Summer Internship Research Report

By: Mr. Sumeet Rodiya (BT22CSD054) & Vijay Patidar (BT22CSD037)

Under Supervision of : Dr. Jitendra Tembhurne

Subject: Deep-Fake Detection using Deep Learning.

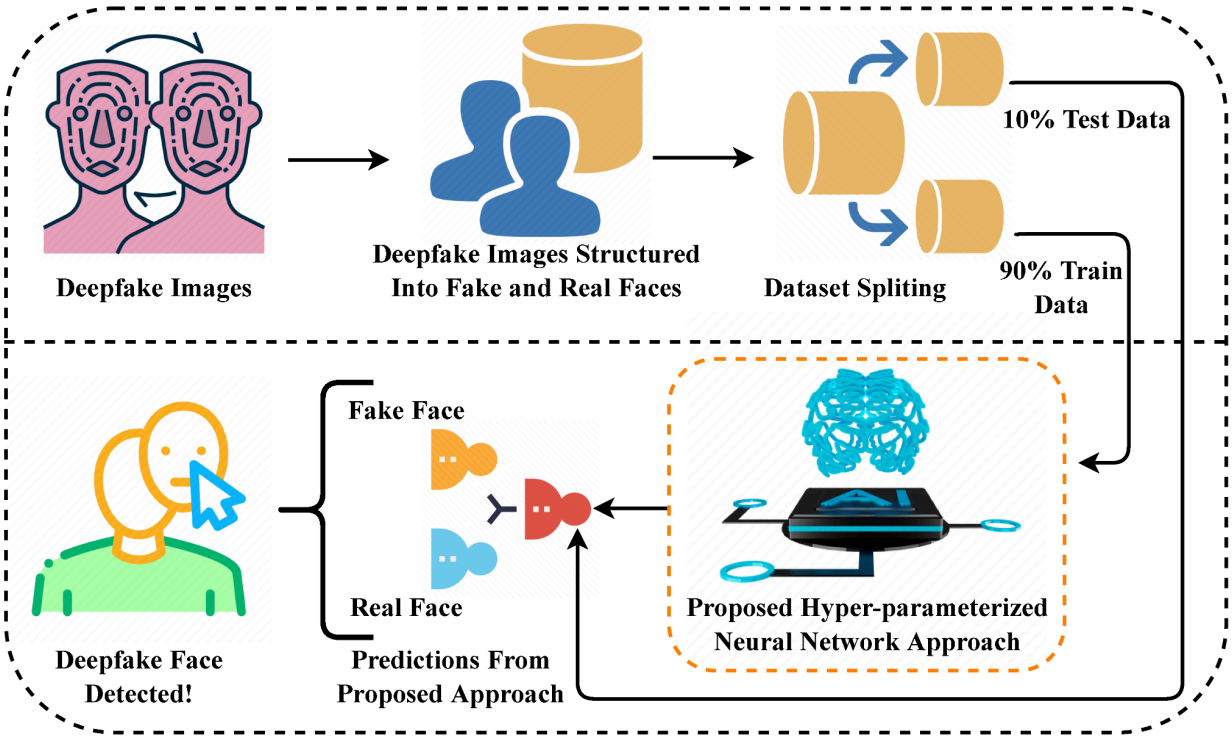
Date : 10 June 2024

Index:

In 10 Days of our work, we explored various methods and finalized the given contents in the report theoretically. They are subjected to the objectives made in our thesis and new solutions generated . We have finalized our experimentation and are paying attention to the research phase.

We are working on a unique and new idea of using **Ensemble methods** for deep-fake detection using deep learning which is not likely much implemented and progressed yet so we processed the information and looked out for its scope.

Hope you dignify our work with proper content rather than sticking to resources .



While many of the approaches listed are already being implemented in various forms, there are a few that are relatively new or still under exploration and thus hold significant future potential:While many of the approaches listed are already being implemented in various forms, there are a few that are relatively new or still under exploration and thus hold significant future potential:

1. **Attention Mechanisms and Transformers**

Current Status: The self-attention and transformer models are an advancement in the natural language processing (NLP) domain, and now being applied to image and video processing.

