Deep Fake Detection Using Hybrid Deep Learning Architecture

A Project Report Submitted in Partial Fulfillment of the Requirements for Award of
the Degree of Bachelor of Technology in Information and Communication Technology

Submitted by
Ishan Mistry 18BIT033
Alister Rodrigues 18BIT041
Khush Joshi 18BIT056
Manali Shah 18BIT060
Mansi Raveshia 18BIT063

Under the Supervision and Guidance of
Dr. Mohendra Roy
Assistant Professor
Pandit Deendayal Energy University

Submitted to
Department of Information and Communication Technology
School of Technology
Pandit Deendayal Energy University (PDEU)
Gandhinagar, INDIA, 382007

Contents

Chapter 1

Introduction

1.1 Problem statement

Detection of Deep Fakes using a hybrid neural network architecture : The Deep Fakes out there need to be kept in check and for that very reason we need to elevate our detection techniques too, in order to tackle them.This project requires studying the various implemented detection schemes and use the insights to devise a new and possibly better hybrid architecture for the detection of Deep Fakes.

1.2 Objective

- Analyze the various prevailing implementations: Study of the various currently popular architectures and learn what exactly is it that makes these architectures good at their job.

- Studying the characteristics of the Deep fakes images and regular images This is essentially crucial and could give us a boost in critically selecting models/layers

- Devising a novel technique if possible for the detection Combining both the stated objectives enable us to think critically and approach solving the problem from a novel perspective.This implementation could then be hosted online on a small web app, made on StreamLit, so that it is available to others for use. This way the project becomes our little contribution back to the Society as well to the Deep Learning community.

1.3 Motivation

People have a tendency to readily believe what they see, quite often overlooking the credibility of the source of the video/image/text.Although people are getting more and more vigilant these days, becoming skeptical towards blindly trusting the media, but there is a caveat to this self established trust.This is a subtle case of misinformation.
Humans are believed to have six degrees of separation between you and your farthest friend. "Six degrees of separation" is the idea that all people on average are six, or fewer, social connections away from each other. As a result, a chain of "friend of a friend" statements can be made to connect any two people in a maximum of six steps.Moreover the situation has been exacerbated by the various channels of media/communication that are consumed these days - News groups/conspiracy groups/Social Media. There is no one to substantiate the validity of these seemingly true sources, as it is a result of the mere misinformation that has been made really easy by the current forms of media.
At first glance, these matters might seem trivial and even not alarming at all, because in most cases the impact of these is not directly visible.But there are times when things could go down-hill and have serious repercussions.For example, at times such information could be used to belittle someone or could be a philippic , aimed to damage the reputation of someone in power.This when combined with the current power of the spreading information can be treacherous.
These subtle but misinformed texts or images are not easy to track and the process could be cumbersome as this spurious information is gleaned into everyday lives to a point that they come out as obvious and evidently become the truth. The facts and figures presented/spread could still be validated with the human expertise in the specific domains - Politics/Sports/Daily Sciences etc. But with advances in Deep Learning, deep fakes have become so realistic that it is nearly impossible for us to visibly differentiate a true video from a deep fake video.This gives rise to the need of having accurate detection techniques for their detection and hence their removal from the web.The bad people keep getting better and better and so shall we, in order to counter them.

1.4 Roadmap

1. Observing the trends/architectures : Literature Survey
2. Start by implementing a few of the current models - Experimenting with them/Looking for vulnerabilities
3. Consolidate our work along with the real world testing

4. Make our findings available to people, in the form of a simple web app and also to the Deep Learning Community in form of a paper

1.4.1 Tentative Timeline

- Week 1 : Settling on the Logistics for Literature Survey and Too
- Week 2 : Literature Survey + Insights from the papers
- Week 3 : Mathematical Formulation of the problem statement
- Week 4 : Coming up with hypothesis/experiments + Setting up environments
- Week 5 : Github Project setup + Modelling the Neural Network + Implementation
- Week 6 : Testing the network + Validation + Conducting Experiments
- Week 7 : Concluding the experiments + Working on the paper + Conference submission
- Week 8 : Submission wrap up + Concluding the project

1.5 Work Done so far

1.5.1 Logistics

- Setting up the System for managing the details from the literature survey
- Zotero Management - for ease of future citations/ reference management

1.5.2 Literature Survey

For literature survey we have started looking for papers and are currently briefly going through the abstracts and will redistribute the papers for in- depth readings as we progress further.

