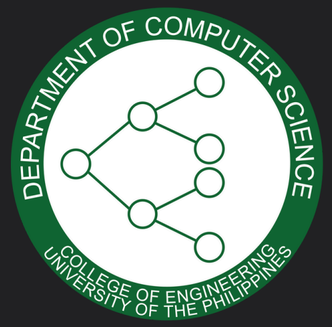


AML

CSG

Investigating the Robustness of Novel Deepfake Detection Transformer Models against Adversarial Attacks

GROUP A1

Photo source: https://www.thetimes.co.uk/article/donald-trump-deepfakes-ai-twitter-g50n7vnbm

What are Deepfakes?

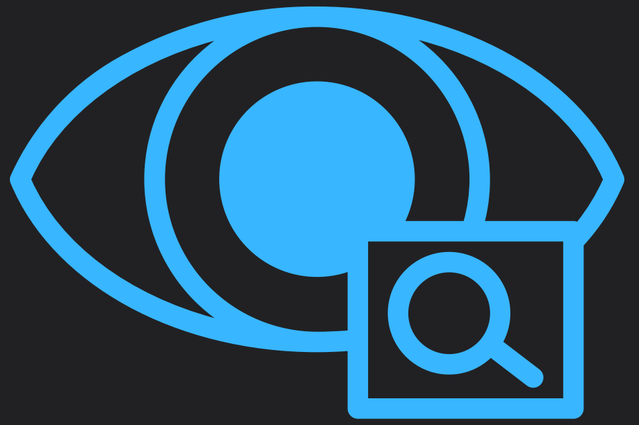
Refers to the term used to describe **synthetic pieces of media** that attempt to **replicate the likenesses** of other people in fictional scenarios

Deepfake

What are Deepfakes?

- **Deepfakes** have become a concerning topic of interest particularly in its use in **visual media** due to its potential for disinformation, online impersonation and harassment, as well as its ramifications in the realm of cybersecurity.

(Westerlund, 2019)



How are deepfakes made?

GANs. Generative Adversarial Networks

(Goodfellow, 2014)

Generator

Generate new media that replicates the original to make it believable.

Discriminator

Reviews the generated media to see if it is indistinguishable from the original piece of media.

Deepfake Detectors

- To combat this, machine learning models trained as **deepfake detectors** were developed to correctly identify the presence of manipulations within the photograph or video.

- A significant portion of the existing implementations for these models utilize Convolutional Neural Networks while **Transformer** models, while not as dominant, have shown **equal to higher accuracy in deepfake detection** comparatively.

(Ahmed et. al., 2022)
(Coccomini et. al., 2022)
(Thing, 2023)

Verifying Robustness Through Adversarial Machine Learning

- Another thing that must be considered when creating these deepfake detection models is their **resilience against malicious inputs**, which is done through the application of **Adversarial Machine Learning**

- Because of their previous dominance in the space, many CNN-based deepfake detectors have already been tested but **Transformer-based** researches are few and far between.

Verifying Robustness Through Adversarial Machine Learning

Research involving applying adversarial attacks on Transformer-based models have been done but only test general image classifiers, not those specifically made for the purpose of deepfake detection.

This paper aims to address this gap in research.

Methodology and Discussion

Published Journal Models

1. Vision Transformers
2. Multi-modal Multi-scale Transformer (M2TR)
3. SeqFakeFormer
4. Identity Consistency Transformer (ICT)

Huggingface Models

1. ViT Deepfake Detection
2. Deepfake Detection Image
3. Detecto: Deepfake Image Detector
4. Deepfake vs. Real Image Detection

