Jyothi Hills, Panjal Road,
Vettikattiri PO, Cheruthuruthy , Thrissur,
Kerala 679531

Jyothi Engineering College
NAAC Accredited College with NBA Accredited Programmes*
Approved by AICTE & affiliated to APJ Abdul Kalam Technological University
A CENTRE OF EXCELLENCE IN SCIENCE & TECHNOLOGY BY THE CATHOLIC ARCHDIOCESE OF TRICHUR
JYOTHI HILLS, VETTIKATTIRI P.O, CHERUTHURUTHY, THRISSUR. PIN-679531 PH : +91- 4884-259000, 274423 FAX : 04884-274777
NBA accredited B.Tech Programmes in Computer Science & Engineering, Electronics & Communication Engineering, Electrical & Electronics Engineering and Mechanical Engineering valid for the academic years 2016-2022. NBA accredited B.Tech Programme in Civil Engineering valid for the academic years 2019-2022.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SEMINAR REPORT

Deep Fakes and Its Detection Techniques

Submitted by

KARTHIK PC
JEC17CS061

Supervised by

Ms. ASWATHY WILSON
Assistant Prof., Dept. of CSE

in partial fulfilment for the award of the degree

of

BACHELOR OF TECHNOLOGY (B.Tech)

in

COMPUTER SCIENCE & ENGINEERING
of

A P J ABDUL KALAM TECHNOLOGICAL UNIVERSITY

CREATING TECHNOLOGY LEADERS OF TOMORROW

DECEMBER 2020

Jyothi Hills, Panjal Road,
Vettikattiri PO, Cheruthuruthy , Thrissur,
Kerala 679531

Jyothi Engineering College
NAAC Accredited College with NBA Accredited Programmes*
Approved by AICTE & affiliated to APJ Abdul Kalam Technological University
A CENTRE OF EXCELLENCE IN SCIENCE & TECHNOLOGY BY THE CATHOLIC ARCHDIOCESE OF TRICHUR
JYOTHI HILLS, VETTIKATTIRI P.O, CHERUTHURUTHY, THRISSUR. PIN-679531 PH : +91- 4884-259000, 274423 FAX : 04884-274777
NBA accredited B.Tech Programmes in Computer Science & Engineering, Electronics & Communication Engineering, Electrical & Electronics Engineering and Mechanical Engineering valid for the academic years 2016-2022. NBA accredited B.Tech Programme in Civil Engineering valid for the academic years 2019-2022.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SEMINAR REPORT

Deep Fakes and Its Detection Techniques

Submitted by

KARTHIK PC
JEC17CS061

Supervised by

Ms. ASWATHY WILSON
Assistant Prof., Dept. of CSE

in partial fulfilment for the award of the degree

of

BACHELOR OF TECHNOLOGY (B.Tech)

in

COMPUTER SCIENCE & ENGINEERING
of

A P J ABDUL KALAM TECHNOLOGICAL UNIVERSITY

CREATING TECHNOLOGY LEADERS OF TOMORROW

DECEMBER 2020

Department of Computer Science and Engineering

JYOTHI ENGINEERING COLLEGE, CHERUTHURUTHY

THRISSUR 679 531

DECEMBER 2020

BONAFIDE CERTIFICATE

This is to certify that the seminar report entitled **Deep Fakes and Its Detection Techniques** submitted by **Karthik PC (JEC17CS061)** in partial fulfillment of the requirements for the award of **Bachelor of Technology** degree in **Computer Science and Engineering** of **A P J Abdul Kalam Technological University** is the bonafide work carried out by him under our supervision and guidance.

Ms. Aswathy Wilson	**Mr. Shaiju Paul**	**Dr. Vinith R**
Seminar Guide	Seminar Coordinator	Head of The Dept
Assistant Professor	Assistant Professor	Professor
Dept. of CSE	Dept. of CSE	Dept. of CSE

DEPARTMENT OF

COMPUTER SCIENCE & ENGINEERING

COLLEGE VISION

Creating eminent and ethical leaders through quality professional education with emphasis on holistic excellence.

COLLEGE MISSION

- To emerge as an institution par excellence of global standards by imparting quality engineering and other professional programmes with state-of- the-art facilities.
- To equip the students with appropriate skills for a meaningful career in the global scenario.
- To inculcate ethical values among students and ignite their passion for holistic excellence through social initiatives.
- To participate in the development of society through technology incubation, entrepreneurship and industry interaction.

DEPARTMENT OF

COMPUTER SCIENCE & ENGINEERING

DEPARTMENT VISION

Creating eminent and ethical leaders in the domain of computational sciences through quality professional education with a focus on holistic learning and excellence.

DEPARTMENT MISSION

- To create technically competent and ethically conscious graduates in the field of Computer Science & Engineering by encouraging holistic learning and excellence.
- To prepare students for careers in Industry, Academia and the Government.
- To instill Entrepreneurial Orientation and research motivation among the students of the department.
- To emerge as a leader in education in the region by encouraging teaching, learning, industry and societal connect

PROGRAMME OUTCOMES (POs)

1. **Engineering Knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
2. **Problem Analysis**: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
3. **Design/Development of Solutions**: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
4. **Conduct Investigations of Complex Problems**: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
5. **Modern Tool Usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
6. **The Engineer and Society**: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
7. **Environment and Sustainability**: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
8. **Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
9. **Individual and Team Work**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
10. **Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
11. **Project Management and Finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
12. **Life-Long Learning**: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

PROGRAMME EDUCATIONAL OBJECTIVES (PEOs)

1. The graduates shall have sound knowledge of Mathematics, Science, Engineering and Management to be able to offer practical software and hardware solutions for the problems of industry and society at large.
2. The grad uates shall be able to establish themselves as practising professionals, researchers or Entrepreneurs in computer science or allied areas and shall also be able to pursue higher education in reputed institutes.
3. The graduates shall be able to communicate effectively and work in multidisciplinary teams with team spirit demonstrating value driven and ethical leadership.

Programme Specific Outcomes (PSOs)

1. An ability to apply knowledge of data structures and algorithms appropriate to computational problems.
2. An ability to apply knowledge of operating systems, programming languages, data management, or networking principles to computational assignments.
3. An ability to apply design, development, maintenance or evaluation of software engineering principles in the construction of computer and software systems of varying complexity and quality.
4. An ability to understand concepts involved in modeling and design of computer science applications in a way that demonstrates comprehension of the fundamentals and tradeoffs involved in design choices.

Course Outcomes (COs)

C418.1 **Presentation Skills in terms of Content** : Students will be able to show competence in identifying relevant information, defining and explaining topics under discussion. They will demonstrate depth of understanding, use primary and secondary sources; they will demonstrate the working, complexity, insight, cogency, independent thought, relevance, and persuasiveness. They will be able to evaluate information and use and apply relevant theories.

C418.2 **Presentation Skills in terms of Organization** : Students will be able to show competence in working with a methodology, structuring their oral work, and synthesizing information. They will make a detailed study on the previous works related to their topic and will present the observations.

C418.3 **Presentation Skills in terms of Delivery** : Students will use appropriate registers and vocabulary, and will demonstrate command of voice modulation, voice projection, and pacing. They will be able to make use of visual, audio and audio-visual material to support their presentation, and will be able to speak cogently with or without notes.

C418.4 **Discussion Skills** : Students will be able to judge when to speak and how much to say, speak clearly and audibly in a manner appropriate to the subject, ask appropriate questions, use evidence to support claims, respond to a range of questions, take part in meaningful discussion to reach a shared understanding, speak with or without notes, show depth of understanding.

C418.5 **Listening Skills** : Students will demonstrate that they have paid close attention to what others say and can respond constructively. Through listening attentively, they will be able to build on discussion fruitfully, supporting and connecting with other discussants.

C418.6 **Argumentative Skills and Critical Thinking** : Students will develop persuasive speech, present information in a compelling, well-structured, and logical sequence, respond respectfully to opposing ideas, show depth of knowledge of complex subjects, and develop their ability to synthesize, evaluate and reflect on information.

Course Outcome							
Programme Outcomes		C418.1	C418.2	C418.3	C418.4	C418.5	C418.6
	1	3	3	3	3	3	3
	2	3	3	3	3	3	3
	3	3	3	3	3	3	3
	4	3	3	3	3	3	3
	5	3	3	3	3	3	3
	6	3	3	3	3	3	3
	7	3	3	3	3	3	3
	8	3	3	3	3	3	3
	9	3	3	3	3	3	3
	10	3	3	3	3	3	3
	11	3	3	3	3	3	3
	12	3	3	3	3	3	3

PO - CO Mapping

PEO - CO Mapping

Course Outcome							
Programme Educational Objective		C418.1	C418.2	C418.3	C418.4	C418.5	C418.6
	1	3	3	1	1	-	2
	2	3	3	3	3	1	3
	3	1	2	3	3	1	3

PSO - CO Mapping

Course Outcome							
Programme Specific Outcomes		C418.1	C418.2	C418.3	C418.4	C418.5	C418.6
	1	3	3	3	3	3	3
	2	3	3	3	3	3	3
	3	3	3	3	3	3	3
	4	3	3	3	3	3	3

Seminar Outcome

1. Studied about the concept of Deep Learning.
2. Studied about different neural networks.
3. Analyzed and compared the general architecture of CNN,RNN and CRNN.
4. Studied about different deep fake detection methods.

Seminar Outcome - CO Mapping

	Course Outcome						
Seminar Outcome		C418.1	C418.2	C418.3	C418.4	C418.5	C418.6
	1	3	3	3	1	3	3
	2	3	3	1	1	3	3
	3	3	3	3	1	3	1
	4	3	3	3	3	1	1
	5	3	1	3	3	1	1

ACKNOWLEDGEMENT

I take this opportunity to express my heartfelt gratitude to all respected personalities who had guided, inspired and helped me in the successful completion of this seminar. First and foremost, I express my thanks to **The Lord Almighty** for guiding me in this endeavour and making it a success.

I take immense pleasure in thanking the **Management** of Jyothi Engineering College and **Dr. Sunny Joseph Kalayathankal**, principal, Jyothi Engineering College for having permitted me to carry out this seminar. My sincere thanks to **Dr. Vinith R**, Head of the Department of Computer Science and Engineering for permitting me to make use of the facilities available in the department to carry out the seminar successfully.

I express my sincere gratitude to **Mr. Shaiju Paul & Dr. Swapna B Sasi**, Seminar Coordinators for their invaluable supervision and timely suggestions. I am very happy to express my deepest gratitude to my mentor **Ms. Aswathy Wilson**, Assistant Professor,Department of Computer Science and Engineering, Jyothi Engineering College for her able guidance and continuous encouragement.

Last but not least, I extend my gratefulness to all teaching and non-teaching staff who directly or indirectly involved in the successful completion of this seminar work and to all friends who have patiently extended all sorts of help for accomplishing this undertaking.

