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

Week 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

Week 2 - Literature Review

2.1 Introduction

The work for this week(24/08/21 - 27/08/21) includes analysis of survey papers.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 deepFake 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

Chapter 3

Week 3 - Research study

3.1 Introduction

This week (30/08/21 - 2/09/21) we have analysed the following papers:

Week 3 - Research Papers		
Research Title	**Publication Date**	**Journal Links**
Deep-Fake Forensics via An Adversarial Game	25 March 2021	Arxiv
Emotions Don't Lie:	1 Aug 2020	Arxiv
Celeb-DF:	16 June 2020	Arxiv
Deep-Fakes Detection Techniques Using Deep Learning	16-19 May 2021	SCIRP
DeepFakes and Beyond:	18 June 2020	Arxiv
Subjective and Objective Evaluation of Deep-Fake Videos	6-11 June 2020	IDIAP
Video Face Manipulation Detection Through Ensemble of CNNs	16 Apr 2020	Arxiv
The DeepFake Detection Challenge (DFDC) Dataset	28 Oct 2020	Arxiv
FaceForensics++: Learning to Detect ManipulatedFacial Images	26 Aug 2019	Arxiv
One Detector to Rule Them All Towards a GeneralDeepfake Attack Detection Framework	1 May 2021	Arxiv
Deepfake Videos in the Wild: Analysis and Detection	11 Sept 2021	Arxiv

3.1.1 Deep-Fake Forensics via An Adversarial Game[19]

Training with samples that are adversarially crafted to attack the classification models **improves the generalization ability** considerably.

They propose a new adversarial training method that attempts to blur out the specific artifacts, by **introducing pixel-wise Gaussian blurring models**.

Most models till now used handcrafted features - **noise variance analysis** and **digital shadow writing analysis** to detect deep fakes.CNNs' came along pretty well but problem of generalization and sensitivity to image quality - still persists.

Previous attempts of improving Generalization include : A more generalizable deep face representation for achieving the goal, by introducing **an auxiliary task of predicting face x-ray**.Some modified the network architecture and introduced an **attention module for the task**.Auto-Encoders are also used for the same.

Traditional Adversarial learning revised

Assume that a classification model (e.g., the deepfake detection model) attempts to minimize the prediction loss $L(x, y; \theta)$ for any given data (x, y) (i.e., an image x paired with its label y), in which θ collects all learnable parameters in the classification model, $x \in \mathbb{R}$h×w×c (h, w, and c represent height, width, and number of channels of x, respectively).

Normally we minimize the empirical risk$\sum_{(x,y)\in\mathbb{D}}[L(x, y; \theta)]$

FSGM suggests to obtain adversarial sample by adding a scaled input-Gradient sign to the original Image i.e :

$$x^{adv} = x + \epsilon \cdot sign(\nabla_x L(x, y; \theta)),$$

The above equation is derived from an optimization problem maximizing the prediction loss of an input obtained by adding a perturbation (whose l norm is no greater than) to a clean image x

Adversarial training plays a zero-sum game which includes an auxiliary process that generates adversarial examples which maximizes the classification loss. The generated adversarial examples can be used instead of the original benign examples or in combination with them for training. For the original images only scenario, we had

$$\min_{\theta} \sum_{(x,y)\in\mathbb{D}} \max_{\delta\in\mathbb{S}} L(x + \delta, y; \theta),$$

Now for the combination examples :

$$\min_{\theta} \sum_{(x,y)\in\mathbb{D}} L(x, y; \theta) + \lambda \cdot \max_{\delta\in\mathbb{S}} L(x + \delta, y; \theta).$$

For $\mathbb{S} \subseteq \mathbb{R}$h×w×c to constrain the noise/disturbance with respect to the clean image.

Spatial Adversarial Learning

Another method includes using calculate a **n adversarial optical flow** to **spatially transform each pixel** of the clean images accordingly.

Let us use $x_{i,j}$ to represent the pixel on the i-th row and j-th column of the original image, an adversarial optical flow vector $f_{i,j} := (\triangle u_{i,j}, \triangle v_{i,j})$ is learned in the method to transform $x_{i,j}$ to the corresponding position $(i', j') = (i + \triangle u_{i,j}, j + \triangle v_{i,j})$on the adversarial image x^{adv}.Such Training is called Spatial Transformed Adversarial Learning(SAT).

