps:

1. **Data Collection:** Gathering diverse dataset of authentic and manipulated images from reliable sources.
2. **Preprocessing:** Tasks include resizing, normalization, and facial landmarks extraction to prepare images for analysis.
3. **Model Selection:** Choosing suitable architectures, including a hybrid model combining CNNs and pretrained transformers.
4. **Training:** Training the model to identify subtle visual cues indicative of manipulated content.
5. **Testing:** Assessing the model's performance using a testing dataset.
6. **Evaluation:** Analyzing metrics such as accuracy, precision, recall, and F1 score to evaluate the model's effectiveness.



Data Collection

Key Points:

- **Gathering Diverse Dataset:** Collecting a wide range of authentic and manipulated images from reputable sources.
- **Reliable Sources:** Stressing the significance of reliable sources to ensure the quality and authenticity of the dataset.

Objective:

- To lay the foundation for robust model training and evaluation by acquiring a comprehensive dataset representative of real-world scenarios.

Preprocessing

Overview:

- Details the preprocessing phase within the deepfake detection framework.
- Highlights essential tasks to prepare the dataset for model training.

Key Tasks:

- **Resizing and Normalization:** Ensuring uniformity in image dimensions and pixel values for consistent processing.
- **Facial Landmarks Extraction:** Identifying key facial features to aid in the detection process.

Objective:

- To optimize the dataset for analysis and model compatibility, enhancing the effectiveness of subsequent stages in the framework.

Model Architecture

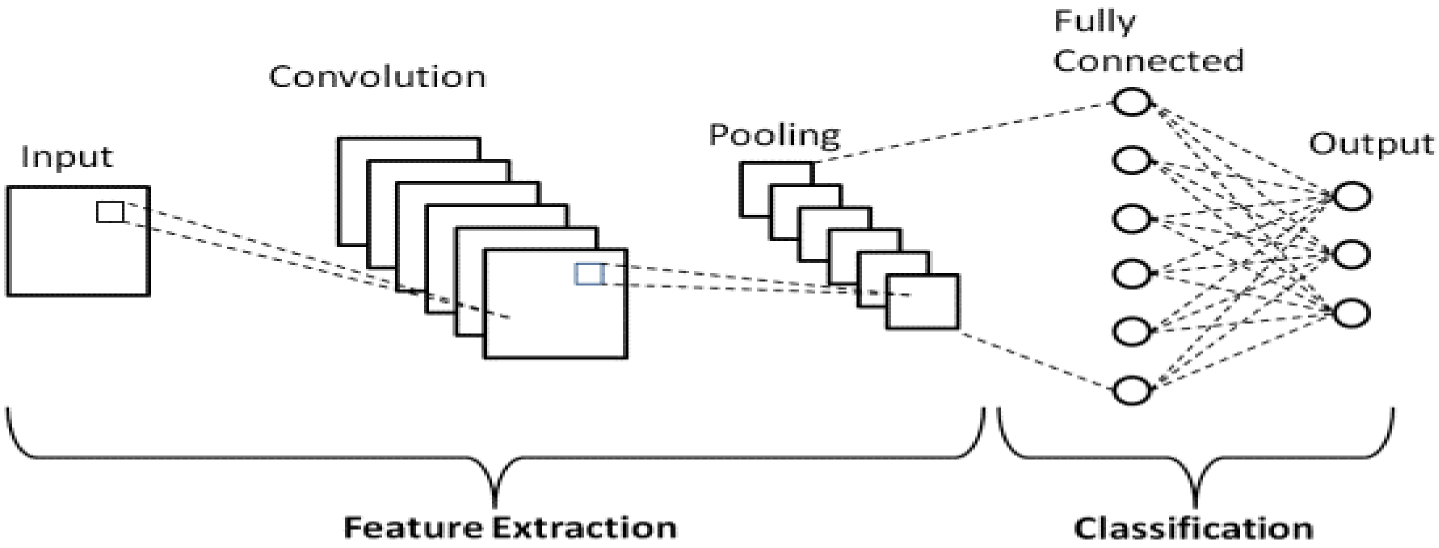
Key Components:

- **Hybrid Model:** Incorporating both CNNs and pretrained transformers to capture spatial and temporal dependencies in the images.
- **Spatial and Temporal Analysis:** Leveraging the strengths of each architecture to effectively discern manipulated content.

Objective:

- To develop a versatile and robust model capable of accurately detecting deepfake videos by leveraging advanced neural network architectures.

Basic CNN Model



https://www.analyticsvidhya.com/blog/2022/03/basic-introduction-to-convolutional-neural-network-in-deep-learning/ [accessed 27 Feb, 2024]

Training

Key Tasks:

- **Feature Learning:** Teaching the model to extract relevant features from the dataset.
- **Parameter Optimization:** Fine-tuning model parameters to enhance performance and accuracy.

Objective:

- To equip the model with the ability to effectively differentiate between authentic and manipulated content through comprehensive training on diverse datasets.

Evaluation Metrics

Key Metrics:

1. **Accuracy:** Measures the overall correctness of the model in classifying authentic and manipulated videos.
2. **Precision:** Indicates the ratio of correctly identified manipulated videos to the total videos classified as manipulated.
3. **Recall:** Reflects the proportion of manipulated videos correctly identified by the model out of all actual manipulated videos.
4. **F1 Score:** Balances precision and recall, providing a single metric to evaluate model performance.

Comparative Analysis

Overview:

- Conducts a comparative analysis of different CNN architectures and pretrained transformer models used in deepfake detection.
- Identifies strengths and weaknesses of each approach to inform model selection.

Key Points:

1. **CNN Architectures:** Discusses various CNN architectures such as VGG, ResNet, and DenseNet, highlighting their performance in deepfake detection.
2. **Pretrained Transformers:** Explores the effectiveness of pretrained transformers like BERT and GPT in capturing semantic features.

Results

Key Findings:

1. **Accuracy Rate:** Provides the accuracy rate achieved by the model in detecting deepfake videos.
2. **Effectiveness:** Highlights the model's effectiveness in accurately distinguishing between authentic and manipulated content.

Implications:

- Demonstrates the practical applicability and reliability of the proposed deepfake detection framework in real-world scenarios.
- Reinforces the significance of robust model training and evaluation in combating the proliferation of deepfake videos.

Conclusion

Summary:

- Recapitulates the key findings and contributions of the study.
- Emphasizes the importance of the proposed deepfake detection method.

Significance:

- Highlights the significance of the research in addressing the growing threat of manipulated media.
- Stresses the need for continued research and development in deepfake detection technology.

Future Directions:

- Suggests potential areas for future research and improvements in deepfake detection techniques.
- Encourages collaboration and innovation in the field to stay ahead of evolving deepfake technology.