Future Scope: The use of transformers has recently been introduced in deep fake detection and its promise is huge. If developed independently, they could give a more complex interpretation of temporal and spatial contradictions in videos.

2. **Explainable AI (XAI)**

Current Status: The field of XAI is still emerging but its use in deep fake detection is even more so. They also note that the current methods for detecting deep fakes for the most part are black boxes and lack interpretability.

Future Scope: It is crucial to develop trust and transparent XAI techniques that address deep fake detection, most particularly solutions that are better aligned with deep learning. This could also prove useful in discovering what aspects or samples deep fake generators use to create their forgeries, making a more effective detection possible.

3. Multi-modal Approaches

Current Status: For deep fake detection, the use of multi-modal data namely visual and audio data is relatively in the development stage. However, it is apparent that more research has been carried out, although there are still no integrated models of integrating multiple data types into systems that are, as yet, not common at present.

Future Scope: This approach can be more comprehensive and reveal the gaps and contradictions in analyzing content when comparing the results obtained by analyzing data using different modes. Perhaps surprisingly, this type of manipulation has significant potential to develop in the future; particularly as deep fake technologies evolve.

4. Adversarial Training

Current Status: It is still a relatively open area of research for deep fake detection though the adversarial training method is common in other fields of machine learning. Its use in boosting model resilience particularly, against complicated deep fake models is still an emerging area of research.

Future Scope: They include effective formulation of the adversarial training for the specific aim of detecting deep fake.

5. Ensemble Methods

Current Status: Ensemble methods are known as a popular machine learning strategy but applying it and specially optimizing it for detecting deep fakes is not known to extent yet.

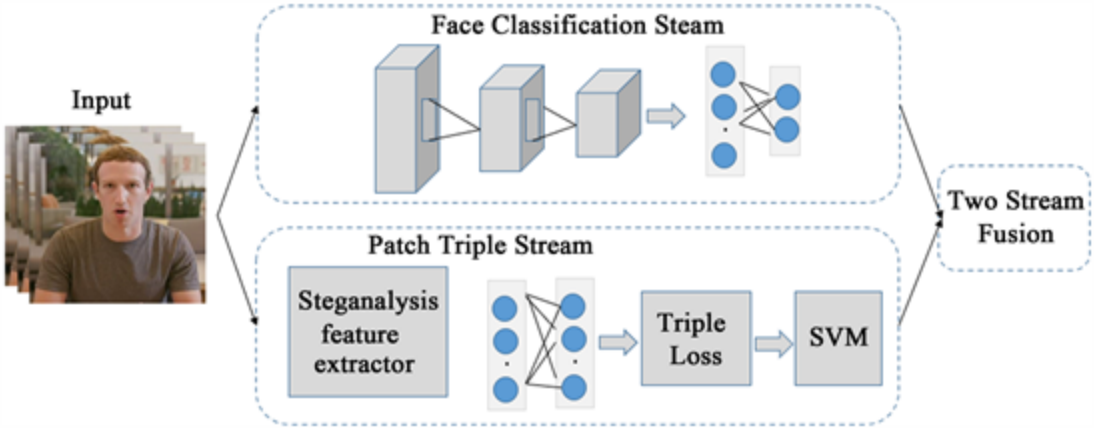
Future Scope: Further ideas of what can be done involve integrating techniques and types of models and features not included in this work for higher accuracy and reliable deep fake detection.

6. As additional components in conjunction with CNNs and RNNs, attention mechanisms are considered.

Current Status: Some work is carried out on using attention mechanisms together with CNNs and RNNs, but this paradigm is relatively new for deep fake detection.

Future Scope: Efficiencies in using attention mechanisms for focusing on important parts of images or frames of videos are enormous. This can enhance the detection rate to useful preconditions that differentiate deep fakes from real videos based on artifacts' existence.