Blurring Adversarial Training

Specifically, given an input image x whose height, width, and number of channels are h, w and c, respectively, we obtain x^{adv} by performing pixel-wise Gaussian blur on x. We use $x^{adv}_{i.j}$ to represent the (i, j)-th pixel of x^{adv}, and we attempt to learn a single-channel map σ^{adv} with a size of hw, each of whose entries (e.g., $\sigma^{adv}_{i,j}$) represents the standard deviation of a Gaussian kernel to be applied to the region centered at the corresponding pixel of image x, i.e $x_{i,j}$.

$$G_{i,j}(u, v) = \frac{1}{2(\sigma^{adv}_{i,j})^2} exp\Big(\frac{u^2 + v^2}{2(\sigma^{adv}_{i,j})^2}\Big),$$

$$x^{adv}_{i.j} = \langle G_{i,j}, \gamma(x_{i,j}, k)\rangle,$$

Here, u and v represent the relative coordinates to the centre pixel in $(x_{i,j}, k)$ - vectorization leads to efficient computation.

σ^{adv} controls the blurriness Larger entries of σ^{adv} should lead to more blurry images and leave less obvious artifacts from the deepfake generator, while on the contrary, smaller entries of σ^{adv} leave more obvious artifacts for the classification model to learn.

We have $x^{adv} \to x$, as all the entries of $\sigma^{adv} \to 0$

The update rule is similar to FGSM :

$\sigma^{adv} = \sigma + \epsilon \cdot \nabla_\sigma L(x^{adv}, y; \theta)$,

Again vectorization would make the process of making updates easier.

Generative Adversarial Training

The parameters of the Discriminator(θ_D) and the Generator(θ_G) are updated in the paper using the weighted noisy rule. The optimization problem formualted is as follows :

$$\min_{D} \max_{G} \sum_{(x,y)\in\mathbb{D}} L(x,y;\theta_D) + \sum_{(x,y)\in\mathbb{D}} L(G(x;\theta_G),y;\theta_D),$$

The introduced generator G can also be considered as an enhancement model for the original deepfake generator(s). The optimization problem above, allows to train a generator G whose goal is opposite to the deepfake detection model, to remove obvious artifacts and synthesize more realistic deepfakes that can invalidate the deepfake detection model.

3.1.2 Emotions Don't Lie: An Audio-Visual Deepfake Detection Method using Affective Cues

A paper that aims to maximize information for learning by extracting and analyzing the similarity between the two audio and visual modalities from within the same video while exploiting video and audio modalities simultaneously. The methods mentioned in the paper included audio/video modality embedding, which utilising 2D convolutional neural network with ReLU in between all layers, and audio/video perceived emotions embedding, which is based on Memory Fusion Networks. OpenFace and pyAudioAnalysis are the tools used to extract faces in videos and audio respectively.

3.1.3 Celeb-DF: A Large - Scale Challenging Dataset for DeepFake Forensics

The paper includes a Celeb-DF dataset which includes 5,639 high quality deepfakes videos of celebrities, generated using improved synthesis process. Consisting of 500+ real videos and 5000+ deepfake videos, the database videos are well diversified among genders, age and ethnic groups. The purpose of this dataset is to compare with other deepfake datasets and to survey current deepfake detection methods on a large scale of videos

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

A survey that primarily concerns itself with the Deep Fake detection using RNNs, CNNs and LSTMs. It explains in detail on the utilisation of deep neural networks on feature extraction in different situations. In video detection, using biological signals to detect various artifacts is made using modality embedded networks and loss functions while spatial and temporal features are extracted by analysing the temporal sequence between frames.

3.1.5 DeepFakes and Beyond: A survey of Face manipulation and Fake Detection

This survey emphasizes on the various types of deepfakes techniques and how to detect these techniques. The discussion is mainly on the manipulation techniques, such as face morphing, de-identification, face generation and swapping, etc.

