Live DeepFake

Yisroel Mirksy, Oleg Brodt, Joshua Cohen, Ruven Levy, Itay Blokh

Abstract—There are an increasing number of live deep fake platforms becoming readily available to the everyday user to easily swap or fake their own and others faces. These everyday users without background in technology or AI have at their fingertips the ability to easily swap and alter their faces. Snapchat for example has live face detection, swapping, and altering capabilities through their filters. FaceApp is another mobile platform that aids in quick face swapping with popular celebrities. Among its direct swapping capabilities, it can also alter the users faces, hair, skin tones and more, all in real time to change the appearance of the target user. Even minor altering such as changing skin tone and hair color, could have major implications when used irresponsibly or maliciously. These attacks are referred to as morphing attacks and identity fraud, and serve as a large threat to Border Protection and Photo ID verification. Next these can also be used in person to person interaction over the internet. One could alter their appearance and give a falso sense of who they are in real time. Following voice cloning and synthesis are becoming just as open and easy to use for the public. Open platforms such as Lyrebird and iSpeech, or even open GitHub repositories allow everyday users to clone voices on their own with little difficulty and repercussions. In this paper, well discuss not only the uses of this technology, but how we have used open source material to create a live synthesized fake of both audio and video in real time. Taking open source GitHub repositories on live deep fakes, and neural net voice cloning, we can create a fully functioning deep fake of both faces and voices to be deployed over a video conferencing platform. Its imperative that the public know that these types of technologies are out in the public and available to use.

Index Terms—Computer Science, Artificial Intelligence, Machine Learning, Generative Adversarial Networks (GAN), Deep Learning, Deep Fake.

1 Introduction

GENERATIVE adversarial networks (GANs) have been applied to many domains, and have shown with great success the ability to generate images [3] and audio samples [4]. GANs have also been used in the public realm from modifying a speech by Speaker of the House, Nancy Pelosi, to swapping Jon Olivers and Jimmy Fallons faces. With the rise in accessibility of GANs for Deep Fakes, its imperative to expose the capabilities of Deep Fakes in doctoring both video and audio in real time.

Previously a large amount of time would be required to fake a video. In a 20-second-long, 25 frames per second video, 500 individual images would need to be edited. This would be incredibly time consuming and difficult task. With the use of Deep Fake, a form of GANs, a similar video could be faked with little to no human intervention after the Deep Fake was synthesized [5].

The same would go for editing an audio sample. In order to edit audio, one would need to manually change the pitch, and tones based on a visualization tool that portrays the wave form of the audio sample. One would do so by cutting, and pasting audio segments, stretching and bending the segments, and recompiling. Even for experts this would be a hard task. The task is even harder when inserting sentences that contain words that dont appear naturally in samples of audio given. With GANs specified for audio synthesis, the insertion of naturally sounding words and short phrases into an audio clip is possible [6] and a relatively easy task.

In the short amount of time that Deep Fakes have been around in the public, stronger and faster algorithms are open to the public to fake audio and video in very little time. This paper introduces a real time Deep Fake that can doctor both a persons face and voice.

- *Yisroel Mirsky is with Telekom Innovation Laboratories, Ben Gurion University, Be'er Sheva, Israel*
  E-mail: ymirsky1@gmail.com
- *Itay Blokh is with Ben Gurion University.*

Manuscript received August 5th, 2019

2 Background

GENERATIVE Adversarial Networks known as a GANs is a system of two separate neural networks working against each other. The system is comprised of a generative model, and a discriminative model.

A generative model is the probability that an observation X is true based on a given target y [9]. A discriminative model is the probability that the target y is true given the observation of X [12]. GANs is a framework that estimates the generator by training the generator and discriminator simultaneously. [11] Take for example a forger and a police officer. The job of the forger is to create a check that is realistic enough to not be caught by a police officer. The job of the police officer is to determine whether the check is real or fake. The police officer is given a data set of checks that are labeled real or fake. The police officer learns the characteristics of real and fake checks in the effort to predict the probability that a new check, the target, is real or fake, given his observed checks. The forger begins handing the officer checks asking if they are real or fake. The officer either accepts them as real or rejects them as fake. The forger learns from this distribution characteristics that will pass as real to the officer. Simultaneously the officer is updating his own understanding of real and fake checks. The goal of the forger is to predict whether a check that he hands to an officer is true before handing it to the officer. As the forger learns which checks pass the test and which don't, he's able

to generate more realistic checks to minimize the probability that a check given to officer will fail the test.

