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. Technology 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 the same as 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.

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.

**Fake image**

An image that has been manipulated to alter the original image or generated by machine learning technique to fits the distribution of 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.

**Examples of manipulated Images over the years**

Figure 1. In the 1930s a Russian leader once took a photo with his secret police official, but after the arrest and allegedly murder of the police Yezhov, this photo was doctored to remove the police.



Figure 2.1 Original photo (right), Yezhov is on the right of the photograph. In manipulated photo (left) he was removed.

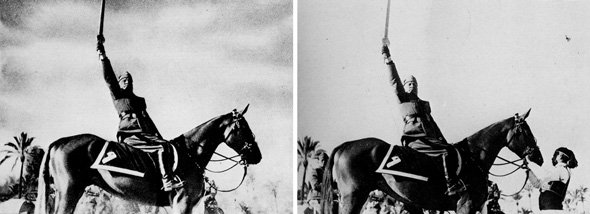
The following images are of Benito Mussolini the Former Prime Minister of Italy

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

**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. The word deepfake is a combination of word deep learning and fake.

**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 which generates data that will come from the same distribution as the training data, the other player is discriminator 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

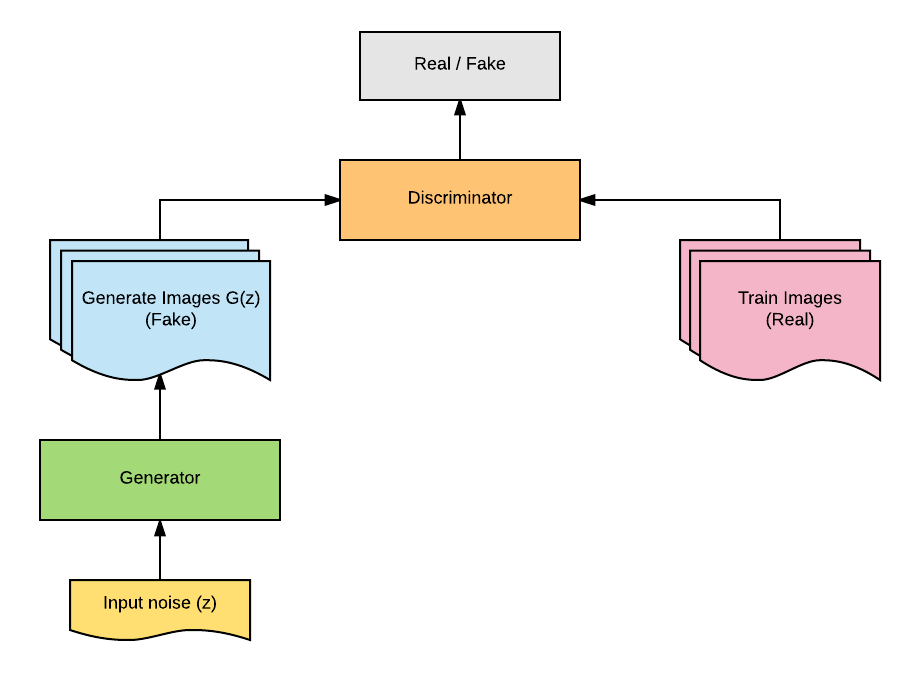


Figure 2.3 shows the the training process of the Generative Adversarial Network