- Deep Learning for Deepfakes Creation and Detection: A Survey
- Deep-fakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward
- Recurrent Convolutional Strategies for face manipulation detection in videos

- An exploratory analysis on visual counterfeits using Conv-LSTM hybrid architecture
- Deepfakes detection technique using deep learning : A Survey
- Deep Fake Detection : Survey of Facial Manipulation Detection Solutions
- Media Forensics and Deep Fakes: An Overview

Chapter 2

Literature Review
24/08/21 - 27/08/21

2.1 Introduction

The survey papers that have been analyzed this week are as follows:

2.1.1 Deep-Fakes Detection Techniques Using Deep Learning: A Survey

This paper primarily concerns with the Deep Fake detection using RNNs, CNNs and LSTMs.This was a very beginner friendly paper and had few architectures mentioned.But this paper proved to be a good starting point for the survey work.

2.1.2 Countering Malicious DeepFakes: Survey, Battleground, and Horizon

This paper was very well documented and was a mega survey as it contained the work involved in around 191 papers.This also has a very unique section - "DeepFake BattleGrounds" - lays the field for early researchers to know the ins and outs of the feud that goes on between detection and generation.

2.2 DeepFake Generation Techniques

In the recent years there have been various algorithms for Deepfakes generation but there was no open source software or tool was available. Methods based on the neural image style transfer becomes the tool for creating the deepfakes videos. Addition to this there are now several open source softwares like FakeApp, DFaker,

Faceswap-GAN, faceswap and DeepFaceLab can generate the deepfakes.The deep-Fake Generation has been classified into 4 major categories :

1. Entire Face Syntheses
2. Attribute Manipulation
3. Identity Swap
4. Expression Swap
5. Other Methods

2.2.1 Entire Face Syntheses

These types of generation techniques take Input as Random Vectors and produce as Output - High Quality non-existent faces.Target images are not needed (latent vector is manipulated for results). The following architectures have been used typically for the Entire face synthesis

DCGAN

First work that does CNN+GAN. Focuses on unsupervised learning - had problems like balancing discrimator and generator

WGAN

This took care of the balance and provided reasonable and efficient approximation of the EM distance. WGAN uses weight clipping to enforce a Lipschitz constraint. To improve the weight clipping operation, they have proposed to penalize the norm of the gradient of the discriminator with respect to its input fake image. The new designs train stably when generating high-quality home images

BEGAN

This took care of the balance and provided reasonable and efficient approximation of the EM distance. WGAN uses weight clipping to enforce a Lipschitz constraint. To improve the weight clipping operation, they have proposed to penalize the norm of the gradient of the discriminator with respect to its input fake image. The new designs train stably when generating high-quality home images

CrammerGAN

Simply using Wasserstein probability can not simultaneously satisfy sum invariance, scale sensitivity, and unbiased sample gradients.Hence came CrammerGANS. It combined the best of the Wasserstein and Kullback-Leibler divergences to propose the Cramér distance.

PGGGAN

Focus on high res images. The images are starting from a low resolution and being detailed step by step with the new layers added in the model. This method is very reasonable in that it can speed up the training as well as greatly stabilize the GAN

BigGAN

The main focus is to generate high resolution diverse images. They have applied orthogonal regularization to enforce the generator to be satisfied with a simple "truncation trick". Thus, the user can control the trade-off between image fidelity and variety by reducing the variance of the generator's input.

StyleGAN

Automatically learn the unsupervised separation of high-level attributes such as pose and human identity. The architecture also leads to stochastic variation in the generated images (e.g., freckles, hair). Furthermore, it enables intuitive, scale-specific control of the synthesis. They have encouraged good conditioning in the mapping from latent codes to images by the new design of generator normalization, progressive growing, and generator regularization.

Glow

A flowbased generative model that uses an invertible 1×1 convolution. The method is based on the theory that a generative model optimized towards the plain log-likelihood objective has the ability to generate efficient realistic-looking synthesis and manipulate large images.