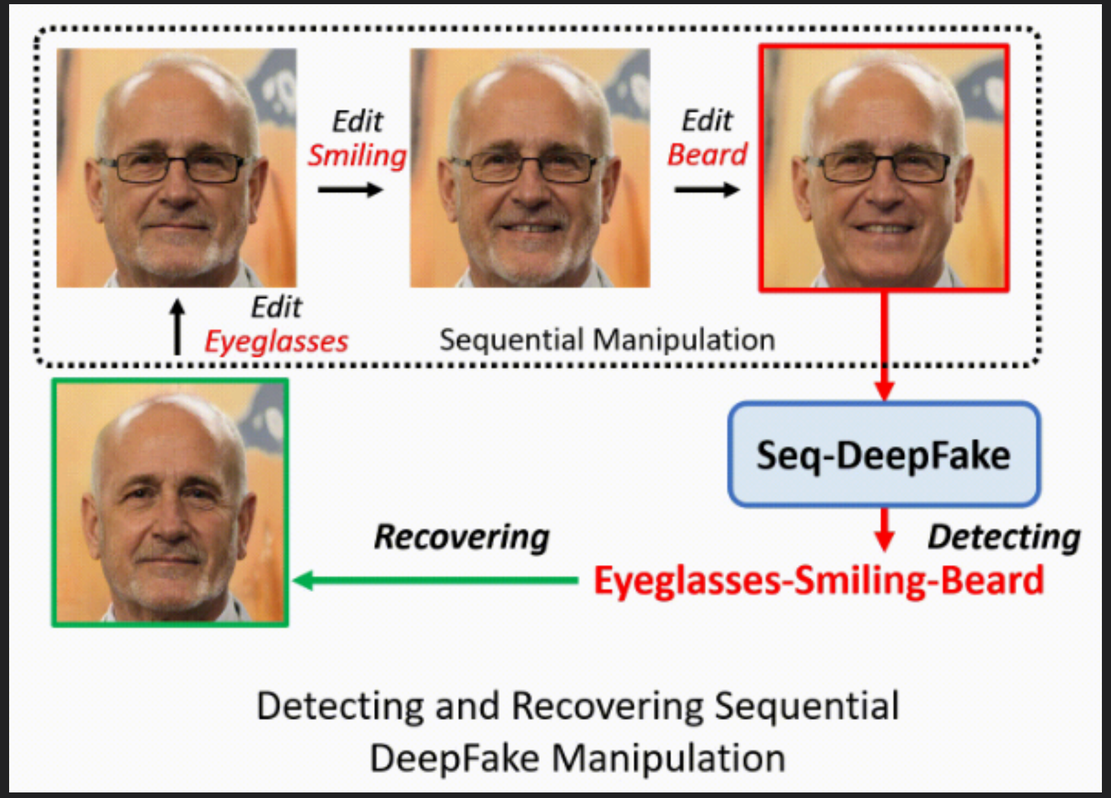
Initial Pool of Transformer Models

Methodology and Discussion

SeqFakeFormer

A Transformer-based model that detects a sequence of manipulations done to create the deepfake as opposed to a binary label indicating if it is real/fake

Stands out as a promising candidate, as it tackles a novel research problem



Detecting and Recovering Sequential DeepFake Manipulation

Shao et. al. (2022)

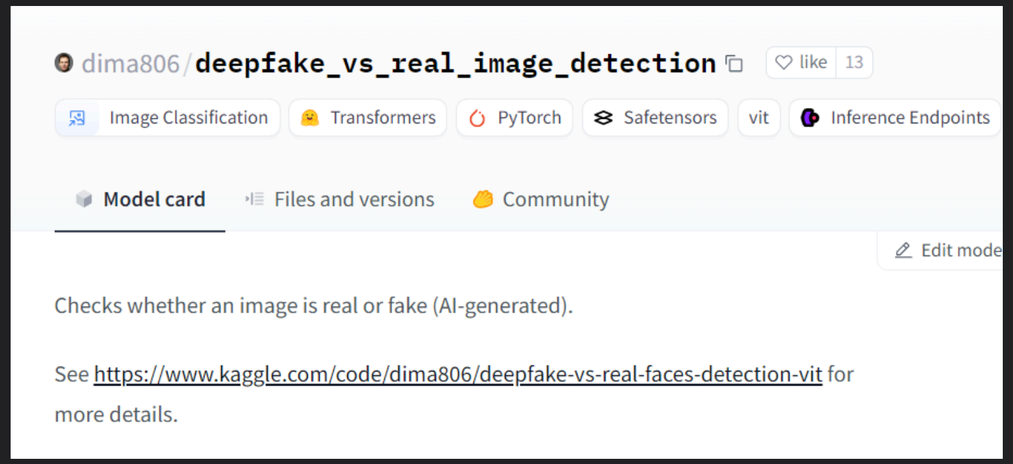
Chosen Models

Methodology and Discussion

Deepfake vs. Real Faces Detection ViT

A ViT-based transformer from huggingface.com made by Dmytro Iakubovskyi

Selected due to its reportedly good performance, well-documented implementation and ease of execution compared to the other huggingface models



Iakubovskyi, D. (2023)

Chosen Models

Methodology and Discussion

Datasets Used

SeqDeepFake

a large-scale dataset created by Shao et. al. comprising of 85,000 manipulated images, annotated with their respective manipulation sequences - generated using 2 methods: sequential facial components and attributes

OpenForensics

dataset obtained from Kaggle sourced from the OpenForensics dataset by Le et al. (2021) containing 115,000 in-the-wild images and 334,000 images of human faces with rich annotations

Methodology and Discussion

SeqFakeFormer Model Testing

- Initial tests involved training the model from scratch but proved to be difficult given the hardware and resources constraints
- Two pre-trained models were used instead, with one trained on facial components and the other with facial attributes
- Problems arose regarding required Python Package which needed to be resolved