ABSTRACT

Deep fake videos are AI-generated videos that look real but are actually fake.Deep fake videos are generally created by face-swapping techniques. It started out as fun but like any technology, it is being misused. In the beginning, these videos could be identified by human eyes. But due to the development of machine learning, it became easier to create deep fake videos. It has almost become indistinguishable from real videos.Deep fake videos are usually created by using GANs (Generative Adversarial Network) and other deep learning technologies. The danger of this is that technology can be used to make people believe something is real when it is not. Smartphone desktop applications like FaceApp and Fake App are built on this process.These videos can affect a person's integrity.

So identifying and categorizing these videos has become a necessity. In this paper, we will deliberate about the different methods for detecting Deep Fake Videos. Hopefully, we will be able to make the internet a safer place.

Keywords - Detection, Identification, Convolutional Recurrent Neural Networks, Recurrent Neural Networks, Convolutional neural networks.

CONTENTS

List of Figures

List of Abbreviations

CNN	: *Convolutional Neural Network*
RNN	: *Recurrent Neural Network*
CRNN	: *Convolutional Recurrent Neural Network*
ROC	: *Receiver Operating Characteristic*
MTCNN	: *Multi Task Cascaded Convolutional Neural Networks*
DF	: *Deep Fake*
AI	: *Artificial intelligence*
ML	: *Machine Learning*
DL	: *Deep Learning*
GAN	: *Generative Adversarial Network*
ROI	: *Region Of Interest*
FER	: *Facial Expression Recognition*

CHAPTER 1
INTRODUCTION

1.1 Overview

Deep Fakes are AI generated fake videos that look real but are actually fake.The word deepfake combines the terms “deep learning” and “fake,” and is a form of artificial intelligence. Deepfakes are falsified videos made by means of deep learning. Deep learning is “a subset of AI,” and refers to arrangements of algorithms that can learn and make intelligent decisions on their own. But the danger of that is the technology can be used to make people believe something is real when it is not.A deep-learning system can produce a persuasive counterfeit by studying photographs and videos of a target person from multiple angles, and then mimicking its behavior and speech patterns. Once a preliminary fake has been produced, a method known as GANs, or generative adversarial networks, makes it more believable. The GANs process seeks to detect flaws in the forgery, leading to improvements addressing the flaws.

It started out as fun but like any technology, it is being misused. In the beginning, these videos could be identified by human eyes. But due to the development of machine learning, it became easier to create deep fake videos. It has almost become indistinguishable from real videos. Smartphone desktop applications like FaceApp and Fake App are built on this process. These videos can affect a person’s integrity. It became necessary to identify these fake videos from real videos. While AI can be used to make deepfakes, it can also be used to detect them. In this paper I discuss some of the methods used in deepfake detection.

1.1.1 Artificial Intelligence

Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving.The ideal characteristic of artificial intelligence is its ability to rationalize and take actions that have the best chance of achieving a specific goal.

Artificial intelligence is based on the principle that human intelligence can be defined in

a way that a machine can easily mimic it and execute tasks, from the most simple to those that are even more complex. The goals of artificial intelligence include learning, reasoning, and perception.

As technology advances, previous benchmarks that defined artificial intelligence become outdated. For example, machines that calculate basic functions or recognize text through optimal character recognition are no longer considered to embody artificial intelligence, since this function is now taken for granted as an inherent computer function.

AI is continuously evolving to benefit many different industries. Machines are wired using a cross-disciplinary approach based in mathematics, computer science, linguistics, psychology, and more.

Types of AI:

- Reactive Machines AI: Based on present actions, it cannot use previous experiences to form current decisions and simultaneously update their memory. Example: Deep Blue
- Limited Memory AI: Used in self-driving cars. They detect the movement of vehicles around them constantly and add it to their memory.
- Theory of Mind AI: Advanced AI that has the ability to understand emotions, people and other things in the real world.
- Self Aware AI: AIs that posses human-like consciousness and reactions. Such machines have the ability to form self-driven actions.
- Artificial Narrow Intelligence (ANI): General purpose AI, used in building virtual assistants like Siri
- Artificial General Intelligence (AGI): Also known as strong AI. An example is the Pillo robot that answers questions related to health.
- AI that possesses the ability to do everything that a human can do and more. An example

is the Alpha 2 which is the first humanoid ASI robot.

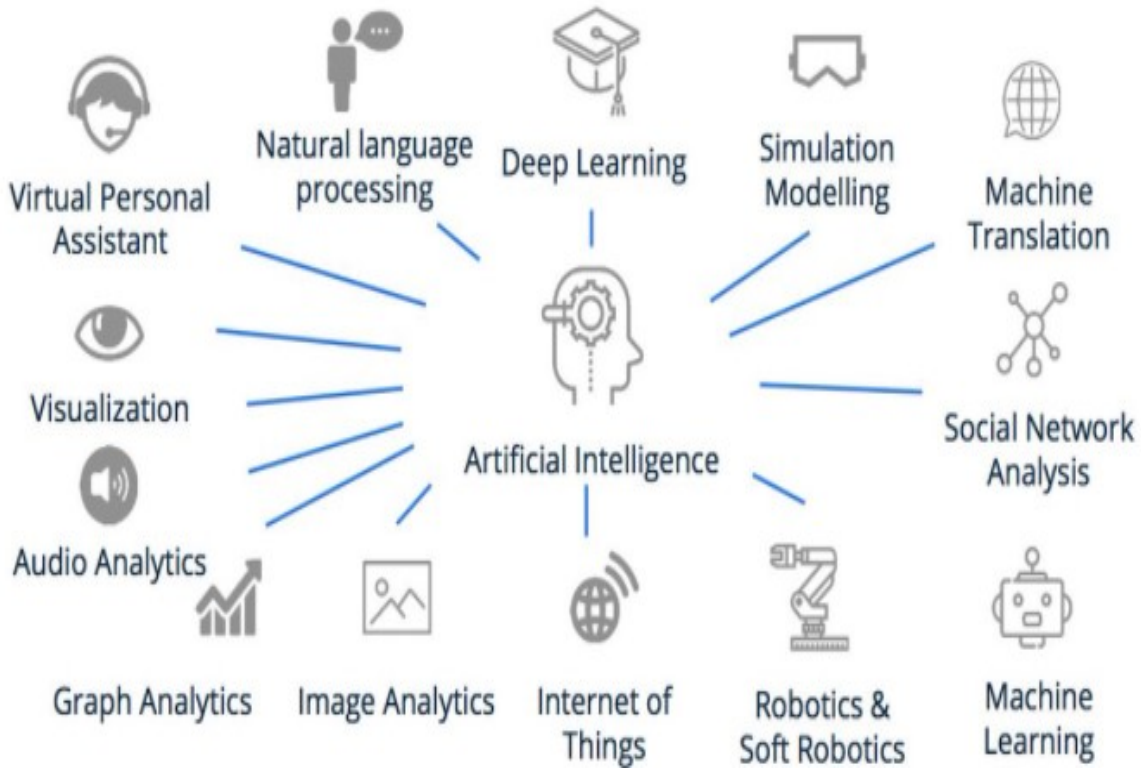

Figure 1.1: Artificial intelligence

1.1.2 Machine Learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

Machine learning algorithms are often categorized as supervised or unsupervised. Su-

pervised machine learning algorithms can apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training data set, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

In contrast, unsupervised machine learning algorithms are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from data sets to describe hidden structures from unlabeled data.

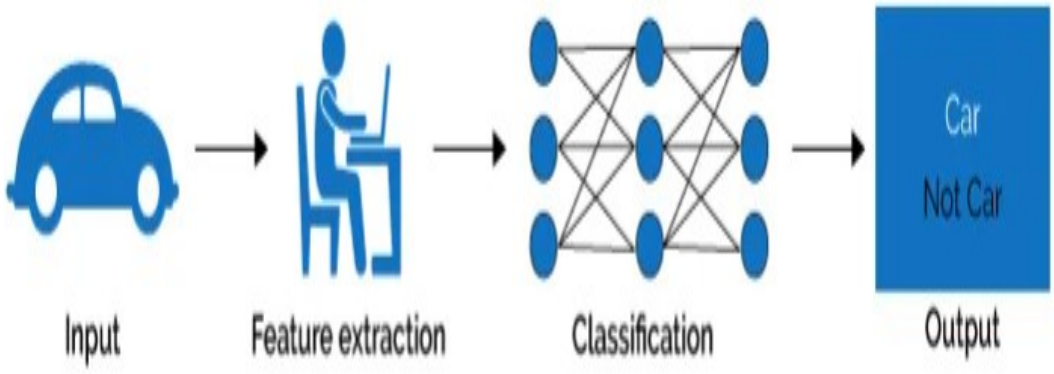

Figure 1.2: Machine Learning

1.1.3 Deep Learning

Deep learning is an artificial intelligence (AI) function that imitates the workings of the human brain in processing data and creating patterns for use in decision making. Deep learning is a subset of machine learning in artificial intelligence that has networks capable of learning unsupervised from data that is unstructured or unlabeled.

Deep learning has evolved hand-in-hand with the digital era, which has brought about an explosion of data in all forms and from every region of the world. This data, known simply as big data, is drawn from sources like social media, internet search engines, e-commerce platforms, and online cinemas, among others. This enormous amount of data is readily accessible and can be shared through fintech applications like cloud computing.

However, the data, which normally is unstructured, is so vast that it could take decades for humans to comprehend it and extract relevant information. Companies realize the incredible potential that can result from unraveling this wealth of information and are increasingly adapting to AI systems for automated support

In deep learning, each level learns to transform its input data into a slightly more abstract and composite representation. In an image recognition application, the raw input may be a matrix of pixels; the first representational layer may abstract the pixels and encode edges; the second layer may compose and encode arrangements of edges; the third layer may encode a nose and eyes; and the fourth layer may recognize that the image contains a face. Importantly, a deep learning process can learn which features to optimally place in which level on its own.