2.1 What are DeepFakes?

DEEPFAKES are Deep Learning algorithms intended for the purpose of superimposing previously existing images or videos onto another image or video. In the recent years, there has been little headway in DeepFakes for live camera feeds. A previously existing GitHub project, called "FaceIt_Live" by Alessandro Cauduro [1], is one of the few existing projects that allows for a real-time DeepFake. In this project, we used a modified version of FaceIt_Live in order to create realistic DeepFakes in real time. For the audio portion of this project, we are using a modified version of Real-Time Voice Cloning by Corentin Jemine [8], which allows us to model a person's voice with only a few seconds of audio. By using Real-Time Voice Cloning, we were able to easily mimic a person's voice. By combining these two individual projects, modifying them for our needs, and using them, we were able to create a live Deep Fake.

2.2 Image to Image translation using Generative Adversarial Networks

IMAGE to image translation is the process by which the structure or semantics of one image is moved to the domain of another image. A form of image to image translation is face swapping. Take for example the face of a person, person A, in one image that is moved onto the face of another person, person B. In image to image translation the semantics of person A, that is the characteristics that define them, are transferred using the domain of person B, that is the facial expression. This allows a human to noticeably distinguish that the new image is of person A while maintaining the mood and expression of person B [2].

2.3 Voice Cloning using GANs

VOICE cloning using a Generative Adversarial Network has two main focuses copying the text and translating the voice and emotion. This is done through a multi-step process using GANs. First, the words are captured using a speech to text translator. This could be one of many translators such as Googles Speech to Text, or others like it. There are then two approaches to cloning a speaker. A multi-speaker generative model is created to be adaptable to different speech patterns, known as a Speaker Adaptation. This is then fine-tuned to model to different speakers with naturally different inflections. Secondly a speaker embedding of the speaker whose voice is to be cloned is created using a similar GAN. The speaker embedding will then be applied to a similar multi-speaker generative model. This creates a cloned model for the new speaker. While the approaches have similarities, speaker embedding (creating model for the speaker first) is shown to be faster, while speaker adaptation (fine-tuning a broad model) proves a more natural sound from the cloned voice B [7].

3 RELATED WORK

WHEN it comes to deep faking a video sequence of a person speaking, there are a few key features that must be taken into account in order to create a video realistic enough to fool an audience. These features include, facial movement and expression, synthesis of both speech and mouth movement, and a realistic voice cloning which carries emotion. A live deep fake, must include these in real time. These key features are demonstrated separately in other works in the area of Deep Fakes.

Audio and video generation have been explored in multiple ways pertaining both to a combined audio video fake, and to separate fakes, consisting of audio from video, and video from audio. The Talking Face Model, is an example of a face image generator, which takes as input an audio sequence, normally a voice that has been trained on, and generates a sequence of facial images of the target which the model has also been trained on. Furthermore the goal of The Talking Face Model was to take a single image of the target and use an audio sequence create those facial images of the target.

4 IMPLEMENTATION

4.1 Video

OUR goal is take an existing live deep fake for faces, train it on our subject and our target, and implement it in real time test the effectiveness of the fake. Using"FaceIt Live", and existing GitHub repository by Alessandro Cauduro [1] we were able to implement a deep fake. FaceIt Live meshes together the image of the subject to the target to a face of the target using the movements of the subject. The FaceIt live code is based on FaceSwap, which recognizes faces in images and swaps them the target face. Following the easy to use instructions on the GitHub Repository, we were able to implement the real time Deepfake Face swap over a live video feed.

