A Project Report

on

DeepFake: A Threat to our society

Submitted in partial fulfillment of the

requirement for the award of the degree of

B.Tech. Computer Science and Engineering

Department Of Computer Science And Engineering

Submitted By

18SCSE1010345 –HIMANSHU YADAV

18SCSE1010060 – SHASHANK SINGH

Under The Supervision of
Prof.(Dr.) Shiv Kumar Verma
Department of Computer Science and Engineering

School Of Computing Science And Engineering

Galgotias University, Greater Noida, India December - 2021

DeepFake: A threat to our society

Abstract- The deepfake make use of deep learning to manipulate images and videos. The fast progress of deep learning techniques with the free access of databases have led to creation of realistic fake content.
This paper provides the large impact of deepfake on our society, different defect methods and methods to overcome this problem by detecting these kinds of manipulations. The different deepfake methods could be using a software like face swap or using sites like deepfakeweb to create a deepfake video. The methods for deepfake detection will be based on two models: (I) image detection models, (II) video detection models, for example a very easy way to detect defects is by fake audio detection.
In addition to this we have also discussed future issues for these kinds of problems.

Introduction- The term "DeepFake" was originated after a reddit user named "deepfakes" claimed in 2017 to have developed a machine learning algorithm that helped him to transpose celebrity faces into porn videos. A sort of artificial intelligence used to create convincing images, audio, and video frauds is known as deep fake.. They can be used to spread false information for example like election propaganda. Deepfakes can be created by using two AI algorithms- one is called generator and the other is called discriminator. The generator helps in the creation of fake multimedia content and discriminator to determine whether the content is fake or real. Together they form to be known as GAN (generative adversarial network). The first step in establishing a GAN is to take desired output and create a training dataset for the generator. Once the generator creates a output we can fed video clips to the discriminator.
As generator gets better at creating fake video clips, the discriminator gets better at spotting them. Before it was difficult to create fake videos, however, now they don't need any skills to create it as AI greatly reduces the effort. Unfortunately, this means that anyone can create a deepfake to promote their hidden agenda. As a result, open software and mobile applications such as ZAO and FaceApp have been released opening this door to anyone. However various deepfake detection software are out there like example- Microsoft revealed a new AI- powered deepfake detection software to combat altered images and videos.
This paper provides in-depth review of digital manipulation techniques applied to facial content e.g. generation of fake news that would provide misinformation in elections. We will cover five types of GAN software (1) Vanilla GAN (2) Conditional GAN (3) Deep convolutional GAN (4) Laplacian Pyramid GAN (5) Super Resolution GAN . We will also discuss future issues for this kind of problem.

Literature survey- In order to become familiar with the background a literature survey was performed. Previous work was thoroughly studied as a result of this survey. It was concluded from literature survey that a lot of work was done in this field.
In this survey, depth review of digital manipulation techniques are introduced, various softwares like ZAO and FaceApp are discussed and the ways to stop them are also covered here. We have also discussed generative adversarial network (GAN) and future issues of this problem.