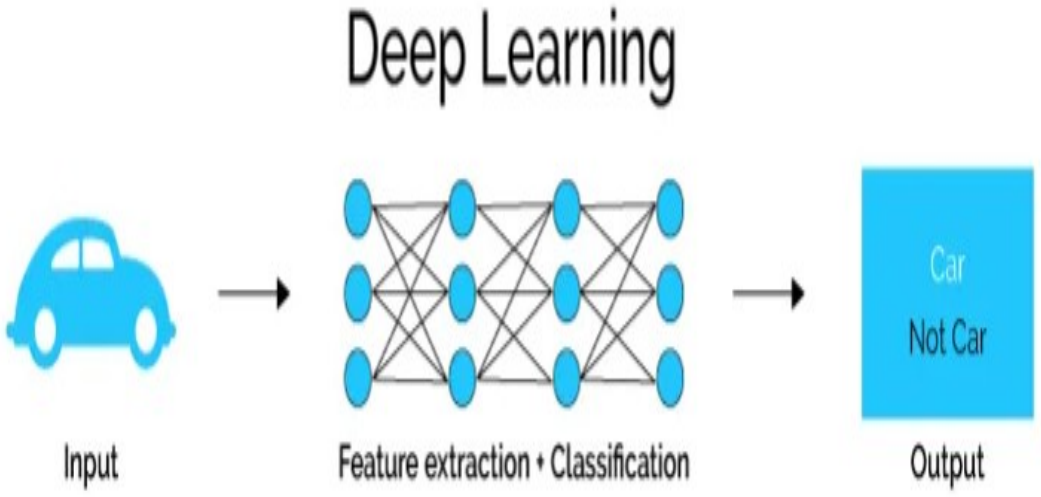

Figure 1.3: Deep Learning

1.1.4 Artificial Neural Networks

An artificial neural network (ANN) is the piece of a computing system designed to simulate the way the human brain analyzes and processes information. It is the foundation of artificial intelligence (AI) and solves problems that would prove impossible or difficult by human or statistical standards. ANNs have self-learning capabilities that enable them to produce better results as more data becomes available. [1]

An ANN has hundreds or thousands of artificial neurons called processing units, which are interconnected by nodes. These processing units are made up of input and output units. The input units receive various forms and structures of information based on an internal weighting system, and the neural network attempts to learn about the information presented to produce one output report

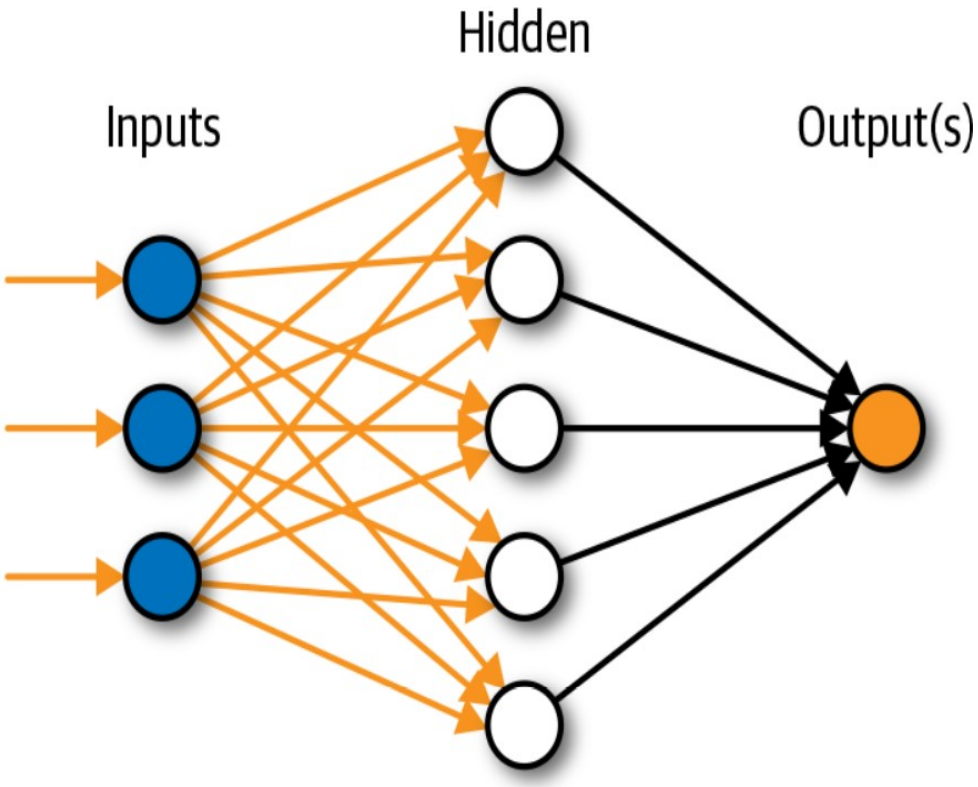

Figure 1.4: Artificial Neural Networks

1.1.5 Recurrent Neural Networks

Recurrent Neural Network(RNN) are a type of Neural Network where the output from previous step are fed as input to the current step. In traditional neural networks, all the inputs and outputs are independent of each other, but in cases like when it is required to predict the next word of a sentence, the previous words are required and hence there is a need to remember the previous words. Thus RNN came into existence, which solved this issue with the help of a Hidden Layer. The main and most important feature of RNN is Hidden state, which remembers some information about a sequence.[8]

RNN have a "memory" which remembers all information about what has been calculated. It uses the same parameters for each input as it performs the same task on all the inputs or hidden layers to produce the output. This reduces the complexity of parameters, unlike other neural networks.

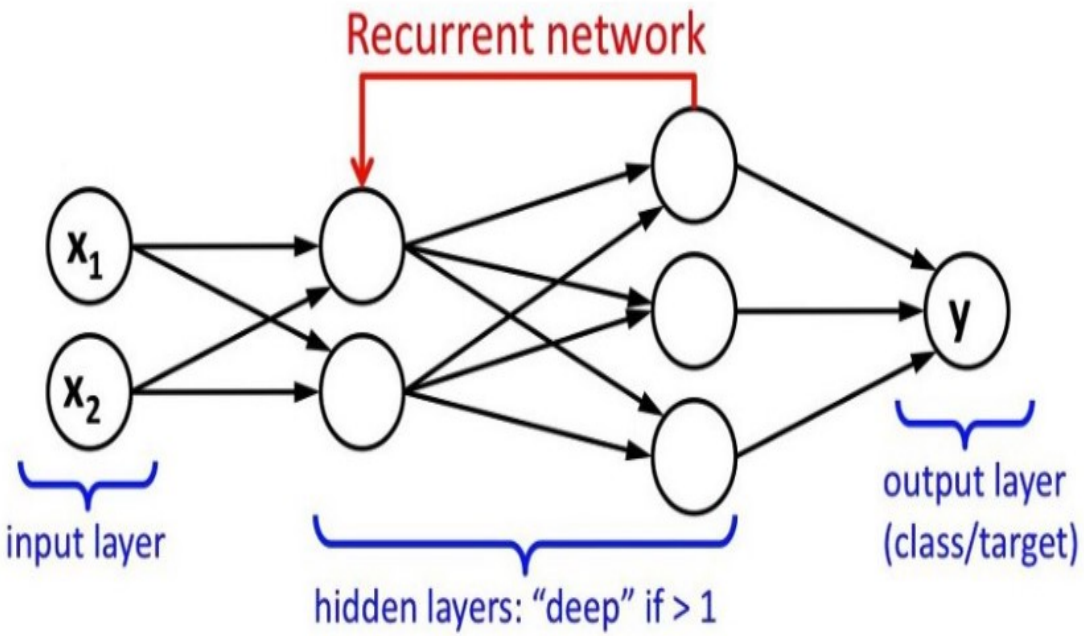

Figure 1.5: Recurrent Neural Networks

1.1.6 Convolutional Neural Networks

A convolutional neural network (CNN) is a specific type of artificial neural network that uses perceptrons, a machine learning unit algorithm, for supervised learning, to analyze data.

CNNs apply to image processing, natural language processing and other kinds of cognitive tasks.[3]

CNNs are a fundamental example of deep learning, where a more sophisticated model pushes the evolution of artificial intelligence by offering systems that simulate different types of biological human brain activity.

CNNs are regularized versions of multilayer perceptrons. Multilayer perceptrons usually mean fully connected networks, that is, each neuron in one layer is connected to all neurons in the next layer. The "fully-connectedness" of these networks makes them prone to overfitting data. Typical ways of regularization include adding some form of magnitude measurement of weights to the loss function. CNNs take a different approach towards regularization: they take advantage of the hierarchical pattern in data and assemble more complex patterns using smaller and simpler patterns. Therefore, on the scale of connectedness and complexity, CNNs are on the lower extreme.

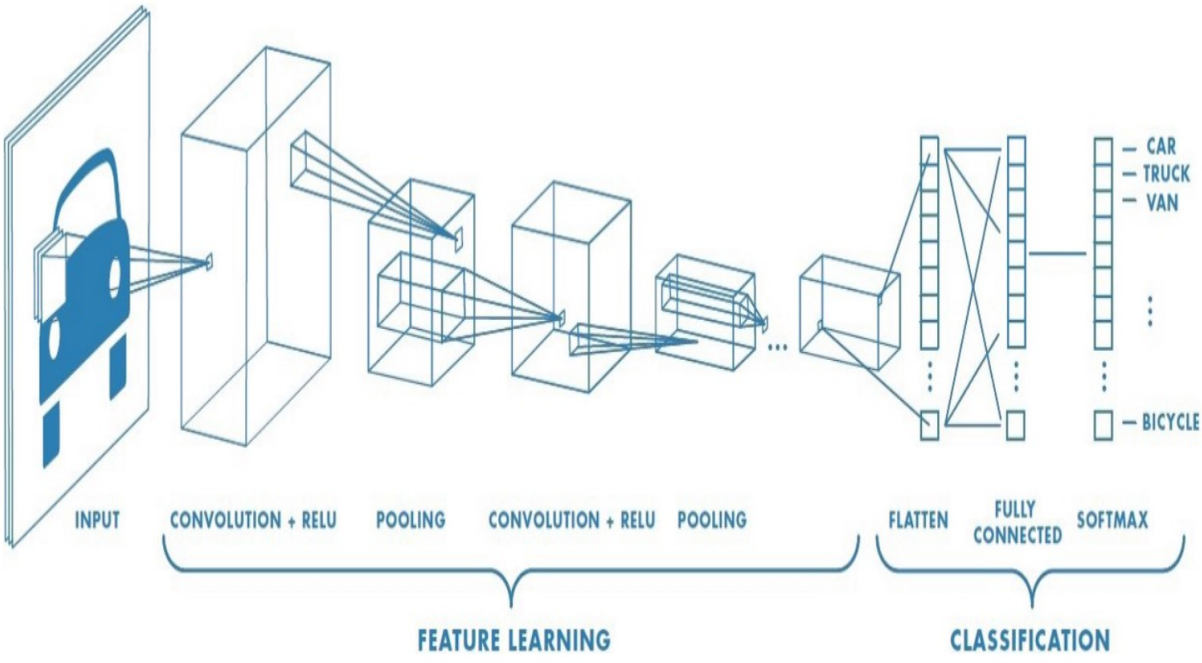

Figure 1.6: Convolutional Neural Networks

1.1.7 DeepFakes

Deepfakes are synthetic media in which a person in an existing image or video is replaced with someone else's likeness. While the act of faking content is not new, deepfakes leverage powerful techniques from machine learning and artificial intelligence to manipulate or generate visual and audio content with a high potential to deceive. [6]

Deepfakes extend the idea of video (or movie) compositing, which has been done for decades. Significant video skills, time, and equipment go into video compositing; video deepfakes require much less skill, time and equipment, although they are often unconvincing to careful observers

The main machine learning methods used to create deepfakes are based on deep learning and involve training generative neural network architectures, such as autoencoders or generative adversarial networks (GANs).

