1. **INTRODUCTION**

Over many years, Images and videos has been considered as a proof of existence of something or a memory of actions that has taken place. But after the emergence of technology that can manipulate digital media, should we still believe what we see. Image/Video processing software like Adobe Photoshop can alter an original image by adding contents that were not originally present in the photo or removing some details, as a result the altered image might give a different impression from the original one to the viewer.

Recent developments in Machine learning, have made availability of technology for manipulating images and videos easily accessible with little knowledge needed. By using deep learning techniques, machines are able to learn the features of the original image/video and then generate a new sample with similar features but not exactly the original sample aiming at eluding the viewer into believing it is real while in fact it was fabricated by the machine, this is known as Deepfake. And once a model is built, it can be used by anyone with their own dataset

The motivation behind creating fake images/videos is to misinform the viewer on the actual content of the image/video. The techniques are used in film productions to create sensational visual effects or in magazines to display perfect photos. The technology can also be used for malicious intent like misleading the public, revenge, influencing a viewers’ emotions and attitude, political influence, defamation.

The existence of social media platforms make spread of false information very fast and the ease of use and inexpensive technologies to manipulate digital media makes this a dangerous prospect, fake images/videos have high negative impact to the society because of their potential to influencing a viewers’ emotions and attitude which might result into dangerous outcomes.

In order to identify these manipulated images/videos, we must first learn how they are generated. Generative Adversarial Network which is technique used to generate deepfakes and then develop an algorithm that can identify fake/doctored images.

1. **BACKGROUND**

The technology to manipulate digital media has evolved and undergone so much growth in recent years, now it is cheaper and readily available for anyone who fancy to use it. This has generated a serious concern for the public safety because manipulated digital media is the main cause of fake news and especially with availability of social media, the spread of the fake news is at exponential rate and sometimes the damage can be irreversible to repair.

**2.1 Fake images**

**Definition of a fake image**

Fake image is an image that has been manipulated/doctored to alter the contents of the original image, this can involve addition/reduction or modification of contents in the original image by using image processing software (e.g. Adobe Photoshop)

A fake image can also be described as a synthetic image generated by machine learning techniques, it is an image consisting of fabricated details but have features that appear similar to the original image (having the same distribution as the original image)

**Pristine Image**

An image that is in its original state and has all information as captured, of which no contents have been removed or added. Only resizing is considered as an operation which does not alter the information present in the image

**Examples of doctored Images**

Here is an image of a cycler and bear, as it can be seen in the photo on the left a bear running towards the cycler, it appears as the bear is chasing the cycler.



Figure 2.3 A bear chasing a cycler on the left is a doctored image of a bear running alone on the road(right)

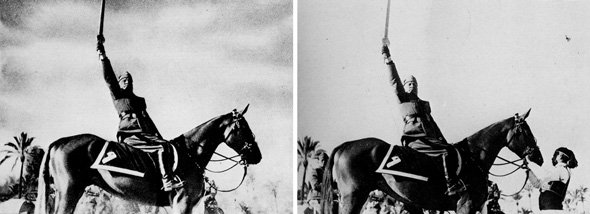
The following images are of Benito Mussolini the Former Prime Minister of Italy

Figure 2.2 He had the horse handler removed on the left photo, so that the photo could look better and more heroic to his people

**2.2 Why manipulation of images/videos?**

Digital media manipulation has been around for many years, it has gained more popularity in recent years because the technology is easily accessible and can be used by anyone with interest, even non-experts. As a result there are social, economic, political and religious implications because of manipulation of images and videos. But not all manipulations have negative impacts, some industries use the technology for increasing creativity and achieving an otherwise impossible tasks if it weren’t for digital media manipulation. I will discuss the negative and positive impacts of image and video manipulation to the society

Implications of manipulated images/videos

1. Political influence

Manipulated images/videos can give a wrong impression of the contents of the media and thus delivering unintended message. If the image/video is believed to be real it can change voters perception of a political party or make them turn against their own party

1. Fake pornography

posting sexual images or videos of another person on the internet without consent by using face-swap technology to digitally manipulate pornography so it looks like other people are present in images or film. When there are many photos of the target person like celebrities it increases the chances of the pornographic photograph or film to look realistic

1. Government propaganda

This involves manipulation of event photos or videos information, ideas, or rumors to deliberately spread non-factual information, ideas or rumors that widely influences opinions, usually to preserve the self-interest of a nation.