Testing of Models

Methodology and Discussion

SeqFakeFormer Model Testing

Facial Components:

Fixed Accuracy Score: 72.657%
Adaptive Accuracy Score: 55.304%

Facial Attributes:

Fixed Accuracy Score: 68.856%
Adaptive Accuracy Score: 49.635%

Testing of Models

Methodology and Discussion

Deepfake vs Real Faces Detection ViT Testing

- Model was utilized through a Jupyter notebook using Python
- Code was slightly modified to include installation of additional required libraries
- Results show slightly lower results than originally reported but are more or less non-significant

Testing of Models

Methodology and Discussion

Deepfake vs Real Faces Detection ViT Testing

class	precision	recall	f1-score	support
Real	0.9920	0.9925	0.9923	38081
Fake	0.9925	0.9920	0.9923	38080
accuracy			0.9923	76161
macro avg	0.9923	0.9923	0.9923	76161
weighted avg	0.9923	0.9923	0.9923	76161

Table 1. Deepfake vs real faces detection ViT Initial Test Results - Classification Report

Testing of Models

Methodology and Discussion

Fast Gradient Sign Method (FGSM)

Popular and efficient

Calini & Wagner L_2 Norm Attack

Slow but effective

Selected Adversarial Attacks

Methodology and Discussion

SeqFakeFormer Initial Testing

- The attempt to generate adversarial examples failed
- Implementations of FGSM and CW-L2 in libraries that were to be used were for CNN models
- Implementations of adversarial attacks for transformer-based multi-label classifiers are hard to find

Application of Adversarial Attacks

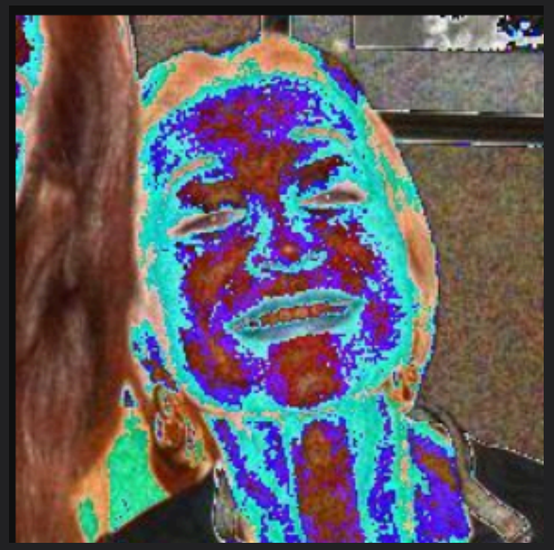
Methodology and Discussion

Deepfake vs Real Faces Detection ViT Testing

- Hugging Face provides an interface for the model originally hosted on Kaggle
- FGSM implementation sourced from CleverHans library
- Custom-made scripts to fit specific model output and obtain a benchmark

Application of Adversarial Attacks

Methodology and Discussion



FGSM Parameters

epsilon = 0.05

norm = 2 (Euclidean distance)

Methodology and Discussion

Null hypothesis:

There is no significant difference on the number of mislabeled images between that of the baseline dataset and the perturbed dataset.

Experiment setup

Baseline (control): Test the model on 50 random samples of 1000 images from OpenForensics

Adversarial: Test the model on 50 random samples of 1000 perturbed images from OpenForensics

Methodology and Discussion

BASELINE

statistic	value
str	f64
"count"	50.0
"null_count"	0.0
"mean"	7.48
"std"	2.822938
"min"	1.0
"25%"	6.0
"50%"	8.0
"75%"	9.0
"max"	13.0

ADVERSARIAL

statistic	value
str	f64
"count"	50.0
"null_count"	0.0
"mean"	229.28
"std"	13.787069
"min"	195.0
"25%"	220.0
"50%"	230.0
"75%"	240.0
"max"	256.0

Measuring the number of mistakes in labeling (fake or real image) across 50 random samples of 1000 images

A significant difference was observed in the performance of the ViT model with regards to correctness of predictions.

The deepfake detector is seen to be very vulnerable against the FGSM L-2 attack!

Recommendations

- Seek domain expertise
- Consider time allotted for conducting preliminary research and for running of heavy scripts
  - Limitations in GPU resources was a recurring obstacle
- Reattempt evaluation of SeqFakeFormer and other transformer models
- Test against other adversarial attack generating methods
- Test against black-box attacks
- Investigate other mediums such as text, video, or audio

Thank You!

Thank you for your attention!

AML GROUP A1

Angelica Raborar

Hans Salazar

ieiaiel Sanceda