2.2.2 Attribute Manipulation

Similar to the prvious techniques, these also don't necessarily require Target images(latent vector manipulated for results).It is known as face editing, which cannot only modify simple face attributes such as hair color, bald, smile, but also retouch complex attributes like gender, age, etc. Examples of such architectures used are as follows :

IcGAN

This can be seen as an extension of cGANs(Conditional GANs) . They have evaluated encoders to map a real image into a latent space and a conditional representation, which allows the reconstruction and modification of arbitrary attributes of real human face images

StarGAN

As previous studies only did image to image translation only for 2 domains - which is cumbersome and time consuming, StarGAN was devised to perfom translation to multiple domains. It allows simultaneous training of multiple different-domain datasets within a single network.

StarGAN2

Maintains diversity of generated images as well as scalability over multiple domains as obtained in StarGAN. They replaced StarGAN's domain label with their domain-specific style code. To adapt the style code, they have proposed two modules: a mapping network(transforms random noise into style codes) and a style encoder(extract style code form a given ref image).

GANimation

StarGAN had the limitation of the content of the datasets, it can only generate a discrete number of expressions. To address this we have a novel GAN condditoning method based on action units (AU) annotations.It defines the human expression with a continuous manifold of the anatomical facial movements. The magnitude of activation of each AU can be controlled independently. Different AUs can also be combined with each other with this method.

AttGAN

Previous methods have attempted to establish an attributeindependent latent representation for further attribute editing. However, since the facial attributes are relevant, requesting for the invariance of the latent representation to the attributes is excessive. Therefore, simply forcing the attribute-independent constraint on the latent representation not only restricts its representation ability but also may result in information loss, which is harmful to the attribute editing. To solve this problem, AttGAN (He et al., 2019b) has removed the strict attributeindependent constraint from the latent representation. It just applies the attribute classification

constraint to the generated image to guarantee the correctness of attribute manipulation. Meanwhile, it groups attribute classification constraint, reconstruction learning, and adversarial learning together for high-quality facial attribute editing.

STGAN

Improvement of AttGAN has selectively taken the difference between target and source attribute vectors as the input of the model. They have enhanced attribute editing by adding a selective transfer unit that can adaptively select and modifying the encoder feature to the encoder-decoder

2.2.3 Identity Swap

This function is able to replace the face in the target image with the face in the source image

2.2.4 Expression Swap

Expression swap is similar to identity swap. It is able to replace the facial expression in the target image with the facial expression in the source image. It is also known as face reenactment. In our investigation, only Face2Face and A2V were attempted by the surveyed DeepFake detection methods.

2.2.5 Other Methods

Style Transfer

1.GatedGAN GatedGAN uses gated networks to transfer multiple styles in a single model. They have added a gated transformer into the encoder-decoder

2. AAMS AAMS has developed an attention-aware multistroke style transfer model. They have enabled using different brush strokes to render the diverse levels of detail. They also have coordinated spatial distribution of visual attention between the content image and stylized image.

In-painting

1. ContextAtten The new architecture proposed by them can synthesize novel image structures as well as explicitly utilize surrounding image features as references to make better predictions.

2. SC-FEGAN A novel image editing system that generates images with the free-form masks, sketches, and color provided by the users.

Rendering

1. CRN CRN (a single feed-forward network, trained end-to-end with a direct regression objective) has proposed a rendering network to produce a photographic image with a two-dimensional semantic specification of the scene.

2. GauGAN GauGAN has proposed a simple yet effective layer for synthesizing photo-realistic images with an input semantic layout. They have proposed to use a spatially-adaptive, learned transformation to modulate the activation in normalization layers with the input layout.

Super Resolution

1. SAN SAN has proposed a second-order attention network for more powerful feature expression and feature correlation learning.

Detection evasive

1. SDGAN SDGAN has proposed using a spectral discriminator to simulate the frequency distribution of the real data when generating images.

2. WUCGAN WUCGAN has shown common up-sampling methods are causing the inability of GANs to reproduce spectral distributions of real images correctly. To overcome this drawback, they have proposed to add a spectral regularization term to the training optimization objective

De -Identification

They mainly obfuscate identities in photos by the head replacement for data privacy.A good area for research!

2.3 DeepFake Detection Techniques

1. The first category of DeepFake detection methods are data-driven, which directly employ various types of DNNs trained on real and DeepFake videos, not relying on any specific artifact.

2. The second category of DeepFake detection algorithms use signal level artifacts introduced during the synthesis process such as those described in the Introduction.