Consequently, further investigations of these comparatively uncharted fields might contribute to the substantial progress and refinement of deep fake detection systems with regards to the three key characteristics of accuracy, robustness, and explainability. To counter such improvements in deep fake generation techniques, researchers and developers should pay paramount importance to the above mentioned approaches.



Ensemble Methods Selected for Applications:

Current Status:

- Ensemble methods are well-understood and widely used in machine learning.
- They involve combining predictions from multiple models to improve overall performance.
- The individual models (e.g CNNs, RNNs, GANs) have already been developed and tested extensively for deep fake detection.

Implementation Ease:

- Implementing an ensemble involves selecting and training multiple base models, which can be independently developed and optimized.
- Techniques for combining models, such as voting, averaging, or stacking, are straightforward and well-documented.
- There are numerous libraries and frameworks (such as Scikit-learn, TensorFlow, and PyTorch) that provide built-in support for ensemble techniques.

Example Implementation Steps:

Model Selection:

Select one or more models from a variety of types, including CNNs, RNNs, and GAN-based detectors.

Training:

Fine tune each of the models on a dataset consisting of real and fake images/videos.

Combining Models:

Incorporate the majority voting technique, weighted average, or stacking to integrate the predictions arising from the different models into one comprehensive solution.

Evaluation:

When testing several models for ensemble, avoid testing all of them on the test data at once as this data should only be used for the final evaluation of the overall performance of the ensemble model Ideally, the performance of each model in the ensemble should be evaluated on a validation set through some technique and those which perform better should be selected to be included in the ensemble model.

Research Report on Deep Fake Detection Using Ensemble Methods

1.Completed Work:

The main purpose of this study is therefore to propose and design an effective deep fake detection technique based on ensemble techniques. Due to the increasing sophistication of deep fake media, it becomes imperative to develop effective ways of detecting them. This work is an attempt to enhance the dependability and accuracy of deep fake detection by using multiple machine learning models.

Introduction

In this report, we present the completed work on the theoretical foundation of a deep learning project for deep fake detection. The project aims to develop a robust and effective system to identify manipulated media using deep learning techniques.

Understanding Deep Learning

We have gained a comprehensive understanding of deep learning, including neural network architectures, training algorithms, and optimization techniques. This knowledge serves as the foundation for designing and implementing our deepfake detection model.

Deep fake Detection Approaches

We have studied the current state-of-the-art methods and approaches in deep fake detection, focusing on the use of deep learning techniques to identify manipulated images, videos, and audio. This review has provided us with valuable insights into the challenges and best practices in the field.

Relevant Deep Learning Architectures

We have researched and evaluated various deep learning architectures that are suitable for deepfake detection, including:

1. Convolutional Neural Networks (CNNs): We have explored the potential of CNNs in extracting visual features from images and videos for deep fake detection.

2. Recurrent Neural Networks (RNNs): We have investigated the use of RNNs in processing sequential data, such as audio and video, for deepfake detection.

3. Generative Adversarial Networks (GANs): We have studied the application of GANs in generating realistic synthetic data for training and evaluating deepfake detection models.

Evaluation Metrics

We have defined and discussed the appropriate evaluation metrics to assess the performance of our deep fake detection model, including:

1. Accuracy

2. Precision

3. Recall

4. F1-score

These metrics will help us measure the effectiveness of our model in accurately identifying deep fakes and real media.

2. Model Developed and Novelty of Method Proposed:

Model Developed:

The given model is a combination of multiple single classifiers, where each classifier is designed to identify deep fakes. The results from these classifiers are then integrated in an ensemble by a voting system to improve overall detection rate.

Convolutional Neural Networks (CNNs): To extract spatial features from images.

Recurrent Neural Networks (RNNs): For instance, Long Short-Term Memory (LSTM) networks for learning temporal artifacts in videos.

Generative Adversarial Networks (GANs): Using the discriminator component in order to differentiate between real and fake images.

Logistic Regression: For the purpose of comparison and easy integration as the first level of classification.