3.1.6 Subjective and Objective Evaluation of Deep-Fake Videos

This paper presents a subjective study to evaluate the difficulty to detect a deepfake between humans and neural networks.60 images were tested, including both the real and the fakes. The original FDFC dataset contains only fake/real classification but the experiment has defined 5 new categories(Very Easy/Easy/Moderate/Difficult/Very Difficult).
c
SOTA Models evaluated are **Xception** and **EfficientNet** (B4 variant) neural network models, pre-trained on two other public databases: **Google and Jiqsaw subset from FaceForensics++** and **CelebDF v2 dataset**

Data and Subjective Evaluation

The subjective evaluation, to gauge human accuracy in detecting Deep-Fakes, was conducted using **QualityCrowd 2 framework** designed for crowdsourcing-based. It claims fairness as the algorithm uses face sections tso identify and hence the humans are also shown the same facial highlights along with the video. Furthermore, **to evaluate the 'trustworthiness'** of the subjects, we used the **'honeypot' method** to filter out scores from people who did not pay attention, ending up with 18.66 answers per video on average.

To check **whether the difference between videos from the five deepfake categories is statistically significant** based on the subjective scores, we performed **ANOVA test**

Evaluation of Algorithms-

If evaluated on the test sets of the same databases they were trained on, both Xception and EfficientNet models demonstrate high accutacy with the area under the curve (AUC) metric equal to almost 100

To compute the performance accuracy, we need to select the threshold. We chose the threshold corresponding to the attack presentation classification error rate (APCER) of 10%, selected on the development set of the respective database.

APCER measures the proportion of attacks, deepfakes in this case, that are incorrectly classified as bonafide or original videos. The counterpart metric is bona fide presentation classification error rate (BPCER), which measures the proportion of incorrectly classified original videos.

$$APCER = \frac{FP}{TN + FP}$$

3.1.7 Video Face Manipulation Detection Through Ensemble of CNNs

This research paper's objective was to be able to run a forensic detector in real-world scenarios and to be able to analyze 4 000 videos in less than 9 hours with minimal resources. Using EfficientNet as the base model, the proposed techniques focuses on frame-by-frame analysis, semantic frame analysis, etc, with ensembling as the proposed method, for better prediction performance. The research improved the already proposed solution by implementing attention mechanisms and siamese training.

3.1.8 The DeepFake Detection Challenge (DFDC) Dataset

The DFDC Dataset is both the largest currently available Deepfake dataset, and one of only a handful of datasets containing footage recorded specifically for use in machine learning tasks. It contains 25tb of raw data, made by 3000+ candidates with a total of almost 50000 videos shot by them, most of them shot in 1080p. The subjects are split in divisions according to training and testing purposes, and all of the face swapped videos are created using a neural network model such as DFAE, MM/NN face swap, NTH, FSGAN, or StyleGAN, and then refined

3.1.9 FaceForensics++: Learning to Detect Manipulated Facial Images

This paper aims to create a benchmark for facial manipulation detection with a human baseline and a large scale dataset and using state of the art forgery detection methods. The paper hopes that as more manipulation methods appear, detection methods must also be developed that can detect forgeries in videos with little to no data.

3.1.10 One Detector to Rule Them All Towards a General Deepfake Attack Detection Framework

The primary focus in this paper is to investigate a training strategy that can detect deepfakes from multiple benchmark datasets and unknown generation methods. In this research, the development of the Convolutional LSTM-based Residual Network (CLRNet) was developed. Using single domain and merge learning techniques and building this network using Convolutional LSTM cells, the network presented an effective and practical transfer learning strategy in detecting multiple deepfakes simultaneously.

3.1.11 Deepfake Videos in the Wild: Analysis and Detection

The need to provide a comprehensive analysis of the various detection methods against a very large deepfake dataset (DF-W) to determine whether these methods are capable of detecting deepfakes in real world scenarios. The test was run with well known detection methods such as Xception, MesoNet, VA, etc. It was found out that the current methods produced poor results and detection performance, thereby noting of unsupervised learning such as transfer learning, which has shown good promise.