1. Visual enhancements

Creative Hollywood cinematic productions, Some great cinematic moments are due to special effects added to the original shot for example in an event an actor is not able to finish acting the film e.g because of illness, death or termination of contract. The creators can use deepfake to fill the gaps for the actor. A good example is when an actor Paul Walker tragically passed away before filming of the movie Fast 7 was finished, the production used deepfake to make his presence available even though it was never him behind the cameras

1. Social media apps

A user can create a version of themselves the way the want it. An overweight person can get a slim version of himself/herself. Recently there is an app called FaceApp going viral, users can create an old/younger version of themselves using the app also they can add or remove some features and appear as they wish.

**2.3 Image Processing**

……….to be filled……….

**2.3.1 CNN**

………to be filled…………

**2.3.2 Deepfake**

A combination of ‘’deep learning’’ and ‘’fake’’ is an Artificial Intelligence technique used to create fake images/videos, it is used to swap existing/original parts of source image/video with a target image/video by using deep learning technique known as generative adversarial network (GAN). Therefore showing an altered version of the original image/video, displaying something/someone that was never actually there or an action that did not take place.

**2.3.2.1 Generative Adversarial Network**

Is an algorithm used in deep learning consisting of two neural networks that acts as two players competing against each other, these networks are trained in adversarial manner. One player is called generative model which generates synthetic data that will come from the same distribution as the training data, the other player is the discriminator model that works to classify whether the data is real (from training dataset) or fake (generated by the generator)

1. Generative model

Generative model (represented as G) to be trained on training data X that is sampled from some true distribution D, is the one given some random distribution Z produces a distribution D’ which is close to D according to some closeness metric. Mathematically, z ∼ Z maps to a sample G(z) ∼D’

1. Discriminator model

Discriminator model uses supervised machine learning technique to classify a data if it is real or fake by assigning a probability. 1 being real and 0 being fake

**The training process**

Generative samples G(z) from generative model which are fakes are fed as input to the Discriminator together with sample from training data which are the real data. The Discriminator then outputs a value indicating whether the image appears to be real or fake. The goal of the generative model is to fool the discriminator that its data are real, and the role of the discriminator is to spot this and distinguish between the real and fake. The discriminator learns by back-propagation, then adjust its bias and weight so that it can output the value of 1 for real data and 0 for fake data. The generator also adjust its bias and weight in an opposite direction as opposed to the adjustment made by the discriminator, because the generator tries to trick the discriminator into outputting a value of 1 for the fake data as well. The GAN model is considered successful when the discriminator outputs 0.5 for the fake data, meaning the discriminator is not sure if it is fake or real which means the generator model produced samples that are very close to the true distribution of the training dataset