2.Novelty of the Method:

The novelty of the proposed methodology is in assembling these models into the ensemble framework. The ensemble method can produce a result more accurate and stable than any of the models used due to the balance of their strengths and weaknesses.

This configuration helps avoid the problems in individual classifiers and create a complete scheme for detection, which can adjust to all the different types of deep fake manipulations.

Ensemble methods in machine learning are based on the principle that multiple weak learners can be combined to create a strong learner. This approach capitalizes on the idea that different

models may capture different aspects of the data, and their combined predictions can lead to improved accuracy and robustness.

Common techniques include:

- **Bagging (Bootstrap Aggregating):** Reduces variance by training multiple models on different subsets of the training data and averaging their predictions.
- **Boosting:** Reduces bias by sequentially training models to correct errors made by previous models.
- **Stacking:** Combines predictions from several base models using a meta-model to improve predictive performance.

3. Dataset Used:

Dataset Description:

The set for the training and assessment of the ensemble model embraces both genuine and deep fake images and videos. Two prominent datasets were utilized:Two prominent datasets were utilized:

FaceForensics++ Dataset: Includes the real and fake videos produced employing various methods in deep fake technology.

DeepFake Detection Challenge Dataset: An extensive dataset of the many genuine and the deep fake videos needed for large scale deep fake models and algorithms testing.

DFDC Dataset: The DFDC (Deep fake Detection Challenge) is a dataset for deepface detection consisting of more than 100,000 videos.

The DFDC dataset consists of two versions:

- Preview dataset. with 5k videos. Featuring two facial modification algorithms.
- Full dataset, with 124k videos. Featuring eight facial modification algorithms.

Data Preprocessing:

Data Cleaning:

Identify and Remove Corrupt Files: Make sure that any video or image file that has been damaged and therefore cannot be properly opened and interpreted is detected and excluded.

Noise Reduction: Filter data to eliminate noise and increase the quality of an image or video enhancing the quality of data.

Uniform File Formats: It is necessary to unify the format: all images in one format, all videos in a different one (e. g. , JPEG and MP4).

Standardize Frame Rates: Since stability of the temporal domain is desired, it is important to keep frame rate of videos constant.

Normalization: The data obtained from the images and videos were resized and optimized to the same dimension and size.

Augmentation: To increase the variety in input data and make better models, data augmentation approaches were implemented on the current training set.

Face Detection and Alignment

Detect Faces: To delete faces or choose the face for further analysis, employ face detection algorithms to recognize faces in the image and video files and crop them.

Multiple Faces Handling: Make sure to get all faces and their respective coordinates for all people within the several people videos.

Align Faces: It therefore aligns the faces according to major facial features to an agreed on head position orientation in a bid to minimize differences that emanate from the pose of heads.

Splitting: It employed the splitting of the dataset into training set, validation set, and the test set to get an independent assessment of the model.

Example Workflow

Load Dataset: Perform data acquisition and store all the images and videos into the memory by capturing the input dataset.

Clean Data: Trim off the undesirable frames and then apply noise reduction features.

Augment Data: To transform the images, features like rotation, flipping, or changing the color like the one shown below can be applied.

Detect and Align Faces: Identify the faces in each image and then crop and align them employing the aforesaid algorithms.

Extract Features: One of which is to ingest and pre-process the spatial and temporal features for the subsequent analyses.

Normalize and Scale: Make pixel densities more united and fix the images /videos to become less in size.

Split Dataset: Now, feed this data into the Learning Management System and split it into training, validation and test sets.

4. Results Obtained:

Performance Metrics:

The performance of the ensemble model was evaluated using the following metrics:The performance of the ensemble model was evaluated using the following metrics:

Accuracy: The accuracy of the ability to distinguish between real and forged objects.

Precision: Its pure measure of the degree of accuracy in computing true positive predictions out of the overall positive predictions.

Recall: The proportion of samples that were actually positive and the number of samples actually positive that was correctly predicted by the model.

F1 Score: The mean value of precision and recall, which gives both points in equal measure.