Chapter 4

Week 4 -Research study

4.1 Introduction

This week (6/09/21 - 9/9/21) we have analyzed and summarized these Papers[1] :

Week 3 - Research Papers		
Research Title	**Publication Date**	**Journal/ Conference**
Two-Stream Neural Networks for Tampered Face Detection	29 Mar 2018	2017-CVPR Workshop
Forensic Face Detection From GANs Using Convolutional Neural Networks	October 2018	Development of Intelligent Meeting Solution
Detecting Both Machine And Human Created Fake Face Images In The Wild	15 October 2018	Signal Processing 174 (2020) 107616
On the Generalization of GANs Image Forensics	December 10, 2019	CRIPAC
Beyond Facial Expressions: Learning Human Emotion from Body Gestures	10-13 Sept 2007	BMVC

[1] The link to the paper for reference has been added as a hyperlink in the column ”Journal/Conference” for the respective paper

4.1.1 Two-Stream Neural Networks for Tampered Face Detection

Current methods for tampering detection depend on a particular source of tampering evidence - for example : local noise analysis fails to deal with tampered images constructed using careful post processing and Color Filter Array (CFA) models cannot deal with resized images.

The Approach : The premise behind local noise estimation based techniques is that the difference between global noise characteristics and local noise characteristics reveals the hidden tampered regions. One stream is a CNN trained to classify if the image is real or fake.Other stream is a patch triplet stream, which is trained on steganalysis features of image patches with a triplet loss, models the traces left by in-camera processing and local noise characteristics.

Assumption: Different imaging devices produce different CFA regions with respect to tampered and authentic images.
One approach was to use light/illumination features from the images and use a SVM to classify tampered parts.But some techniques might use only small sections to change hence not much global change in the illumination.

The steganalysis feature (low level features) is a local descriptor based on coocccurrence statistics of nearby pixel noise residuals obtained from multiple linear and non-linear filters. First attempt was made for the above technique by modelling it as a **Gaussian Model**, which was later improved by treating the task **as anomaly detection** which used **a discriminative learning autoencoder outlier removing method based on steganalysis features**

Noteworthy for CNNs : **The performance degrades significantly when multiple post processing techniques are applied to tampered regions.**

For steganlaysis they have a Deep CNN that is based on triple loss - in order to ensure that two parts of the image from the same image are closer in the learned embedding whereas the difference between those from different images are large.

This is implemented by simply constraining the distances between the features of the patches by some margin m .

Formally, given Image patch x_a (anchor patch), a patch x_p (positive patch) from the same image, and x_n (negative patch) from a different image.

$$d(f(r(x_a)), f(r(x_p))) + m < d(f(r(x_a)), f(r(x_n)))$$

where $r(x)$ is the steganalysis features of patch x, $f(r(x))$ - is the embedding we want to learn of x, and $d()$ is the sum of squares distance measure.

Loss function is hence :

$$L(f) = \sum_{a,p,n} \max(0, m + d(f_a, f_p) - d(f_a, f_n))$$

In laymen terms, the triplet network tells if the two patches under study are from the same image or not .This means if the patches are from the same image then they are closer in distance based on the learned embedding and if they are farther then they come from a different image - hence tampered.

Combined metric : Steganlysis + SVM

$$F(q) + \lambda \frac{1}{N_q} \sum_{x \in q} S(x)$$

where N_q is the number of patches inside face q

Key approaches this two stream approach was tested against are as follows :

1. **Face Classification stream**

2. **Patch Triplet Stream**

3. **Steganalysis features + Linear SVM**

4. **CFA Pattern** -
   This method estimates the CFA pattern and uses a GMM algorithm to classify the variance of prediction error using the estimated CFA pattern. The output of this method is a local level tampering probability map. For the face region, an average probability is calculated as the final score.

5. **Improved DCT Coefficient** -
   This method estimates the DCT coefficients for all the regions in the given image to find the singly JPEG compressed regions and classifies them as tampered regions. The output of this method is a probability map indicating tampering

The end product analysed was the Class Activation Maps(CAMs) from the trained GoogLeNet.

4.1.2 Forensics Face Detection From GANs Using Convolutional Neural Network

Various models were generated based on GANs such as Deep Convolutional GANs (DC-GAN) which generates image size of 64 X 64 and Progressive training (PG-GANs) which generate image of size 256 X 256 and 1024 X 1024. Network Architectures like AlexNet, VGG-Net, ResNet, SENet are the backbone of face feature extraction and representation. The architecture used here is VGGFace deep feature extraction through VGG-Net. The VGG-Net consists of five-layer blocks including convolutional and max-pooling layers in each block for feature extraction task. The fully connected layer connects to the K-way SoftMax (where K is the number of classes) to output the correct probability of the corresponding face identities. After extraction of features fine tuning is performed by adding a new fully connected layer after feature representation blocks, which is connected in 2-way SoftMax for real/fake binary classification.