4.2 Audio

VOICE cloning is another popular deep-fake study area, especially when referring to voice cloning, in the context of real time. This begs the question, what does it mean to fake a voice in real time? The delay time to create and synthesize a sentence from scratch, should this take 5 seconds, or closer to .5 seconds? Where is the input sentence coming from if a conversation is taking place? These were the major considerations we took when analysing different voice cloning techniques, and when we finally chose and adapted a voice cloning program to fit the specs of our project. We chose to use the "Real Time Voice Cloning" Repository on GitHub, by Corentin Jemine.

Following the instructions on the repository, a real time voice cloning sampler was constructed. In order to clone a voice, a voice sample of at least 10 seconds was required to imported once. Then every time a new sentence or phrase needed to be created that phrase would need to be vocoded. The vocoding process took a constant time of 7 seconds, regardless of the length of the sentence. Sentences with fewer words would be stretch, and or unecessary pauses would be made in the middle. This was due to the fact the program was created a spectogram for a constant length of time.

The first step in advancing the real time aspect of this voice cloner was to speed up to rate at which vocoder creates the wave file. This was done by skipping every other iteration through the generation loop. The audio was output two times faster. This in turn sped up the rate at which the clone spoke, and so the output was slowed down to half speed in order to maintain that real human sound and speed.

4.3 Synthesized Fake

WITH a voice cloner and live face swap programs in place, the next step would be to combine the two to create a live deep fake attack to be tested. The attack would be over a live video feed with a single person, or group of people on a video conference call.

The video and audio fakes remain separate and such two attackers were needed to create a realistic synthesized deep fake. Attacker one would sit in front of the camera, he or she would have their face trained for the face swap prior to the test. They would act as the controller of the head model that projects over the video call. Attacker two sits next to attacker one out of the scope of the camera. Attacker two types sentences into the UI of the real time voice cloner to be vocoded by the program and project over the audio of the video conference. Attacker one must be ready to read in time with the vocoder to appear to be speaking on his own.

APPENDIX A
PROOF OF THE FIRST ZONKLAR EQUATION

Appendix one text goes here.

APPENDIX B

Appendix two text goes here.

ACKNOWLEDGMENTS

The authors would like to thank...

REFERENCES

[1] A. Cauduro, Face_It Live (2018), GitHub Repository, https://github.com/alew3/faceit_live
[2] H. Dong, P. Neekhara, C. Wu, Y. Guo, Unsupervised Image-to-Image Translation with Generative Adversarial Networks, 2017
[3] P. Isola, J.Y. Zhu, T. Zhou, A.A. Efros; The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1125-1134
[4] C. Donahue, J. McAuley, M. Puckette, Adversarial Audio Synthesis, ICLR 2019.
[5] Li, Yuezun, et al. In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. ArXiv.org, Cornell University, 11 June 2018, https://arxiv.org/abs/1806.02877.
[6] Zeyu Jin, Gautham J. Mysore, Stephen DiVerdi, Jingwan Lu, and Adam Finkelstein. VoCo: Text-based Insertion and Replacement in Audio Narration. ACM Transactions on Graphics (July 2017).
[7] Arik, Sercan O., et al. Neural Voice Cloning with a Few Samples. ArXiv.org, Cornell University, 12 Oct. 2018, https://arxiv.org/abs/1802.06006.
[8] C. Jemine, Real-Time Voice Cloning (2019), GitHub Repository, texttthttps://github.com/CorentinJ/Real-Time-Voice-Cloning
[9] T. Jebara, Machine learning: Discriminative and Generative. Boston: Kluwer Academic, 2004.
[10] T. Jebara. Discriminative, Generative, and Imitative Learning (2005).
[11] S. Vaerenbergh, I. Goodfellow. Ian Goodfellow: Generative Adversarial Networks (NIPS 2016 Tutorial). 18 Jan. 2018, www.youtube.com/watch?v=HGYYEUSm-0Q&feature=youtu.be.
[12] Generative Model. Wikipedia, Wikimedia Foundation, 31 May 2019, https://en.wikipedia.org/wiki/Generative_model.