Project Design-

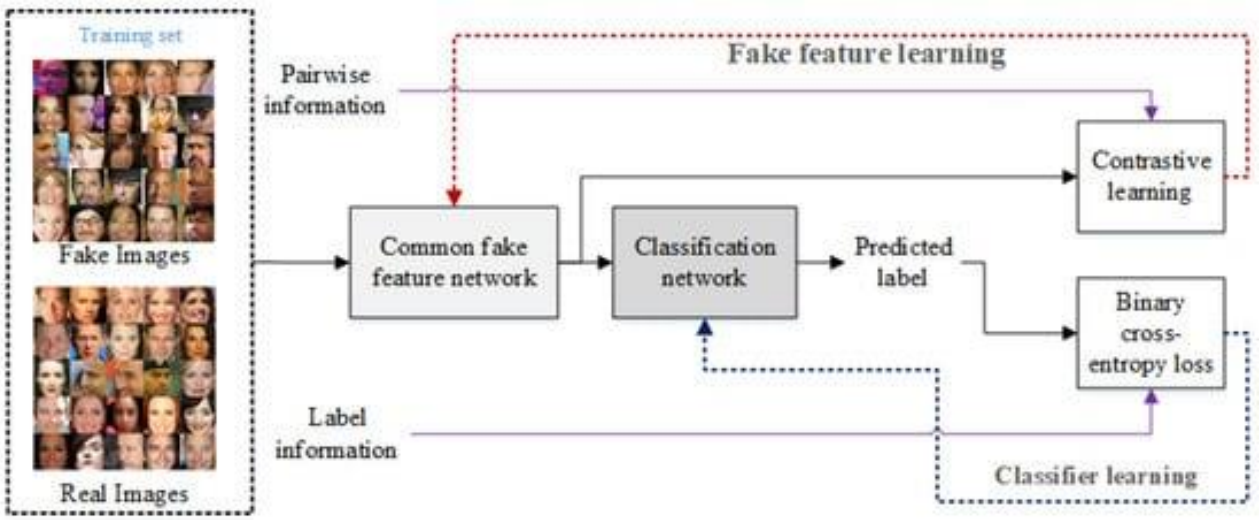

Fig 1. Flowchart for deepfake detection

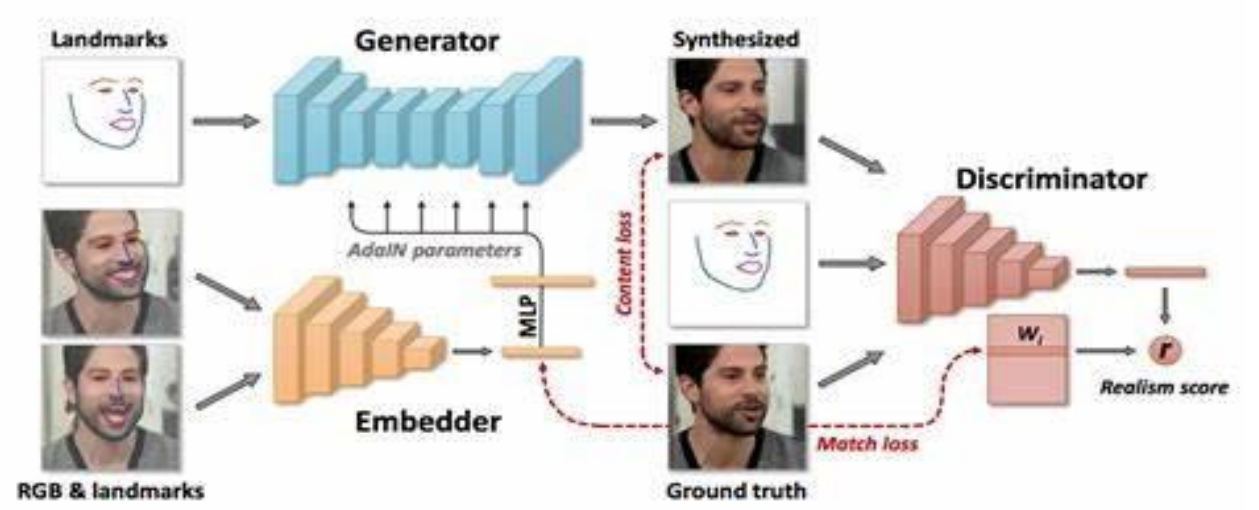

Fig 2. Deepfake architectural diagram

Advantages of DeepFake-

1. **Bringing back the loved ones-** Instead of using CGI and VFX we can bring back dead celebrities by using deepfake. This method is cheaper and easier than using CGI/VFX.
2. **More Realistic scenes in movies-** With tremendous acting talents, Henry Cavill brought Man of Steel to life in a Justice League film, but there was an issue. He was also on set for Mission: Impossible: Fallout at the same time. In MI:6, his character wore a moustache. But, as you know, Superman doesn't have a moustache! His moustache was not allowed to be shaved by the studio. The shootings took place at the same time. As a result, they opted to use CGI to remove his moustache. And in this manner, he was able to play both roles without difficulty, although it was a disaster. The CGI was abysmal and strange-looking. It rattled the Internet and turned into a humorous figure. When a YouTube user used Deepfakes to get rid of his moustache, it went viral.
3. **Chance of getting education from its masters-** Imagine a world where we can really study physics from Albert Einstein at any time as well as from anyplace. Deepfake makes the unachievable achievable. Learning from the masters is a great way to stay motivated. We can make it more efficient, but there's still a long way to go.
4. **Training of bots**- We can train bots to detect deepfake videos and have them working as a discriminator.

Disadvantages of DeepFake-

1. **Scamming -** Sadly, there are numerous ways to use this powerful technology to scam people. We've recently seen some examples of this.
2. **Generating fake news**- Despite the fact that he was not confirmed to be in the film, footage of actor Andrew Garfield on the set of New Way Home has emerged. Fans expected him to appear in the film, but neither Sony nor Marvel made an official announcement. After this footage was leaked, the internet went crazy. Everyone was debating whether or not this video was genuine. Some believed that it was a Deepfake, but nobody really knew for sure. It was extremely accurate.
3. **Spreading misleading news via politician-** These days, it's difficult to tell what's real and what's not. We are so easily persuaded by what we see or hear in the media. And, if you see a public figure discussing an issue, you're not inclined to wonder if they're "real" or "fake." However, some can use Deepfake. Someone with nefarious motives can readily manipulate the media. This can result in warfare, anarchy, and even starvation.