4.1.3 Detecting Both Machine And Human Created Fake Face Images In The Wild

Diverse detection techniques available but most based on meta data or artifacts that can be cloaked. Paper hopes to human created and machine created face images with neural networks using ensemble methods.
Steps:

1. Developing ensemble based neural network classifier for GANs detection

2. Developing neural network classifier to detect sophisticated human-created fake face images

3. Proposing a fully automated effective end-to-end fake face detection pipeline

The Approach : This work aims on distinguishing GANs generated faces from real faces, and developing a fundamental enabling classifier technology to detect with high accuracy. They also design ensembles of various CNN-based classifiers to detect GAN-created face images.
For detection of human made fake images, there aren't many available datasets related. Hence, a sample dataset was made by the team, taking inputs from google and bing image search and then using noncommercial reuse with modifications search settings. Hard samples taken such as people with strong make up, glasses, hats etc. In detection, meta data is omitted due to easy manipulation of it and only RGB channel information is utilized. The detection method is divided into two stages, 1) to perform pre-processing to crop and filter face regions, 2) Once cropped and

aligned faces are obtained, the classifier model is trained to distinguish fake images created by humans from unmodified real images.

Detection algorithms and neural networks used for distinguishing human-created fake faces from real faces

- SeetaFace engine, MTCNN (Best performance), YOLO, Dlib

- Noise filtering algorithm used to reduce false positives of detected face regions

Evaluation - Different deep learning models were trained using keras python deep learning library and performance was evaluated using the previously mentioned test dataset. Various ensembles of methods were used to provide diversity to the discriminators. Out of all the neural networks, ShallowNet outperformed the other neural networks. In human created image detection, it was found out that XceptionNet outperformed the other DNN while the others didn't perform well. It also turned out that with more training examples the higher resolution images can yield better result on the methods used.

Conclusion - The preliminary results gave promising results in detection deepfakes, especially in GANs generated fake images. With more training data, the performance can only be improved.

4.1.4 On the Generalization of GAN Image Forensics

Image preprocessing methods were involved Gaussian Blur and Gaussian Noise in training phase to enhance the generalization ability of the forensic convolutional neural networks model. Purpose of general forensic method is to enhance high frequency pixel noise and to focus on the clues in low level pixel statistics. Purpose of general forensic method is to enhance high frequency pixel noise and to focus on the clues in low level pixel statistics. Image preprocessing step in the training stage is used to destroy low level unstable artifact GAN images and force the forensic discriminator to focus on more intrinsic forensic clues. The kernel size of Gaussian blur is chosen randomly from 1,3,5 and 7 for each training batch. The standard gaussian noise is randomly set between 0 and 5. For the four convolutional layer , we use Batch Normalization except the first layer, and use Leaky Rectified Linear Unit activation functions that introduce non-linearities. The Loss function and and Optimization algorithm are Binary Cross Entropy Loss and Adaptive Moment Estimation respectively.

4.1.5 Beyond Facial Expressions: Learning Human Emotion from Body Gestures

In this paper, affective body gesture analysis in videos was investigated, a relatively understudied problem.Spatial-temporal features are exploited for modeling of body gestures. They also present to fuse facial expression and body gesture at the feature level usingCanonical Correlation Analysis.The current spatial-temporal features based video description does not consider the position relations of cuboids detected.