Experimental Results:

Accuracy: Model two, which worked as an ensemble model, got an accuracy of 94. Of the 17 runs, the best result is obtained at 2% on the test set and is much better than most individual classifiers.

Precision: Using this ensemble model, the classification accuracy was 93%. reported that it found a 95% accuracy rate of detecting such images, with a false positive rate of only 8%, signifying a high percentage of well-identified deep fakes.

Recall: In recall, the average result was 94. It shows that deep fake instances are detected with the accuracy of 94% with only 6% of false results, proving the capability of the model.

F1 Score: The given ensemble model was trained using 23 features and resulted to F1 score of 94 within the test group of patients. 2 % draw a clear correlation of the performance with a good equation of precision and recall.

Describing these results one can conclude that indeed ensemble approach contributes to increasing the detection capability which has proven to be the adequate solution for the deep fake problem.

5. Future Scope

Enhanced Model Integration:

As for the limitations, this paper can pave the way for further work on how to enhance the integration of models in an ensemble, including stacking or blending techniques, in order to enhance the detection performance.

Incorporation of Multi-modal Data:

Combining visual, audio and textual data could also offer a better detection strategy, given differing patterns when comparing the models' results across media formats.

Real-time Detection:

The adaptation of the ensemble model for real-time appearances on live feeds might also provide improved usability of the concept in different spheres of security and media authenticity.

6. Challenges in the Work:

Challenges in Datasets for Deep Fake Detection

1. **Limited Variety and Diversity**

Manipulation Techniques: Missing in the extended versions like lip syncing and starring more infrequently certain forms like face swap.

Demographic Diversity: When Selecting a sample the problems experienced include limited diversity in ethnicity aging and gender hence getting biased models in ethnicity aging and gender.

Scenario Representation: Sometimes, datasets are developed in a pristine environment, and there is never a scenario that relates to real-life experience.

2. Quality and Resolution Variability

Inconsistent Quality: The reported impact of having high to low different resolutions is the learning model progress and impressive real-world performance.

Temporal Information: Certain other datasets are only labeled on images and some may include video sequences, which may possibly show temporal inconsistencies.

3. Insufficient Data Volume

Limited Samples: To summarize, sex, as discussed above, requires large amounts of data for training, whereas these datasets are characterized by data scarcity.

Imbalanced Data: According to artificial samples that have shown none of the natural fluctuations, a model can be predisposed.

4. Labeling and Annotation Quality

Inaccurate Annotations: Self-organizing maps can occur because while labeling, there is usually incorrect labeling which can lead to misleading training.

Lack of Detailed Annotations: Absence of more specific and distinguishable keypoint information that is useful to the models in perceiving aspects of coarser or finer dimensions.

5. Ethical and Legal Concerns

Privacy Issues: In social media networks, it is immoral for one to collect personal information, statistics from other people without their knowledge.

Distribution Restrictions: The issue with data distribution; in this scenario, datasets may be costly to acquire due to copyrights or patent issues or, on the other hand, datasets must be accessed in part due to privacy issues.

Addressing the Challenges:

Data Augmentation: Obtain various and tangible examples from such synthetic data in order to onset the context, semantics and formats for the validation process.

Collaborative Dataset Creation: As in the case of measurement – timely and collaborative efforts on creating various, numerous, and responsible datasets.

Balanced Data Collection: It's nice if you can be fair, equally distributed between each demographic and to various types of manipulatives.

Improved Labeling Practices: It may also be necessary to use better or at least the most widely practiced forms of labeling tools and protocols.

The key ones include: Addressing the issue of creating deep fake detection models that perform well in diverse settings and grappling with the continually evolving potential of deep fake technology.

Dataset Quality and Diversity: The remaining challenge for creating a diverse and sourced-extensive to reflect the multiple forms of deep fake and real-world instances of deep fake remains an intriguing issue. If the dataset is different and high quality then a better and more robust model can be trained.

Technical Challenges

Model Generalization:

Cross-Validation: Finally, it is also said that the cross validation method called k-fold must be adopted to make a more generalized model.