4. **Privacy Problem -** We all have accounts on social media. Facebook, Instagram, Twitter and Snapchat are all popular social media platforms. Every day, we all generate massive amounts of data. We all post images taken from various perspectives and in various moods. This could cause some privacy issues. A malicious Deepfaker can simply gain access to your images, take them without your consent, and utilise them.

Generative adversarial network- GAN is a class of machine learning frameworks designed by Ian Goodfellow and his colleagues in June 2014

GAN is a framework used for creating deepfakes. It consists of 2 Parts:

1. Generator
2. Discriminator

Different types of GANs:

1. **Vanilla GAN:** This is the most basic GAN kind. Simple multi-layer perceptrons serve as the Generator and Discriminator in this example. The algorithm in vanilla GAN is really simple: it uses stochastic gradient descent to try to optimise the mathematical equation.
2. **Conditional GAN (CGAN):** CGAN is a deep learning approach that includes several conditional parameters. In CGAN, the Generator has a variable 'y' that is used to generate the corresponding data. Labels are also included in the Discriminator's input to aid the Discriminator in distinguishing between real and falsely generated data.
3. **Deep Convolutional GAN (DCGAN**): DCGAN is one of the most widely used and successful GAN implementations. It is made up of ConvNets rather than multi-layer perceptrons. Convolutional stride is used instead of max pooling in the implementation of ConvNets. In addition, the layers are not completely joined.
4. **Laplacian Pyramid GAN (LAPGAN**): The Laplacian pyramid is a linear invertible picture representation made up of a series of octave-spaced band-pass images and a low-frequency residual. This method employs a large number of Generator and Discriminator networks as well as different Laplacian Pyramid levels. This method is popular because it creates extremely high-quality photos. The image is initially down-sampled at each layer of the pyramid, then up-scaled at each layer in a backward pass, gaining some noise from the Conditional GAN at these layers until it reaches its original size.
5. **Super Resolution GAN (SRGAN):** SRGAN is a method of creating a GAN in which a deep neural network is combined with an adversarial network to produce higher resolution images, as the name suggests. This sort of GAN is especially beneficial for upscaling native low-resolution images to improve their details while minimising mistakes.