3. Third category is based on inconsistencies exhibited by the physical/physiological aspects in the DeepFake videos

Third category is based on inconsistencies exhibited by the physical/physiological aspects in the DeepFake videos

2.3.1 Spatial Based

Image Forensics based detection

Differences between synthesised faces and real faces are revealed in the chrominance components, especially in the residual domain. Idea was proposing to train a one-class classifier on real faces by leveraging the differences in the chrominance components for tackling the unseen GANs. However, performance against perturbation attacks like image transformations is unknown. Similarly in tackling fake videos, researchers borrowed ideas from traditional video forensic by leveraging the local motion features captured from real videos to spot the abnormality of manipulated videos

DNN-based detection

Completely data driven by utilising DNN models by extracting spatial feaatures to improve effectiveness and generalization detection. Severely weak against adversarial attacks with additive noises. Existing studies to leverage DNN to identify deep fakes are categorised by 1) Improving generalisation abilities, 2) Investigating artifact clues and 3) Empowering CNN models

Obvious artifact clues

Generated deepfakes exhibit obvious artifacts due to limitations in AI and can be leveraged. A full convolutional approach is applied for training classifiers. Two vectors from the aforementioned two networks are compared for detecting the identity-to-identity discrepancies. This approach also has a good generalization ability across GANs

Detection and localisation

By locating manipulated regions that provide evidence for forensics, it was found that the imperfection of upsampling methods exhibits obvious clues for detection

and forgery localization where the manipulated area could be precisely marked

Facial image preprocessing

Some studies propose preprocessing the facial images before sending them to binary classifiers for discrimination..Layer-by-layer neuron behaviors provide more subtle features for capturing the differences between real and fake faces. This provides a new insight for spotting fake faces by monitoring third-party DNN-based neuron behaviors, which could be extended to other fields like fake speech detection.

2.3.2 Frequency based

GAN-based artifacts

Investigating imperfect designs of existing GANs which provides obvious signals, it was observed that the internal value of the generator is normalized which limits the frequency of saturated pixels. Then, a simple SVM-based classifier is trained to measure the frequency of saturated and under-exposed pixels in each facial image for discriminating fake faces.

Frequency Domain

Severe artifacts introduced due to the upsampling techniques in GANs, to which a classifier with a simple linear model and a CNN based model can acheive promising results on the entire frequency spectrum.

2.3.3 Biological Signal based

Visual-Audio inconsistency

Specific words give that involve in lips touching is found inconsistent in fake videos. Lip sync inconsistencies are strong but not solid evidence towards deep fakes

Visual Inconsistency

Indicates that synthesized faces are inconsistent and unnatural. In noticing general behavioural patterns in humans and deepfakes, a lot of artifacts can be found

Biological Signals in video

Biological signals like heartbeat rhythms and monitoring blood flow to observe subtle color changes in the skin

2.3.4 Other Detectors

Distributed ledger technologies (DLT)

Leveraging distributed ledger technologies (DLT) to combat digital deception, user behavior clues like the eye-gaze for DeepFake detection. Finding the artifacts which exist in the specific facial region could improve the detection performance by a large margin than the entire face.

2.4 DeepFake Evasion Techniques

With the fast improvement of DeepFake finders, specialists begin focusing on plan techniques to dodge the phony faces being distinguished.

In particular, given a genuine or phony face, avoidance strategies map it to another one that can't be accurately arranged by the cutting edge DeepFake identifiers, stowing away the phony appearances from being found.

We can generally separate all techniques into three kinds:-

1. The first type is based on the adversarial attack.
2. The second type of methods focus on removing the fake traces in the frequency domain. These methods mainly focus on the mismatching between real and fake faces in the frequency domain while neglecting other potential factors that may make fake faces be identified easily.
3. The third kind of methods regard evasion as a general image generation process and use advanced image filtering or generative models to mislead Deep-Fake detectors.

2.5 Datasets

- https://www.idiap.ch/en/dataset/deepfaketimitDeepfakeTIMIT dataset
- https://github.com/NVlabs/styleganFlickr-Faces-HQ, FFHQ
- https://generated.photos/100K-Faces
- https://github.com/NVlabs/ffhq-dataset
- https://paperswithcode.com/dataset/casia-webfaceCASIA-WebFace

- https://www.tensorflow.org/datasets/catalog/vgg$_face2VGGFace2$