**Project Plan**

Project Title

Research Question

Objectives and Motivations

Summary and Background

References

**Project Title**

Deepfake Face Detection of Human Face Images by Generative Adversarial Network

Detection for Human Face Images by Generative Adversarial Networks

**Research Questions**

**.** How do different evaluation metrics such as precision, recall, and F1-score reflect the performance of deepfake face detection models, and which metrics are most suitable for assessing their effectiveness in real-world scenarios?

**.** How effective are Generative Adversarial Networks (GANs) in generating realistic deepfake human faces compared to traditional methods?

**.** What are the key challenges in accurately detecting deepfake faces generated by GANs amidst rapidly advancing deepfake technology?

**.** What are the reasons behind discriminator loss becoming zero after some iteration?

**Objectives**

Explored methods to distinguish real and fake faces generated by DCGANs.

Used CNN for face detection.

Used confusion matrices and AUROC curves as tools to analyze and visualize the performance of the models.

These tools are proved effective in enhancing GAN-generated face detection.

**Motivations**

Creating tools to protect us from tricky computer stuff that can be used the wrong way.

Making computer models smarter to recognize real faces and not get manipulated.

Helping to keep the internet safer by finding and stopping fake faces using smart tools.

This technological advancement is crucial for safeguarding individuals and institutions against the adverse effects of deepfakes, ensuring the integrity of digital media, and fostering trust in digital communications.

**Background**

What are GANs?

GANs stand for Generative Adversarial Networks:

**Generative**: Capable of generating new data instances.

**Adversarial**: Involving competition or conflict between two entities, in this case, neural networks.

**Network**: A system of interconnected nodes or units, discussion of the neural network architecture used in GANs.

**Background cont…**

In 2014, Ian Goodfellow, along with his colleagues, proposed the concept of Generative Adversarial Networks. The paper titled "Generative Adversarial Nets" was published by Goodfellow and his collaborators. It's like a game between two networks: one, called the generator, makes fake stuff, and the other, called the discriminator, decides if it's real or fake. They both get better by playing this game, with the generator trying to make more realistic things and the discriminator getting better at telling what's real.

**Deepfake:**

The term "deepfake" is a combination of “deep learning” and “fake”. Deepfake refers to a type of media that is created using deep learning techniques, particularly with the use of artificial intelligence models such as Generative Adversarial Networks (GANs). These AI algorithms analyze and learn patterns from large datasets, allowing them to generate highly realistic and convincing fake content, often involving the modification of images, videos, or audio. Fake videos and pictures, called deepfakes, make people worried because they can be used to spread lies, steal someone's identity, and invade privacy. To deal with this, people are working hard to create tools and methods that can find and reduce the harm caused by these fake contents.

**Summary**

Utilize the FFHQ dataset for training and evaluating the models containing real facial images.

Employ Generative Adversarial Networks (GANs) type, namely DCGANs, to generate synthetic facial images.

Choose deep learning model for the identification of deepfake faces that is Convolutional Neural Networks (CNNs).

Calculate accuracy values for each model to measure their performance in distinguishing between real and deepfake images.

Use confusion matrices and Area Under the Receiver Operating Curve (AUROC) to check how well the models are doing. This helps us understand and measure how good the models are at telling real images from fake ones.

**Loss Function**

The basic GANs loss function is:

where:

*D* is the discriminator function.

*G* is the generator function.

x represents real data.

z represents noise sampled from a distribution 𝑝𝑧.

This loss function encourages the generator to produce samples that are indistinguishable from real data according to the discriminator.

**Method for Detection**

Use model (CNN) to find fake images.

Training process undertaken for each model to learn features and patterns from the dataset.

Accuracy values obtained as performance metric after training.

To understand their performance, we used a chart (confusion matrix) that

shows if they were right or wrong.

Another chart (AUROC curve) helped us see how well they could tell fake from real images.

Our research highlights challenges in deepfake techniques and solutions found by face detection algorithms.

**References**

Do, N. T., Na, I. S., & Kim, S. H. (2018). Forensics face detection from GANs using convolutional neural network. *ISITC*, *2018*, 376-379.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega- Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, *64*, 131-148.

Korshunov, P., & Marcel, S. (2018). Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*.

Kohli, A., & Gupta, A. (2021). Detecting deepfake, faceswap and face2face facial forgeries using frequency cnn. *Multimedia Tools and Applications*, *80*, 18461-18478.