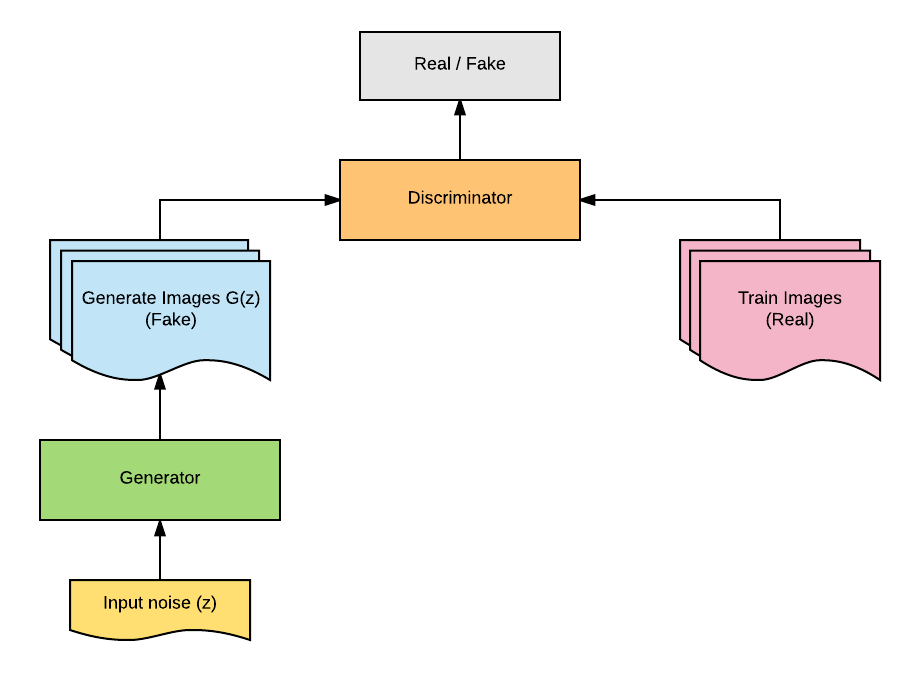


Figure 2.3 shows the the training process of the Generative Adversarial Network

1. **STATE OF THE ART**

**Current methods for identifying fake images**

There are three kind of methods for detecting fake images, each method has their pros and cons. I will define and explain about the current detection techniques in each method

**Feature-based** — where there is a kind of artifact in a certain (or multiple) types of forgeries — mostly applicable for methods that look for part of the image with contents that has been manipulated.

**Supervised learning** — using deep learning classifiers (mostly CNN) to learn certain types of fake images applicable for content-based feature as well as generative models, e.g GANs.

**Unsupervised learning** — an attempt to capture some essence of a genuine image, to detect new kinds of forgeries (that the model hasn’t seen before). It can be seen as a kind of anomaly detection.

* 1. **Feature based**

Techniques in this method uses content-based features on an image to determine if it has been manipulated. Techniques in this method can be placed into five categories

1. **Camera based techniques**

In 2004, Hany Farid and Alin Popescu have published the work ‘’Exposing Digital Forgeries in Color Filter Array Interpolated Images’’ about identifying fake images, using digital artifacts. Digital cameras have different artifacts stemming either from the photographing hardware, software or from the compression techniques, which are image specific. Farid and Popescu have used a specific camera filter (CFA) to identify fake parts of the images

Most digital cameras, for example, employ a single sensor in conjunction with a color filter array (CFA), and then interpolate the missing color samples to obtain a three-channel color image. This interpolation introduces specific correlations which are likely to be destroyed when tampering with an image.

1. **Pixel-based techniques**

Detection of statistical anomalies introduced at the pixel level, in the paper ‘’Pixel based Image Forensic Technique for copy-move forgery detection using Auto Color Correlogram’’ by Ashwini V Malviyaa and Siddharth A Ladhakeb.

The proposed method has the following steps:

* Preprocess the image by filtering noise and divide the image into M x N blocks.
* Subject each block to 8Z affine transformation.
* Feature extraction of each block by extracting Auto Color Correlogram of each block.
* Finding match using similarity measure to detect the forgery.

**c) Format-based technique** (to be filled)

**d) Physics-based technique** (to be filled)

**e) Geometry-based technique** (to be filled)

**3.2 Supervised learning**

In this work, researchers design a special convolutional layer, which by putting a constraint on the filters, is intended to capture manipulations instead of image semantic content. Tested on different retouching methods such as median filtering, Gaussian blurring and more, the method reached >95% accuracy. This work is intended to be universal, however, it’s designed allegedly limits it to photoshopping and tampering, not to GANs and the like.

(to be filled)

* 1. **Unsupervised learning**

There are several techniques in unsupervised learning method,

1. **Self-consistency** Fighting Fake News: Image Splice Detection via Learned Self-Consistency by Minyoung Huh, Andrew Liu, Andrew Owens, Alexei A. Efros 2018 (to be filled)
2. **Forensic transfer** (to be filled)
3. **Noise print** (to be filled)
4. **The deep fake spin** (to be filled)
5. **GANs artificial fingerprints** (to be filled)