- Autoencoders

  Essentially, autoencoders for deepfake faces in images run a two-step process. Step one is to use a neural network to extract a face from a source image and encode that into a set of features and possibly a mask, typically using several 2D convolution layers, a couple of dense layers, and a softmax layer. Step two is to use another neural network to decode the features, upscale the generated face, rotate and scale the face as needed, and apply the upscaled face to another image. Training an autoencoder for deepfake face generation requires a lot of images of the source and target faces from multiple points of view and in varied lighting conditions.

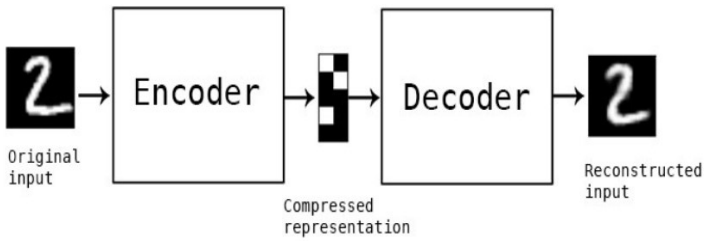

Figure 1.7: Autoencoders example

- GANs

Generative adversarial networks can refine the results of autoencoders, for example, by pitting two neural networks against each other. The generative network tries to create examples that have the same statistics as the original, while the discriminative network tries to detect deviations from the original data distribution.

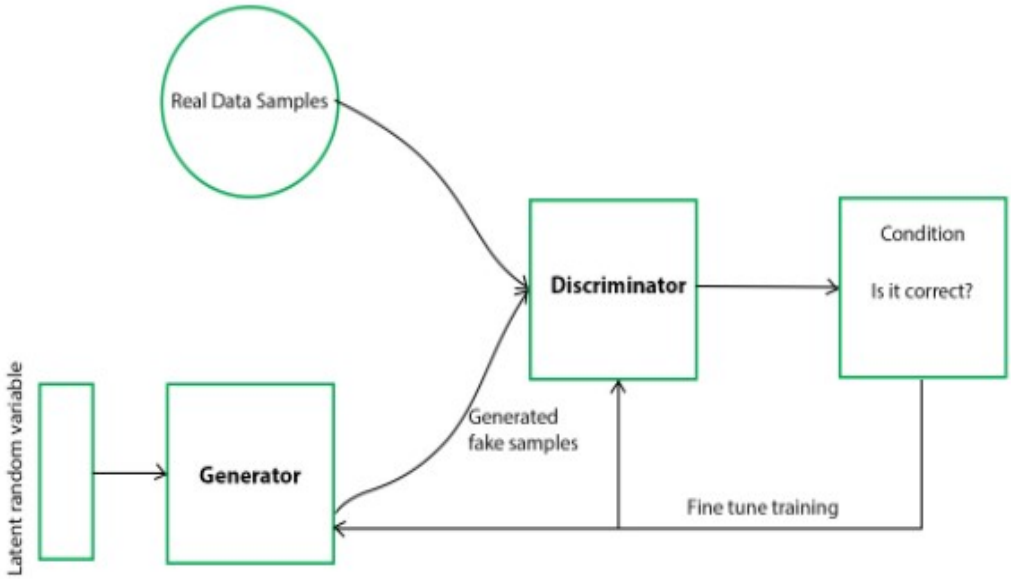

Figure 1.8: Generative Adversarial Networks

1.2 Objective

The main objective of this seminar is to introduce different methods used to identify deepfake videos and classify them as real and fake videos based on their features using deep learning.These methods identify fake videos from real videos and there by prevent the usage of these videos in creating political distress, blackmailing, fake terrorism events, etc.

1.3 Organization Of The Report

The report is organised as follow:

- **Chapter 1:Introduction**
  Gives an introduction about DeepFakes

- **Chapter 2:Literature Survey**
  Explains different DeepFake detection methods

- **Chapter 3: Detecting Deepfakes With Metric Learning**

- **Chapter 4:Effective and Fast Deepfake Detection Method Using Haar Wavelet Transform**

- **Chapter 5:Deepfake Video Detection Using Recurrent Neural Network**

- **Chapter 6:Conclusion**
  summarizes the different DeepFake detection techniques that were discussed

- **References**
  Reference papers are included for future purposes

CHAPTER 2

LITERATURE SURVEY

2.1 Joint Face detection and Facial Expression Recognition with MTCNN

MTCNN or Multi-Task Cascaded Convolutional Neural Networks is a neural network which detects faces and facial landmarks on images. Humans interact with each other mainly through speech,but also through body gestures, to emphasize certain parts of their speech and to display emotions. One of the important ways humans display emotions is through facial expressions which is key to nonverbal communication between humans. Computers and other electronic devices in our daily lives will become more user-friendly if they can adequately interpret a person's facial expressions, thereby improving human-machine interfaces. Facial expression recognition can be implemented in all computer interfaces.Convolutional Neural Networks (CNNs) have gained much popularity in recent years for vision-related applications and have the potential to achieve some of the higher accuracies in FER. However, most of the available face detection and facial expression recognition methods ignore the inherent correlation between these two tasks., but there are several existed works attempt to jointly solve face detection and face alignment.

In this paper, they exploited the inherent correlation between face detection and facial expression recognition based MTCNN and implement the two tasks. The input into our system is an image; then, we use this framework to predict the face location and facial expression label which should be one these labels: anger, happiness, fear, sadness, disgust and neutral, and compare to a variety of other recent high performing detectors.[9]

2.1.1 Overview

The MTCNN, has inherited correlation between the face detection and alignment to boost up their performance. It essentially consists of three parts: (1) a Proposal Network (P-Net) for generating a list of the candidate windows. Then, it is used for classifying face and non-face and estimate bounding box regression vectors to face location, and non-maximum suppression

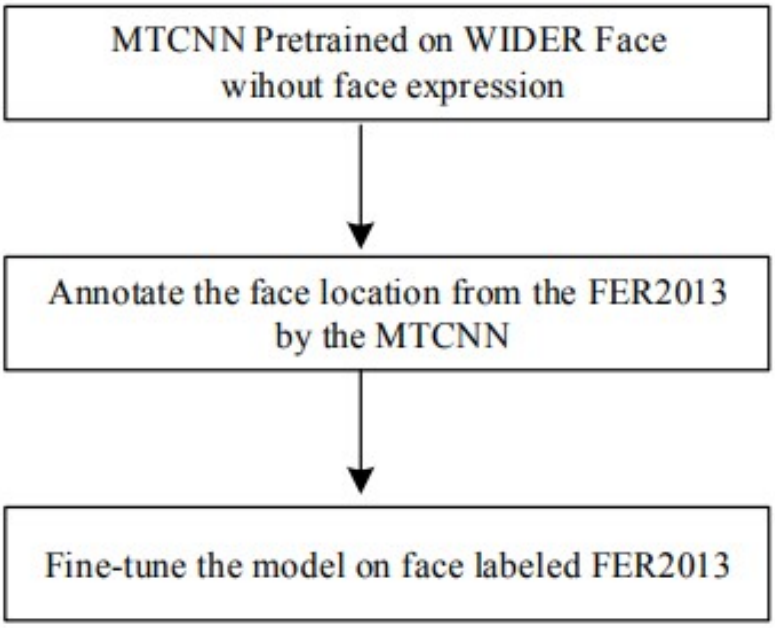

Figure 2.1: flow chart of training procedure

(NMS) candidate merge. (2) a Refine Network (R-Net), unlike P-Net, it doesn't obtain the region proposals but rejects a lot of false candidates. (3) It is similar to R-Net, called O-Net. In this network, it will output five facial landmarks' positions

2.1.2 Training

.Our method follows the similar deep learning framework of MTCNN, It is similar to R-Net, called O-Net. In this network, we aim to classify the emotion of facial expressions. According we propose to modify the MTCNN architecture for facial expression recognition and train our facial expression recognition model by following the proposed procedure as shown in Fig. 2.1

2.1.3 Experiments

Network is trained and tested using the Caffe open source framework for deep convolutional neural networks and reported experiments on comparison of facial expression recognition and also on performance of top face detectors

2.2 Deepfake Video Detection through Optical Flow based CNN

Recent advances in visual media technology have led to new tools for processing and, above all, generating multimedia contents. In particular, modern AI-based technologies have provided easy-to-use tools to create extremely realistic manipulated videos.Such synthetic videos, named Deep Fakes, may constitute a serious threat to attack the reputation of public subjects or to address the general opinion on a certain event.Deep learning techniques are escalating technology sophistication regarding creation and processing of multimedia contents. A new phenomenon, known as Deep Fakes (DF), has recently emerged: it permits to quite simply create realistic videos where people faces, or sometimes only lips and eyes movements, are modified in order to likely simulate the presence of another subject in a certain context or to make someone speak coherently with a different and, probably compromising, speech. The effects can be straightforwardly imagined when this fake information is deliberately used to harm a person such a public figure or a politician, or even an organization like a political party. The impact of Deep Fakes can also be amplified by the action of social networks that deliver information quickly and worldwide.

In this work, a new forensic technique able to discern between fake and original video sequences is given; unlike other state-of-the-art methods which resorts at single video frames, we propose the adoption of optical flow fields to exploit possible inter-frame dissimilarities. This paper present a sequence-based approach dedicated to investigate possible dissimilarities in the temporal structure of a video. [2]

2.2.1 Method

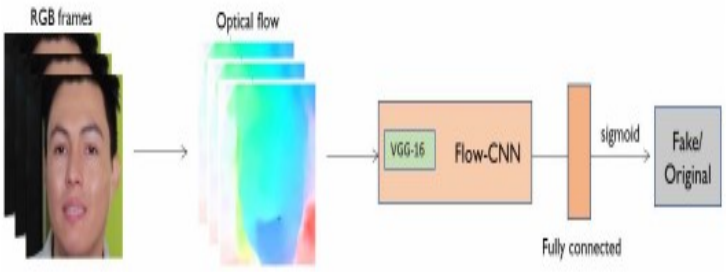

Figure 2.2: Proposed architecture

Basic architecture of this method is given in fig 2.2. A structure has been built up to understand the actual effectiveness of optical flow fields to distinguish a deepfake from an original video. Optical flow is a vector field which is computed on two consecutive frame f(t) and f(t + 1) to extract apparent motion between the observer and the scene itself.The hypothesis is that the optical flow is able to exploit discrepancies in motion across frames synthetically created with respect to those naturally generated by a video camera.It should be more appreciable in the optical flow matrices, the introduction of fake and unusual movements of the lips, eyes and in general of the whole face. So, for this reason, for each frame f(t), at a certain time t, a forward flow OF(f(t), f(t + 1)) is extracted using the CNN model for optical flow called PWC-Net . This technique is based on pyramidal processing and warping and on the use of a cost volume processed by the CNN itself to estimate the optical flow.Successively, the computed forward flow OF(f(t), f(t + 1)) is given as input to a semi-trainable CNN named Flow-CNN, based on some pre-trained network.

