Ben-Gurion University of the Negev

Faculty of Engineering Science

Department of Software and Information Systems Engineering

Project in Offensive Artificial Intelligence Course

OAI Final Project - Robustness of Real Time Deepfakes

Author:
Amit Kama

Author:
Oren Shvartzman

June 11, 2022

Contents

1 Introduction

Deepfake is a general term that encompasses the use of deep learning algorithms in order to create synthetic media, in which one subject in an existing visual and/or audio content is usually replaced with another's likeness. While fraudulent content has been around for some time, recent advances in machine vision have posed a major threat to the trust and transparency of the media. Using powerful machine-learning and deep-learning techniques, deepfakes can now manipulate or generate visual and audio content that can be more easily misleading.

In recent years, deepfakes have garnered widespread attention for their uses in spreading fake news, committing financial fraud, creating pornographic materials, and many other disturbing uses. This has led to a significant need to identify and restrict their use.

Ever since the introduction of deepfakes, researchers in deep learning have increasingly focused on this area of research. In particular, they propose methods, as well as practical implementations of deepfakes is in various fields. Among other methods, in [1], Siarohin et al. propose the first order motion model for image animation. Their framework enables generating a video sequence, in which an object in a source image is animated according to the motion of a driving video, without using any annotation or prior information about the specific object to animate. According to them, once trained on a set of videos depicting objects of the same category, the method can be applied to any object of it. Based on this method, a number of real time photorealistic avatars have recently been developed, one of which we will explore in this work.

However, real time avatars are far from perfect, as they are not robust when it comes to edge cases. This includes facial gestures in the driving video, objects in the source or target media that make it difficult to identify facial boundaries, and many more.

Note that these limitations can be create visual glitches and distortions that can be detected with the naked eye as well as by software, and therefore can be utilized for deepfakes detection. This observation drives a growing amount of research dealing with those inaccuracies for dual use – correcting them in order to improve the deepfakes' credibility, or exploiting them to distinguish deepfakes from real content.

In this work, we evaluate the robustness of real time deepfakes by im-

plementing first order motion model-based avatarify and examining various edge cases on it. After demonstrating failures in the implementation, we also provide a method to utilize them for deepfakes detection.

2 Methods

In this section we will briefly introduce the implementation on which our work relies on, and the First Order Model, on which the Avatarify Python consists of.

2.1 First Order Model

2.2 Avatarify Python

3 Experiments and Results

In this section we introduce the experiments and results. A detailed explanation of the visual glitches and distortions that we managed to create and detect is provided, as well as the results of our proposed method to utilize them for deepfakes detection.

3.1 Avatarify Python Robustness

3.2 Failures Detection

4 Discussion

Today, deepfakes are bla bla bla..

References

[1] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *CoRR*, abs/2003.00196, 2020.