Spatial-Temporal Features - We extract spatial-temporal features by detecting space-time interest points in videos. we calculate the response function by application of separable linear filters. Assuming a stationary camera or a process that can account for camera motion, the response function has the form:

$$R = (I \cdot g \cdot h_{ev})^2 + (I \cdot g \cdot h_{od})^2$$

where $I(x, y, t)$ denotes images in the video, $g(x, y; \sigma)$ is the 2D Gaussian smoothing kernel, applied only along the spatial dimensions (x, y), and h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters applied temporally, which are defined as

$$h_{ev}(t; \tau, \omega) = -cos(2\pi t\omega)e^{\frac{-t^2}{\tau^2}}$$

$$hod(t; \tau, \omega) = -sin(2\pi t\omega)e^{\frac{-t^2}{\tau^2}}$$

In all cases we use $\omega = \frac{4}{\tau}$. The two parameters σ and τ correspond roughly to the spatial and temporal scales of the detector. Each interest point is extracted as a local maxima of the response function.Any region with spatially distinguishing characteristics undergoing a complex motion can induce a strong response, while region undergoing pure translational motion, or areas without spatially distinguishing features, will not induce a strong response.

- At each detected interest point, a cuboid is extracted which contains the spatiotemporally windowed pixel values. The side length of cuboids is set as approximately six times the scales along each dimension, so containing most of the volume of data that contribute to the response function at each interest point. After extracting the cuboids, the original video is discarded, which is represented as a collection of the cuboids. To compare two cuboids, different descriptors for cuboids have been evaluated, including normalized pixel values, brightness gradient and windowed optical flow, followed by a conversion into a vector by flattening, global histogramming, and local histogramming. As suggested, we adopt the flattened brightness gradient as the cuboid descriptor. To reduce the dimensionality, the descriptor is projected to a lower dimensional PCA space. By clustering a large number of cuboids extracted from the training data using the K-Means algorithm, we derive a library of cuboid prototypes. So each cuboid is assigned a type by mapping it to the closest prototype vector. we use the histogram of the cuboid types to describe the video.

Recognition : SVM -

The Support Vector Machine (SVM) classifier to recognize affective body gestures. For the cases where it is difficult to estimate the density model in high-dimensional space, the discriminant approach is preferable to the generative approach.
Given a training set of labeled examples $(x_i, y_i), i = 1, ..., l$ where $x_i \in \mathrm{R}^n$ and $y_i \in 1, 1$, a new test example x is classified by the following function:

$$f(x) = sgn(\sum_{i=1}^{l} {}_{i} y_i K(x_i, x) + b)$$

where $_i$ are Lagrange multipliers of a dual optimization problem that describe the separating hyperplane, $K(,)$ is a kernel function, and b is the threshold parameter of the hyperplane. The training sample x_i with $_i > 0$ is called the support vector, and SVM finds the hyperplane that maximizes the distance between the support vectors and the hyperplane. Given a non-linear mapping ϕ that embeds the input data into the high dimensional space, kernels have the form of

$$K(x_i, x_j) = h \cdot \phi(x_i) \cdot \phi(x_j) \cdot i.$$

Adavntage of using SVM here is that it allows domain-specific selection of the kernel function, and the most commonly used kernel functions are the linear, polynomial, and Radial Basis Function (RBF) kernels.

Canonical Correlation Analysis for Fusing Facial and Body Gestures: :

CCA is a statistical technique developed for measuring linear relationships between two multidimensional variables. CCA to find corresponding points in stereo images was applied. CCA to model the relation between an object's poses with raw brightness images for appearance-based 3D pose estimation was suggested. A method using CCA to learn a semantic representation to web images and their associated text was proposed. Given two zero-mean random variables $x \in R_m$ and $y \in R_n$, CCA finds pairs of directions w_x and w_y that maximize the correlation between the projections x and y are called canonical variates.

$$\rho = \frac{E[xy]}{\sqrt{E[x^2]E[y^2]}} = \frac{E[W_x^T xy^T W_y^T]}{\sqrt{E[W_x^T xy^T W_x^T]E[W_x^T yy^T W_y^T]}} = \frac{W_x^T C_{xy} W_y}{\sqrt{W_x^T C_{xx} W_x W_y^T C_{yy} W_y}}$$

where $C_{xx} \in \mathbb{R}^{m\times m}$and $C_{yy} \in \mathbb{R}^{n\times n}$are the within-set co-variance matrices of x and y respectively and $C_{xy} \in \mathbb{R}^{m\times n}$ denotes their between-sets co-variance matrix.