2.3 Deepfake Detection with Clustering-based Embedding Regularization

In recent months, AI-synthesized face swapping videos referred to as deepfake have become an emerging problem. False video is becoming more and more difficult to distinguish, which brings a series of challenges to social security. Some scholars are devoted to studying how to improve the detection accuracy of deepfake video.At present, intelligent video recognition based on deep learning has been applied to various fields of the country and society, such as network content supervision, intelligent video surveillance and autonomous vehicles. People post photos and videos every day on popular social software and websites such Weibo, Facebook, Instagram and Twitter. However, manipulation of visual content has now become ubiquitous. In the past two years, AI-algorithms for face swap and many other video manipulation methods have been proposed.Many such tools are publicly available as open source software, such as DeepFake, Face2Face, FaceSwap, FakeApp.r. The continuous advancement of video tampering technology and the improvement of video quality have brought great challenges to deepfake detection.

In this work, we propose a new method with clustering based embedding regularization for deepfake detection. They use open source algorithms to simulate the process of the process of generating artifacts during the deepfake generation process using simple image processing operations on a image, and make it as a generated example. We train the Xception network for classification, using positive samples, negative samples, and generated samples as input samples. Class number is set to 3 during training process, and in the testing process the

generated samples are also classified as negative samples to improve the classification effect. At the same time, a regularization loss is added during the training process to ensure the inter-class distance and intra-class smoothness of the embedded space.[12]

2.3.1 Experiment

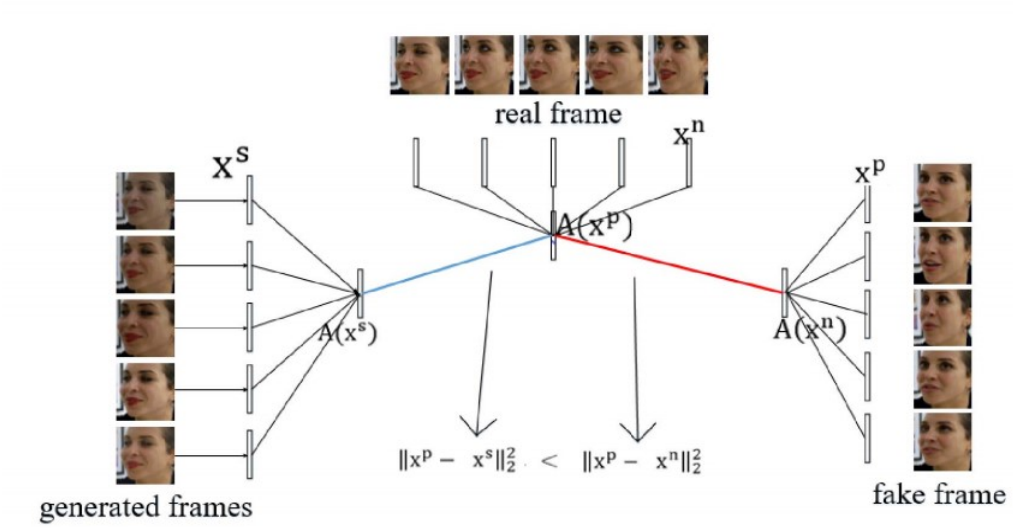

Figure 2.3: The training process of a batch size sample

for positive and provided fake samples, we first extract each frame of the videos, and then perform face recognition on each frame and resize it to 224 * 224. Then we use the generation algorithm to get simulated samples. In order to ensure the randomness and diversity of the samples, we add a random size to the identified face rectangle, and then also resize it to 224 * 224 pixel. Generate the same size simulation sample 5 times for each picture as described above. We choose the Xception networks for training. Each iteration takes the same size of positive samples, negative samples, and simulated samples randomly. We selected three deepfake datasets, including UADFV, Celeb-DF, DeepFakeDetection.

2.4 Deepfake Video Detection Using Recurrent Neural Networks

In recent months a machine learning based free softwaretool has made it easy to create believable face swaps invideos that leaves few traces of manipulation, in what areknown as “deepfake” videos.Scenarios where these real-istic fake videos are used to create political distress, black-mail someone or fake terrorism events are easily envisioned.This paper proposes a

temporal-aware pipeline to automat-ically detect deepfake videos. Our system uses a convolutional neural network (CNN) to extract frame-level features.These features are then used to train a recurrent neural net-work (RNN) that learns to classify if a video has been sub-ject to manipulation or not. We evaluate our method againsta large set of deepfake videos collected from multiple video websites.

Proposed system is composed of a convolutional LSTM structure for processing frame sequence 2 essential components in a convolutional LSTM: CNN for frame feature extraction and LSTM for temporal sequence analysis. CNN generates a set of features for each frame from an unseen test sequence Features of multiple consecutive frames are concatenated and passed on to the LSTM for analysis. Finally an estimate of the likelihood of sequence being either a deep fake or real video is obtained[4]

2.4.1 Architecture

Dataset is selected with 50% of deep fake videos and 50% of real videos. Random split is used to generate 3 disjoint set which is used for training, validation and testing. Balanced splitting on real videos and fake videos are done. This made sure that final set has exactly 50% of each class. Resize every frame to 299 X 299. Length of input sequence is controlled using subsequence sampling. This allows us to identify how many frames are necessary per video to have an accurate detection. Optimizer is set for end to end training of the complete model.

2.5 Detecting Deepfakes with Metric Learning

With the arrival of several face-swapping applica-tions such asFaceApp, SnapChat, MixBooth, FaceBlenderandmany more, the authenticity of digital media content is hangingon a very loose thread. On social media platforms, videos arewidely circulated often at a high compression factor. In thiswork, we analyze several deep learning approaches in the contextof deepfakes classification in high compression scenarios anddemonstrate that a proposed approach based on metric learningcan be very effective in performing such a classification.Metric learning is an approach based directly on a distance metric that aims to establish similarity or dissimilarity between objects. While metric learning aims to reduce the distance between similar objects, it also aims to increase the distance between dissimilar objects.Metric learning can be very effective in classification of deep fakes.It learns to enhance the feature space distance between the cluster of real and fake videos embedding vectors.We use Multitask Cascaded CNNs (MTCNN) to extract faces out of frames MTCNN is a neural network which detects

faces and facial landmarks on images.[7]

2.5.1 Architecture

Final architecture uses triplet network to discriminate between fake and real videos. MTCNN extracts faces from frames. Facenet generates 512 dimension embeddings for each face in the feature space. As facenet is developed for face recognition, each unique face occupies a small cluster in the feature space. Using Semi-Hard triplets, the embeddings of fake frames are distinctly separated through triplet loss.

2.6 Effective and Fast DeepFake Detection Method Based on Haar Wavelet Transform

DeepFake using Generative Adversarial Networks (GANs) tampered videos reveals a new challenge in today's life. With the inception of GANs, generating high-quality fake videos becomes much easier and in a very realistic manner. Therefore, the development of efficient tools that can automatically detect these fake videos is of paramount importance.Haar wavelet is a sequence of rescaled "square-shaped" functions which together form a wavelet family. This method takes advantage of the fact that deep fake algorithms are only able to generate fake faces with specific size resolution. In order to fit the source face in the original video, a blur function must be added. This leaves traces due to the blur inconsistency between the transformed region (ROI) the surrounding area. This method detects inconsistency by comparing the blurred area of ROI the surrounding context with a dedicated Haar wavelet transform function.[11]

2.6.1 Architecture

Linear blurred images can be described as G = H * F + N. Where G, F N represent noisy image and Blur function is represented by H matrix. The blur extent is measured by testing the edge features in an image. If blur is present, both the edge sharpness its type will be changed. This will indicate whether the image has been manipulated or not.Different types of edges are present in an image. Blur extent is identified by taking sharpness of Roof-structure G Step-Structure.Sharpness of edge is indicated by parameter alpha(α).

If alpha is large, it means the edge is sharper. Blurred images will not have A Step-Structure or Dirac-Structure. Percentage of G Step-Structure Roof-Structure is used to set the

value of blur extent. By comparing blur extent of ROI with blur extent of rest of image, we can find whether the image is manipulated or not.

2.7 Literature Survey Conclusions

2.7.1 Joint Face detection and Facial Expression Recognition with MTCNN

- The final validation accuracy obtained is 60.7%.

- Introduced a new method for face detection and facial expression recognition using deep learning techniques.

2.7.2 Deepfake Video Detection through Optical Flow based CNN

- The final validation accuracy obtained is 81.61%.

- The idea to exploit optical flow field dissimilarities as a clue to discriminate between deepfake videos and original ones has been introduced and investigated.

2.7.3 Deepfake Detection with Clustering-based Embedding Regularization

- The final validation accuracy obtained is 98.43%.

- Proposed a new method with clustering based embedding regularization for deepfake detection.

- This model achieved good results on UADFV, Celeb-DF and DeepFake Detection datasets including low-quality videos and high quality videos.

2.7.4 Deepfake Video Detection Using Recurrent Neural Networks

- Extracted features from frame using CNN.

- LSTM is used process the sequence generated by CNN.

- Using convolutional LSTM, we can accurately identify whether a video has been manipulated or not.

2.7.5 Detecting Deepfakes with Metric Learning

- Extracted faces from video frames.
- Used triplet network to discriminate between fake and real videos FaceNet is used to generate face embeddings.
- Using semi-hard triplet networks, embeddings of fake frames are distinctly separated through triplet loss.

2.7.6 Effective and Fast DeepFake Detection Method Based on Haar Wavelet Transform

- Based on the fact that the DeepFake method can only produce face images of limited resolutions and fixed size .
- These images then need to be further blurred and transformed to match the faces in the original video.
- This added blur and transformations on the ROI leave special characteristics in the resulting DeepFake videos.
- This can be effectively captured by detecting the differences between ROI and the rest of the image using Haar Wavelet transformation.