Generator- The generator helps in the creation of fake multimedia content. By incorporating input from the discriminator, the generator element of a GAN learns to create bogus data. It learns to make the discriminator classify its output as real.

The section of the GAN that trains the generator consists of the following elements:

- random input
- generator network, which converts the random input into a data instance
- discriminator network, which organizes the generated data
- discriminator output
- generator loss, which penalizes the generator for not fooling the discriminator

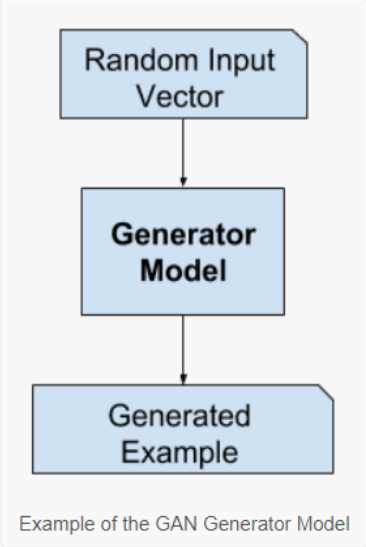

Fig 3

Discriminator- It determines whether a content is real or fake. As generator keeps getting better at creating fake videos, the discriminator gets better at spotting them.

The training data for the discriminator originates from two places:

- **Real data** instances, such as real-world photographs of people.
- **Fake data** instances created by the generator, which the discriminator uses as positive examples during training. During training, the discriminator uses these situations as negative examples.

During discriminator training:

1. The discriminator identifies both genuine and fake data from the generator during discriminator training.
2. The discriminator is penalised if a real instance is misclassified as fake or a fake instance is misclassified as real.
3. The discriminator's weights are updated via back propagation from the discriminator loss via the discriminator network.

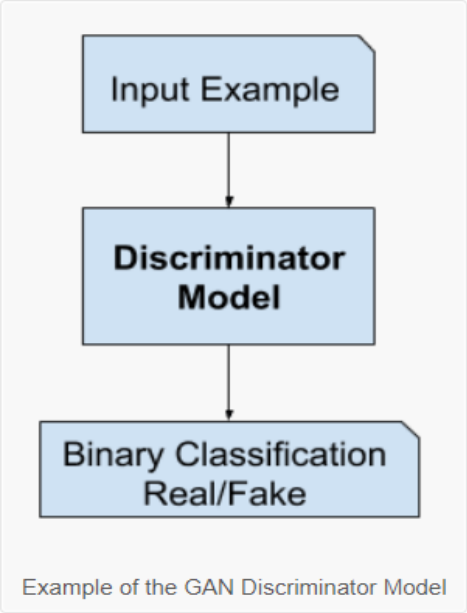

Fig 4

Convergence - Since the discriminator can't identify the difference between real and phony, as the generator improves with training, the discriminator's performance deteriorates. The discriminator has a 50% accuracy if the generator succeeds flawlessly. To make its forecast, the discriminator essentially flips a coin.

This trend causes a difficulty for the GAN's overall convergence: the discriminator feedback becomes less useful over time. If the GAN continues to train after the discriminator has given fully random feedback, the generator will begin to train on trash feedback, and its own quality will deteriorate.

Convergence is generally a temporal rather than a permanent state for a GAN.

Methods to prevent deepfake

1. **Legal protections:** Passing of laws that ban the creation of deepfake videos. We can ask government of India to create a new bill regarding this

2. **Deepfake audio detection:** Many deepfake films can be identified merely by their audio, as it is difficult to copy audio in these types of videos. As a result, we can employ audio detection to guard against deepfake. We can even use AI to detect whether or not a video is false by capturing the lip movement.