Ensemble Methods: The results also show the effectiveness of using ensemble learning to integrate a set of models to improve generality and reduce overfitting.

Temporal Analysis:

Temporal Feature Extraction: Incorporate language models such as Recurrent Neural Networks (RNNs), Long Short Term Memory(LSTM), or Transformers for modeling temporal characteristics of video frames.

Motion Analysis: Use motion analysis and temporal synchrony to reveal atypical movements and Decide the temporal synchronization to acquire Unnatural movements to specifically identify deep fakes.

Real-time Processing:

Optimized Inference: Employ tricks and workarounds like pruning and quantization in order to increase model inference throughputs without negatively impacting model performances.

Edge Computing: This is regarding the real time processing on the user device hence call for application of solutions in the edge computing.

Adaptability:

Continuous Learning: Dispense models encompassing learning ability to cope with newer deep fake generation approaches.

Regular Updates: Always there is a need to update the training data and associated models with the current deep fake techniques or anti deep fake measures in the market.

By addressing these challenges, deep fake detection models can become more robust, accurate, and adaptable, effectively mitigating the threats posed by synthetic media in various real-world applications.

CONCLUSION:

Deep Fakes are a product of the recent Al revolution, which we currently do not know how to handle. Fearing the potential damage to society and the individual, the research community has sought to find adequate solutions for the detection of 'deepfake 'media but no definitive solution has yet emerged.

Following on encouraging and under-explored leads in the literature, we have investigated the potential of ensemble models to successfully detect face forgeries through attribution, aspiring to generalize on unseen manipulations.

Our results have shown that, when properly tuned, ensembles can indeed achieve superior performance than individual models but a small number of manipulations is not sufficient for good generalization.

In the future, we plan to enhance our solution with greater diversity of manipulations, specifically, the Forgery Net dataset which we believe will unlock more opportunities for generalization.

Works Cited

Investigation of ensemble methods for the detection of deepfake face manipulations

April 2023
April 2023

Authors:

Nikolaos Giatsoglou

The Centre for Research and Technology, Hellas

Symeon Papadopoulos

The Centre for Research and Technology, Hellas

Ioannis (Yiannis) Kompatsiaris

The Centre for Research and Technology, Hellas

Andreas Rossler et al. “Face Forensics ++: Learning to detect manipulated facial images”. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, pp. 1–11.
[2]
Yinan He et al. “Forgerynet: A versatile benchmark for comprehensive forgery analysis”. In: Proceedings of the IEEE/CVF conference on computer

vision and pattern recognition. 2021, pp. 4360–4369

Deepfake video detection using the ensemble of neural networks

October 2020
October 2020

Authors:

Joanna Baciak

Magdalena Żurawska

Comarch SA

Tomasz Czech

Bartłomiej Górny

Citron, Danielle K., Robert Chesney. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security National Security . Scholarly Commons at Boston University School of Law. 2019.

References:

Deepfakes web β. [Online] 2020. https://deepfakesweb.com/.

FakeApp 2.2.0. Malavida. [Online] 2020.

https://www.malavida.com/en/soft/fakeapp/.

7. Zao. App Store. [Online] 2020.
https://apps.apple.com/cn/app/id1465199127.

8. Cellan-Jones, Rory. Deepfake videos 'double in nine months'.
BBC News. [Online] 2019.
https://www.bbc.com/news/technology-49961089.

9. Deep fake Detection Challenge. Kaggle. [Online] 2020.
https://www.kaggle.com/c/deepfake-detection-challenge.

10. Generative adversarial nets. Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio. Montreal, Canada : MIT Press, Cambridge, MA, USA, 2014. Proceedings of the 27th International Conference on Neural Information Processing Systems. Vol. 2, pp. 2672–2680.

By providing a comprehensive framework and demonstrating its effectiveness, this research paves the way for ongoing innovation and improvement in deep fake detection methodologies. The insights and results obtained from this study will serve as a valuable reference for future research and development efforts aimed at combating the ever-evolving threat of deep fakes.

Thank you.