CHAPTER 3

Detecting Deepfakes with Metric Learning

With the arrival of several face-swapping applications such as FaceApp, SnapChat, MixBooth, FaceBlender and many more, the authenticity of digital media content is hanging on a very loose thread. On social media platforms, videos are widely circulated often at a high compression factor.With the rapid increase of online streaming platforms, there is a dire need to check the authenticity of the videos. In Youtube alone, 300 hours of videos are uploaded every minute.On a daily basis, 5 billion videos are watched and 1 billion hours are streamed, that's Facebook and Netflix streaming combined. The rise of deepfakes in recent years seriously raises concerns about the authenticity of digital content by media and other online streaming platforms. Generative architectures are excellent for aiding in boosting the performance of deep learning architectures by satisfying the need for large datasets, and in general to explore the creative power of deep learning. However, such approaches have also resulted in Deepfakes, which are now been utilized for nefarious purposes to manipulate the images of politicians, famous actors, etc.Many politicians and actors are becoming victims of Deep-fakes. For criminal purposes, forensic videos are altered using novel methods such as faceswap and faceswap-GAN.[7]

In this work, we analyze several deep learning approaches in the context of deepfakes classification in high compression scenarios and demonstrate that a proposed approach based on metric learning can be very effective in performing such a classification. Using less number of frames per video to assess its realism, the metric learning approach using a triplet network architecture proves to be fruitful. It learns to enhance the feature space distance between the cluster of real and fake videos embedding vectors.Metric learning is an approach based directly on a distance metric that aims to establish similarity or dissimilarity between objects. While metric learning aims to reduce the distance between similar objects, it also aims to increase the distance between dissimilar objects. Metric learning can be very effective in classification of deep fakes. It learns to enhance the feature space distance between the cluster of real and fake videos embedding vectors. We use Multitask Cascaded CNNs (MTCNN) to extract faces out of frames. MTCNN is a neural network which detects faces and facial landmarks on images.

3.0.1 Methodology

We use MTCNN to extract faces out of frames.Based on the success of detection of fake and real videos of XceptionNet architecture from FF++paper, we started off with it for video classification. To combat the classification in low-resolution videos, we further analyzed several methods using recurrent neural networks, convolutional 3D networks, and, then finally metric learning approach. Our architecture and methods involved are discussed in the following subsections.

MTCNN

One can crop out images using Proposal,Refine and Output Networks. The proposal network detects faces across multiple resolutions, following which, the refine network suppress the overlapping boxes using nonmax suppression.Finally, the output network gives the bounded face using five landmarks. [10]

Figure 3.1: Extraction of face from frame using MTCNN algorithm.

Transfer Learning

To make use of previous knowledge of architecture from one problem onto another problem is known as transfer learning. In this work, we used Xception architecture to learn the crucial feature of real and fake faces. Xceptionnet based on Inception V3 uses the Inception module, with modification of the spatial convolutions to depth wise separable convolutions. After separating each channel, 1x1 depth wise convolutions help network to capture the cross-channel cor-relations. Compared to Inception architecture convolutions,depth wise separable convolution differs in two ways: 1)Xception modules perform channel-wise convolutions first,then, 1x1 convolution, compared to Inception where the 1x1is performed earlier, and, 2) There's no non-linearity after depth wise separable convolutions. With this, the number of layers is reduced from 159 layers in Inception V3 to 126layers in Xception architecture. Although Xception Net was quite successful in capturing dissimilarities in high-resolution images, the performance drops significantly for low resolution compressed images.

Sequence Classification

Recurrent Neural networks capture the information along the temporal domain. An output vector from the previous step is fed into the next step to learn the relation of features across time domain. Their ability to connect a sequence of input frames over a period of time makes them significantly helpful for video classification purposes. LSTM supersedes the RNN as it can retain information for a long sequence of frames.we used a sequence of 16 and 32 frames per video to learn the inconsistency across the temporal domain. Our architecture contains a single layer LSTM of 32 units and then a prediction layer.

3D Convolution

3D convolution model employs 3D filters that pick up the knowledge of spatio temporal features from the videos,in contrast to 2D convolution, where the temporal domain is collapsed. In deepfakes, while transferring the appearance to target candidate, if the target candidate has a pose that's not happened in source candidate video then there's a discrepancy.To capture the spatial and temporal irregularities, we took 32frames per video into consideration. The architecture com-prises of 3x3x3 filters in convolution layers and 2x2x2 poolsize in maxpooling layers.[5]

Triplet Network

Triplet network is a type of metric learning where the similar features are grouped together and different features are placed large apart in the feature space. The network applies loss to cluster similar features together and dissimilar features farther apart in the feature space. Loss function of triplet is defined as follows :

- $L(A,P,N) = \max(\| f(A) - f(P) \|^2 - \| f(A) - f(N) \|^2 + \alpha, 0)$
- α : margin (hyperparameter)
- A : anchor
- P : positive
- N : negative

There are three different types of triplet generation methods based upon the distance between anchor, positive and negative embedding vectors;Easy Triplets, Semi-hard Triplets and Hard Triplets.

Easy Triplets: In this case, the distance between negative and anchor embedding is greater than the distance between anchor and positive embedding plus margin, i.e.d(a,p)+margin < d(a,n). Hence, the loss propagated is zero and it does not help the network to learn anything.

Semi-hard Triplets: Distance between anchor and negative is between the distance between anchor and positive,and, distance between anchor and positive plus margin,i.e.d(a,p)< d(a,n)< d(a,p)+margin. The loss propagated is positive and zero in this scenario.

Hard Triplets: The distance between the anchor and negative is less than the distance between the anchor and positive plus margin, i.e.d(a,n)< d(a,p) + margin Hence, the loss propagated backward is always positive in this case.

3.0.2 Architecture

final architecture uses a triplet network to discriminat ebetween the fake video and real video embedding vectors in both Celeb-DF and FF++ datasets. MTCNN extract faces from the frames, then, facenet generates 512 dimension embeddings for each face in the feature space.

As facenet is developed for face recognition, each unique face occupies a small cluster in the feature space. Then, we generate semi-hard triplets via online triplet mining. Using these triplets, the embeddings of fake frames and positive frames are distinctively separated through triplet loss.

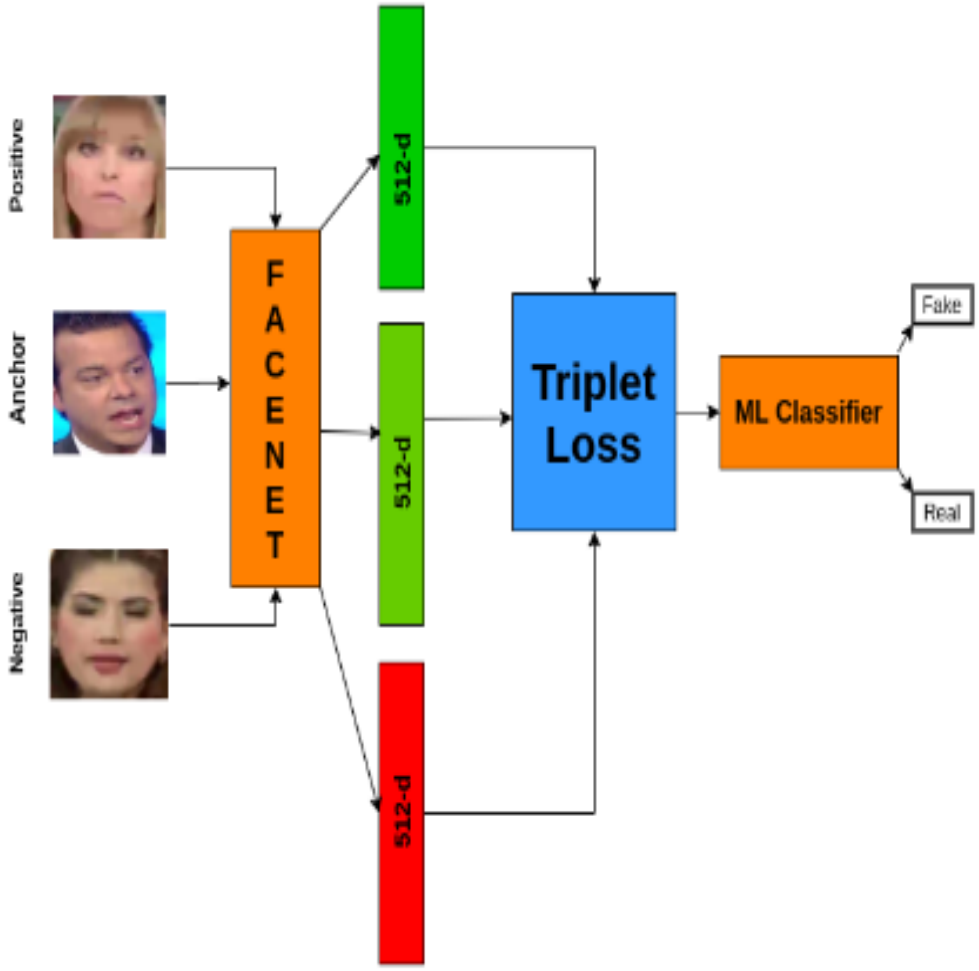

Figure 3.2: riplet architecture used for clustering and classification of fake andreal videos embeddings.