3. **Use of blockchain technology:** This technology has the potential to make our internet information less vulnerable to hackers. Furthermore, the traceable keys associated with each block on a blockchain may aid in establishing which files are genuine and which have been tampered with. Moreover, blockchains are resistant to a wide range of security vulnerabilities that centralised data storage are susceptible to. Although distributed ledgers are not currently capable of storing large amounts of data, they are suitable for storing hashes and digital signatures. Individuals may, for example, use the blockchain to digitally sign and confirm the authenticity of a video or sound document that is personal to them. The more people that sign that film with their digital signature, the more likely it is to be deemed a true record. This is clearly not the best option. Additional procedures will be required to assess and factor in the capability of those who vote on a document.

4. **Train computers to spot fakes:** Using obvious antiquities or heuristic analysis, some of the current imperfect deepfakes can now be identified. Microsoft has unveiled a new method for detecting faults in synthetic media. The Defense Advanced Research Projects Agency, or DARPA, is working on a tool called SemaFor that aims to discover semantic inadequacies in deepfakes, such as a snapshot of a man with anatomically incorrect teeth or a person wearing socially unusual jewellery.

5. **Make teleconference calls private**- Ensure that all video conversations and webinars are only accessible to trustworthy personnel by implementing security measures such as password protection. If other photographs or videos of company executives are available online, this won't completely eliminate the risk, but it will reduce the quantity of material available to scammers, making it more difficult to create a convincing deepfake.

Future of Deepfake - The creation and detection of deepfakes will become increasingly fierce in the future. Deepfakes will continue to grow and expand, and problems such as a lack of details will be addressed. New algorithms that can give higher levels of realism and run in real time have already been seen. As a society, we must educate ourselves to only trust content from trustworthy sources. With more public knowledge, we may be able to learn how to mitigate the bad effects of deepfakes, coexist with them, and benefit from them in the future.

Conclusion- In this report we have talked about what is deepfake, why is it majorly bad for our society, it's advantages and disadvantages. We also talked about the algorithm on which deepfake runs on which is Generative adversarial network (GAN) and talked about it's two components generator and discriminator. The generator helps in the creation of fake multimedia content. The generator component of a GAN learns to generate fake data by incorporating discriminator feedback. It learns to manipulate the discriminator so that its output is classified as real. Discriminator determines whether a piece of content is genuine or fake. As generator keeps getting better at creating fake videos, the discriminator gets better at spotting them. We talked about the convergence problem which GAN may be facing sometimes

We also talked about the different types of GAN which includes Vanilla GAN, Conditional GAN, Deep Convolutional GAN, Laplacian Pyramid GAN, Super Resolution GAN and methods to prevent deepfake which included making of new laws, audio detection, use of blockchain technology, training of computers to spot fakes. Lastly we talked about the future of deepfake. To sum up this research, deepfakes allow for the fabrication of media, often without consent, resulting in psychological trauma, political instability, and corporate upheaval. Deepfakes' weaponization might have a huge influence on the economy, personal liberty, and national security. To raise awareness and stimulate growth and innovation, deepfake danger models, harm frameworks, ethical AI principles, and reasonable legislation must be developed through collaborations and civil society oversight.

References-

[1] Advantages and disadvantages from - Advantages and Disadvantages of DeepFake Technology | aiTechTrend

[2] Different types of GAN from - Generative Adversarial Network (GAN) - GeeksforGeeks

[3] Generator from- The Generator | Generative Adversarial Networks | Google Developers

[4] Discriminator from - The Discriminator | Generative Adversarial Networks (google.com)

[5] Convergence from - GAN Training | Generative Adversarial Networks | Google Developers

[6] Methods to prevent deepfake from - Best Ways to Prevent Deepfakes (analyticsinsight.net)

[7] Pictures from-

Fig 1 from- Applied Sciences | Free Full-Text | Deep Fake Image Detection Based on Pairwise Learning | HTML (mdpi.com)

Fig 2 from- Detecting Deepfake Videos: An Analysis of Three Techniques | DeepAI

Fig 3 and Fig 4 from- Introducing GAN - BLOCKGENI