Evaluation :
To evaluate the algorithms' generalization ability , a 5-fold cross-validation test scheme in all recognition experiments is adopted. That is the data set is divided randomly into five groups with roughly equal number of videos, and then used the data from four groups for training and the left group for testing; the process was repeated five times for each group in turn to be tested. The average recognition rates here. In all experiments, the soft margin C value of SVMs set to infinity so that no training error was allowed. With regard to the hyper-parameter selection of RBF kernels, we carried out grid-search on the kernel parameters in the 5-fold cross-validation.

Chapter 5

Week 5 - Recapitulation

5.1 Papers studied

This week we have (20/09/21 - 24/09/21) we have been working on the preparing a coherent summary for all the insights from Week 2 through Week 4,which is still a work in progress.Hence we decided to finish off the remaining papers.

1. **Deepfake Video Detection Using Convolutional Vision Transformer**
2. **DeepFakes and Beyond: A survey of Face manipulation and Fake Detection**
3. **The DeepFake Detection Challenge (DFDC) Dataset**
4. **Video Face Manipulation Detection Through Ensemble of CNNs**
5. **Deep-Fake Forensics via An Adversarial Game**
6. **Deep-Fakes Detection Techniques Using Deep Learning: A Survey**
7. **Subjective and Objective Evaluation of Deep-Fake Videos[5]**
8. **Two-Stream Neural Networks for Tampered Face Detection[10-29]**
9. **Countering Malicious DeepFakes: Survey, Battleground, and Horizon**
10. **ArcFace: Additive Angular Margin Loss for Deep Face Recognition**
11. **Celeb-DF: A Large - Scale Challenging Dataset for DeepFake Forensics**
12. **Deepfake Videos in the Wild: Analysis and Detection**

13. **Deepfakes Detection with Automatic Face Weighting [8]**

14. **Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations**

15. **Forensics Face Detection From GANs Using Convolutional Neural Networ**

16. **On the Generalization of GAN Image Forensics**

17. **Beyond Facial Expressions: Learning Human Emotion from Body Gestures**

18. **Deep Learning for Deepfakes Creation and Detection: A Survey**

19. **Deep-Fake: a social construction of technology perspective**

20. **DeepFakes: perspectives on the future "reality" of advertising and branding**

21. **Emotions don't lie: An audio visual**

22. **DeepFake detection methods using affective cues.**

23. **Generalization Of Audio DeepFake Detection**

24. **Deepfake Video Detection Using Recurrent Neural Networks**

25. **Deepfake Videos in the Wild: Analysis and Detection**

26. **Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News[10-4]**

27. **Detecting GAN generated Fake Images using Co-occurrence Matrices**

28. **FaceForensics++: Learning to Detect Manipulated Facial Images**

29. **MesoNet: a Compact Facial Video Forgery Detection Network**

30. **One Detector to Rule Them All Towards a General Deepfake Attack Detection Framework**

5.2 Remaining Papers from the list

5.2.1 One Detector to Rule Them All Towards a General Deepfake Attack Detection Framework

Focuses on developing a generalized approach to detect multiple types of DFs, including deepfakes from unknown generation methods such as DeepFake-in-the-Wild (DFW) videos.A Convolutional LSTM-based Residual Network (CLRNet), which adopts a unique model training strategy and explores spatial as well as the temporal information in a deepfakes is used. A unique method is proposed an effective, multiple training strategy (Merge- vs. Transfer-learning).

One of the main observations include that - the insights that have been learned in one domain (e.g., FaceSwap) can be used to enhance the generalizability in another domain (e.g., Face2Face).Also the paper investigated the use of an autoencoder as a method for learning better semantic representation with zero-shot or few-shot learning.

5.2.2 MesoNet: a Compact Facial Video Forgery Detection Network

This paper follows a deep learning approach and presents two networks, both with a low number of layers to focus on the mesoscopic properties of images.

The Meso4 network begins with a sequence of four layers of successive convolutions and pooling, and is followed by a dense network with one hidden layer. To improve generalization, the convolutional layers use ReLU activation functions that introduce non-linearities and Batch Normalization to regularize their output and prevent the vanishing gradient effect, and the fully-connected layers use Dropout to regularize and improve their robustness.

MesoInception4 ,on the other hand, is an alternative structure consists in replacing the first two convolutional layers of Meso4 by a variant of the inception module.