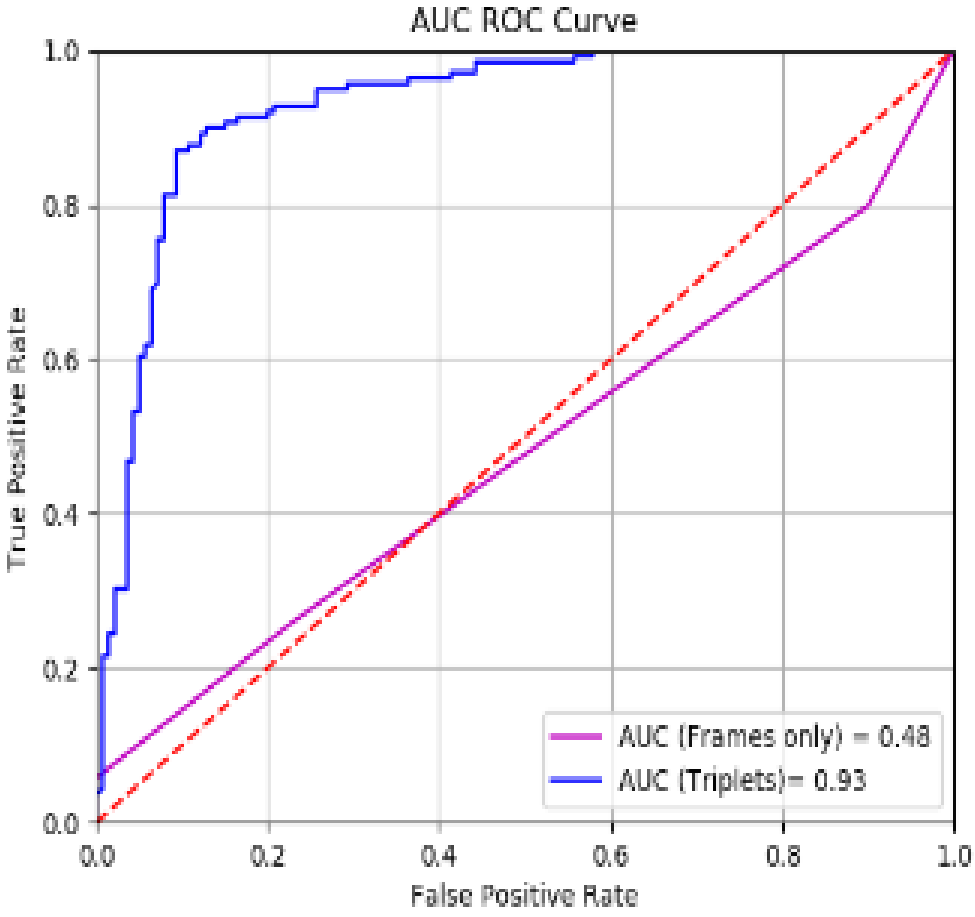

Figure 3.3: AUC ROC Curve plots of frames only and triplets network

CHAPTER 4

Effective and Fast DeepFake Detection Method Based on Haar Wavelet Transform

DeepFake using Generative Adversarial Networks (GANs) tampered videos reveals a new challenge in today's life. With the inception of GANs, generating high-quality fake videos becomes much easier and in a very realistic manner. Therefore, the development of efficient tools that can automatically detect these fake videos is of paramount importance.he proposed DeepFake detection method takes the advantage of the fact that current DeepFake generation algorithms cannot generate face images with varied resolutions, it is only able to generate new faces with a limited size and resolution, a further distortion and blur is needed to match and fit the fake face with the background and surrounding context in the source video. This transformation causes exclusive blur inconsistency between the generated face and its background in the outcome DeepFake videos, in turn, these artifacts can be effectively spotted by examining the edge pixels in the wavelet domain of the faces in each frame compared to the rest of the frame. A blur inconsistency detection scheme relied on the type of edge and the analysis of its sharpness using Haar wavelet transform is used.Haar wavelet is a sequence of rescaled "square-shaped" functions which together form a wavelet family.

4.0.1 Detection

Linear blurred images can be described as G = H * F + N (1). Where G, F N represent noisy image and Blur function is represented by H matrix. The blur extent is measured by testing the edge features in an image. The edge feature is one of these features that can be used. If the blur is present, both the edge sharpness and its type will be changed and that will indicate whether the face image has been manipulated or not.[11]

Different types of edges are present in an image. Generally, there are three classes of edge type: Dirac-Structure, Step-Structure, and Roof-Structure.The Step-Structure type is divided according to the change of intensity whether it is gradual or not into: "A Step-Structure"

and “G Step-Structure”. Every image has all types of edges more or less, most of the G Step-Structure and Roof-Structure are sharp enough. In case it is blurred, the edges lose their sharpness. The sharpness parameter is measured by the sharpness parameter alpha.This method detects whether a face image is blurred or not based on Dirac-structure and A Step-Structure. A blur extent is identified by taking sharpness of Roof-Structure and G Step-Structure into account. The sharpness of the edge is indicated by the parameter alpha , if alpha is larger, means the edge is sharper.

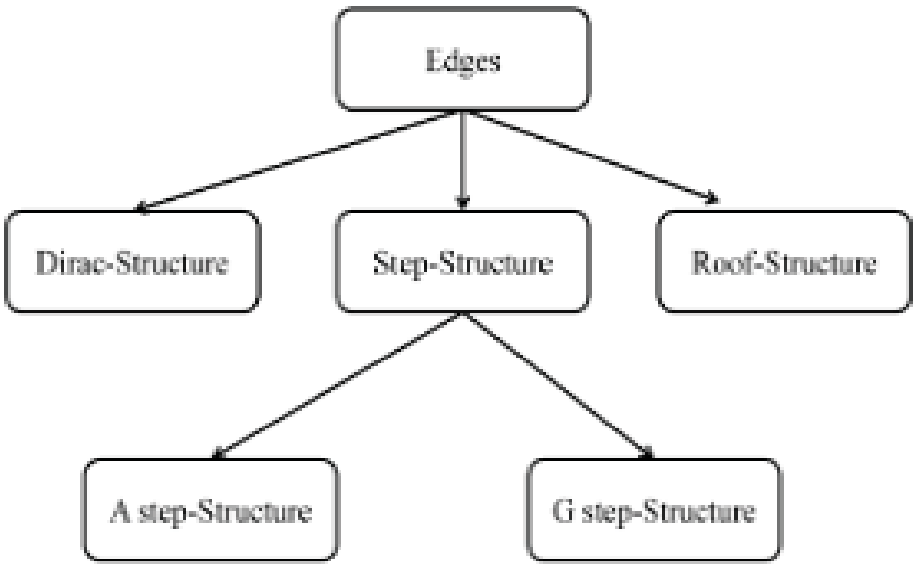

Figure 4.1: Edge Classification

4.0.2 Flow chart of deepfake detection using Haar Wavelet

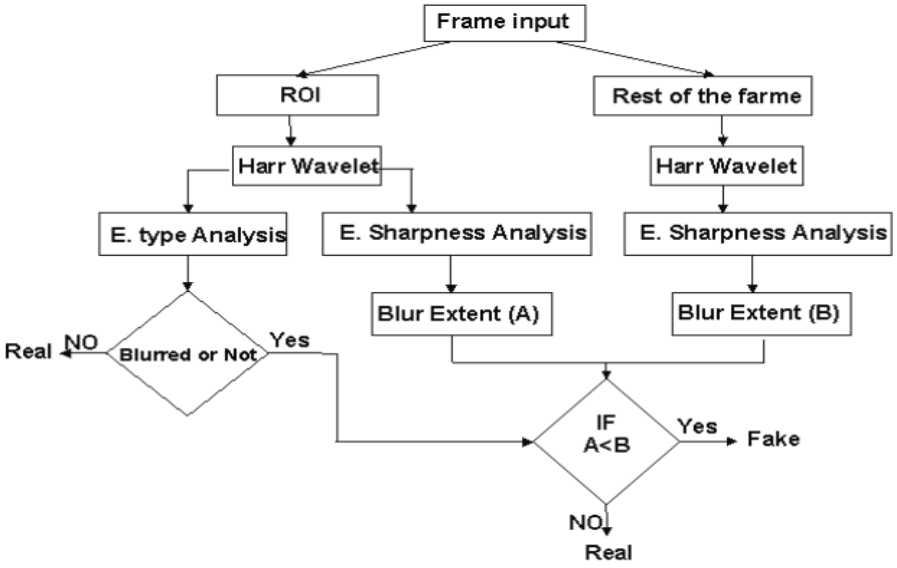

Figure 4.2: DeepFake detection Algorithm using Haar Wavelet

4.0.3 The Effect of Blur on Different Edges

The noise factor N in "(1)" can be neglected since there is a small noise ratio in photos usually acquired from digital cameras.The main blur functions H is a convolution operation that affects the equation and will change the edge property. Take into consideration that there will be no Dirac-Structure or A Step-Structure in the blurred image. On the other hand, both Roof-Structure and G Step-Structure will tend to lose their sharpness (less $\alpha value$).

4.0.4 Sharpness Detection and Edge Type

Discrete Haar functions can be described as functions determined by sampling the Haar functions at 2n points. A matrix form can represent the function in a convenient means. Every row in the matrix H(n), have a discrete Haar sequence Haar(w, t) each alone, where the index (w) represents the number of the Haar function, the discrete point of the function determination interval is identified by index (t). By the following recurrence relation , any dimension of the Haar matrix can be gained;

$$H(n) = \begin{bmatrix} H(n-1) & \otimes [1 \quad 1] \\ 2^{\frac{n-1}{2}} I(n-1) & \otimes [1 \quad 1] \end{bmatrix}, H(0) = 1$$

Where $H(n)$ - the discrete Haar functions of degree 2^n matrix, $I(n)$ - identity matrix of degree 2^n.

Figure 4.3: Recurrence relation for Haar matrix

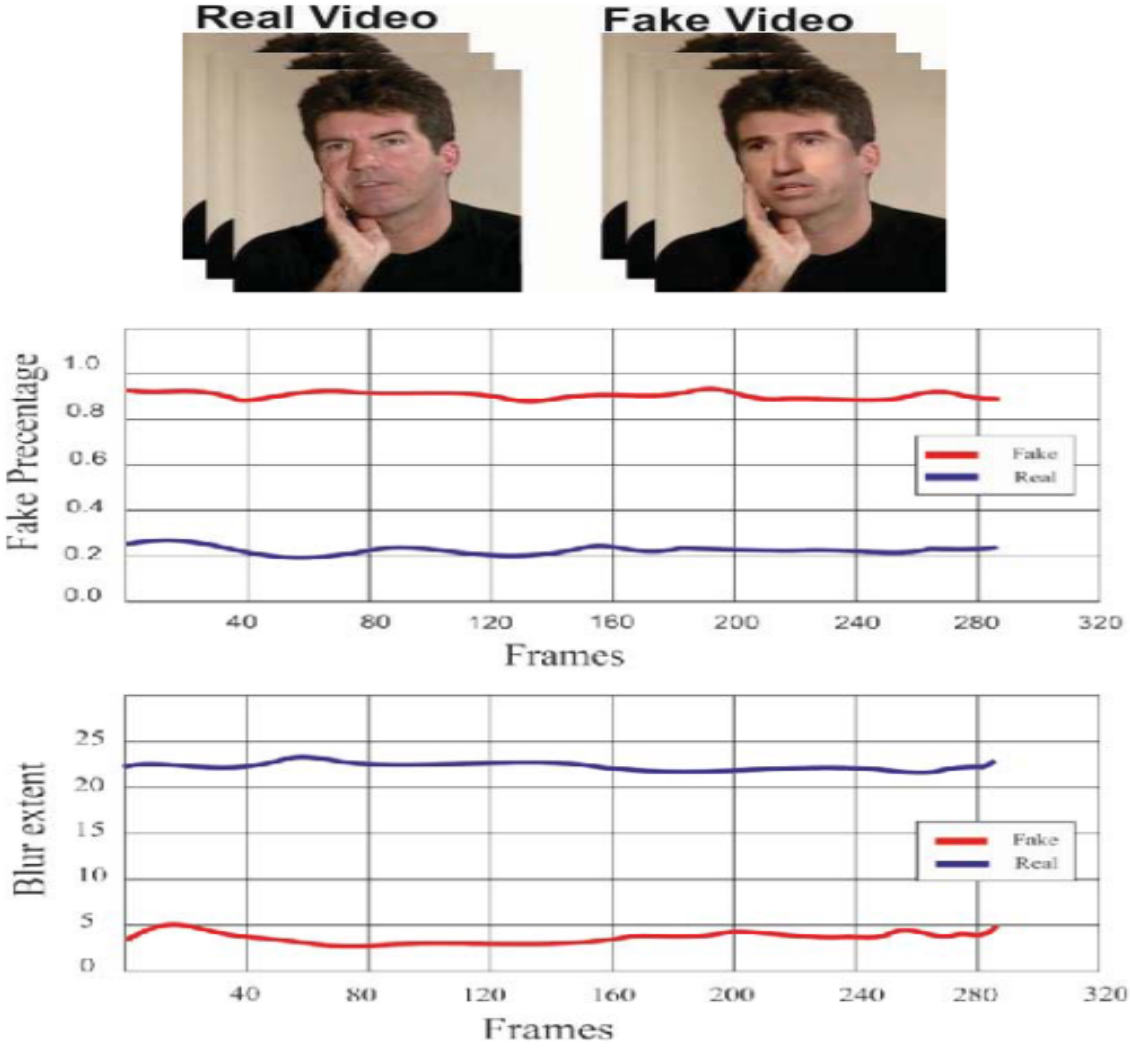

Figure 4.4: Example of the proposed method on one of the DeepFake generated videos from the "UADFV" datase

CHAPTER 5

Deepfake Video Detection Using Recurrent Neural Networks

In recent months a machine learning based free software tool has made it easy to create believable face swaps in videos that leaves few traces of manipulation, in what are known as "deepfake" videos. Scenarios where these realistic fake videos are used to create political distress, black-mail someone or fake terrorism events are easily envisioned.This paper proposes a temporal-aware pipeline to automatically detect deepfake videos. Our system uses a convolutional neural network (CNN) to extract frame-level features.These features are then used to train a recurrent neural net-work (RNN) that learns to classify if a video has been subject to manipulation or not.

This paper propose a two-stage analysis composed of a CNN to extract features at the frame level followed by a temporally-aware RNN network to capture temporal inconsistencies between frames introduced by the face-swapping process.A collection of 600 videos have been used to evaluate the proposed method, d, with half of the videos being deepfakes collected from multiple video hosting website.

5.0.1 Creating Deepfake Videos

It is well known that deep learning techniques have been successfully used to enhance the performance of image compression. Especially, the autoencoder has been applied for dimensionality reduction, compact representations of images, and generative models learning . Thus, autoen-coders are able to extract more compressed representations of images with a minimized loss function and are expected to achieve better compression performance than existing image compression standards. The compressed representations or latent vectors that current convolutional autoencoders learn are the first cornerstone behind the faceswap-ping capabilities of . The second insight is the use of two sets of encoder-decoders with shared weights for the encoder networks.[4]

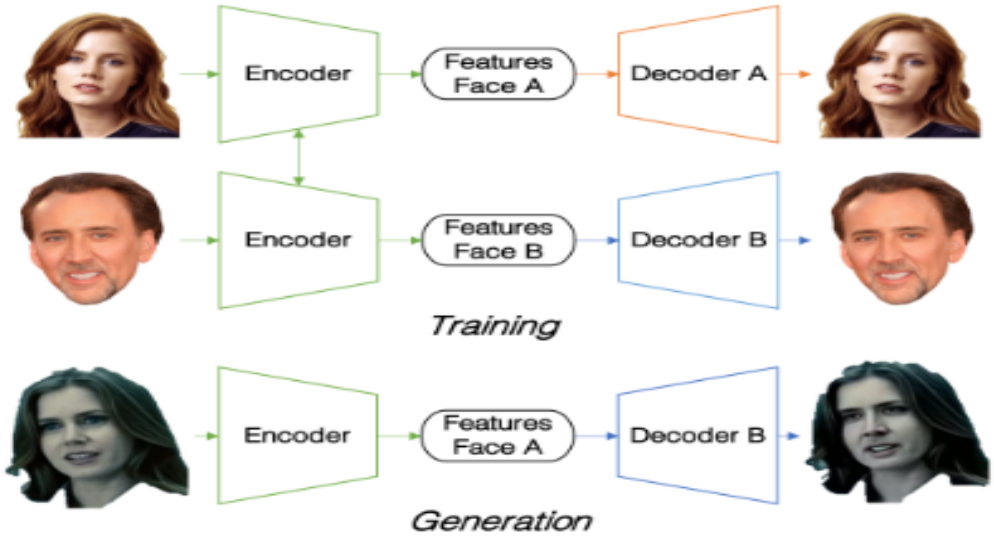

Figure 5.1: Generation of deepfakes using auto encoders

5.0.2 Training

Two sets of training images are required. The first set only has samples of the original face that will be replaced, which can be extracted from the target video that will be manipulated. This first set of images can be further extended with images from other sources for more realistic results. The second set of images contains the desired face that will be swapped in the target video. To ease the training processof the autoencoders, the easiest face swap would have both the original face and target face under similar viewing and illumination conditions. However, this is usually not the case. Multiple camera views, differences in lightning conditions or simply the use of different video codecs makes it difficult for autencoders to produce realistic faces under all conditions. This usually leads to swapped faces that are visually inconsistent with the rest of the scene. This frame-level scene inconsistency will be the first feature that we will exploit with our approach.

If two autoencoders are trained separately on different sets of faces, their latent spaces and representations will be different. This means that each decoder is only able to decode a single kind of latent representations which it has learnt during the training phase. This can be overcome by forcing the two set of autoencoders to share the weights for the encoder networks, yet using two different decoders.

5.0.3 Video Generation

When the training process is complete, we can pass a latent representation of a face generated from the original subject present in the video to the decoder network trained on faces of the subject we want to insert in the video.The decoder will try to reconstruct a face from the new subject, from the information relative to the original subject face present in the video. This process is repeated for every frame in the video where we want to do a face swapping operation. This is usually a second source of scene inconsistency between the swapped face and the re-set of the scene. Because the encoder is not aware of the skin or other scene information it is very common to have boundary effects due to a seamed fusion between the new face and the rest of the frame.

5.0.4 Recurrent Network for Deepfake Detection

The proposed system is composed by a convolutional LSTM structure for processing frame sequences. There are two essential components in a convolutional LSTM:

- CNN for frame feature extraction.
- LSTM for temporal sequence analysis

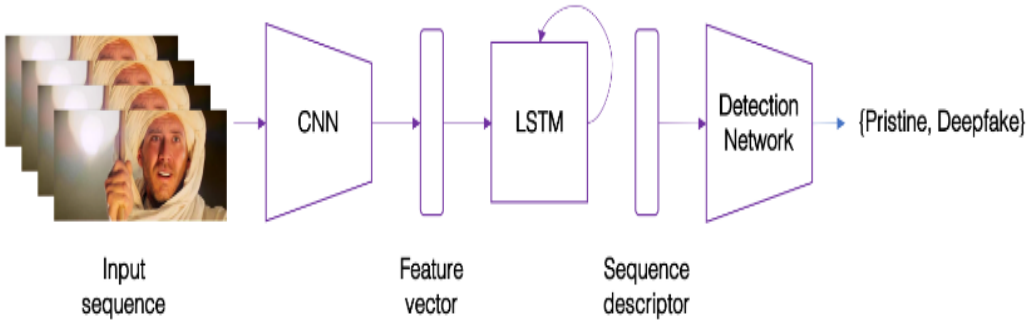

Figure 5.2: Overview of our detection system.

5.0.5 Convolutional LSTM

A convolutional LSTM is employed to produce a temporal sequence descriptor for image manipulation of the shot frame. Aiming at end-to-end learning, an integration of fully-connected layers is used to map the high-dimensional LSTM descriptor to a final detection probability. Specifically, our shallow network consists of two fully-connected layers and one dropout layer to minimize training over-fitting. The convolutional LSTM can be divided into a CNN and a LSTM, which we will describe separately in the following paragraphs.

CNN for Feature Extraction.This method uses InceptionV3 with fully connected layers.The 2048 dimensional feature vectors after the last pooling are used as the sequential LSTM input.

LSTM for Sequence Processing.Sequence of CNN feature vectors of input frames are input and a 2-node neural network with the probabilities of the sequence being part of deep fake video or real video.LSTM model takes a sequence of 2048-dimensional ImageNet feature vectors. LSTM is followed by 512 fully-connected layers with 0.5 chance of dropout. Finally softmax layer is used to compute the probabilities of the frame sequence being either pristine or deep fake.

5.0.6 Architecture

Dataset is selected with 50% of deep fake videos and 50% of real videos. Random split is used to generate 3 disjoint set which is used for training, validation and testing. Balanced splitting on real videos and fake videos are done. This made sure that final set has exactly 50% of each class. Resized every frame to 299 X 299. Length of input sequence is controlled using subsequence sampling. This allows us to identify how many frames are necessary per video to have an accurate detection. Optimizer is set for end to end training of the complete model.

CHAPTER 6

Conclusion

- AI generated videos that look real but are actually fake are called Deep Fake videos
- Generative Adversarial Networks are used to create deep fake videos
- Haar Wavelet Transform detects whether a video is fake or not by comparing the blurred area of ROI the surrounding context
- Metric Learning uses triplet loss to distinctly identify embeddings of fake frames
- Recurrent Neural Network uses CNN to extract frame-level features and RNN to identify whether a video has been manipulated or not

REFERENCES

[1] S Agatonovic-Kustrin and R Beresford. Basic concepts of artificial neural network (ann) modeling and its application in pharmaceutical research. *Journal of pharmaceutical and biomedical analysis*, 22(5):717–727, 2000.

[2] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo. Deepfake video detection through optical flow based cnn. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1205–1207, 2019.

[3] Reagan L Galvez, Argel A Bandala, Elmer P Dadios, Ryan Rhay P Vicerra, and Jose Martin Z Maningo. Object detection using convolutional neural networks. In *TENCON 2018-2018 IEEE Region 10 Conference*, pages 2023–2027. IEEE, 2018.

[4] D. Güera and E. J. Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2018.

[5] Kamal Jnawali, Mohammad R Arbabshirani, Navalgund Rao, and Alpen A Patel. Deep 3d convolution neural network for ct brain hemorrhage classification. In *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, page 105751C. International Society for Optics and Photonics, 2018.

[6] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018.

[7] A. Kumar, A. Bhavsar, and R. Verma. Detecting deepfakes with metric learning. In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2020.

[8] Francesco Visin, Kyle Kastner, Kyunghyun Cho, Matteo Matteucci, Aaron Courville, and Yoshua Bengio. Renet: A recurrent neural network based alternative to convolutional networks. *arXiv preprint arXiv:1505.00393*, 2015.

[9] J. Xiang and G. Zhu. Joint face detection and facial expression recognition with mtcnn. In *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, pages 424–427, 2017.

[10] Jia Xiang and Gengming Zhu. Joint face detection and facial expression recognition with mtcnn. In *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, pages 424–427. IEEE, 2017.

[11] Mohammed Akram Younus and Taha Mohammed Hasan. Effective and fast deepfake detection method based on haar wavelet transform. In *2020 International Conference on Computer Science and Software Engineering (CSASE)*, pages 186–190. IEEE, 2020.

[12] K. Zhu, B. Wu, and B. Wang. Deepfake detection with clustering-based embedding regularization. In *2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC)*, pages 257–264, 2020.