Deepfake: An Deep Learning Approach to Replace Human Face in Videos(July 2019)

500ping

***Abstract*—Deepfake is controversial, the advent of this technique brings both praise and criticism. From moral perspective, it does pose detrimental effects on personal privacy, whereas from angle of computer vision, deepfake proposes a new method of image fusion and enhancement, which is likely to be utilized in industries of film and journalism. I am going to analyze the history and current situation of deepfake technique then make use of a deepfake software named DeepFaceLab to deploy two groups of experiment. A conclusion will be conducted from experimental results.**

***Index Terms*— Deepfake, DeepFaceLab**

I. INTRODUCTION

Deepfake (a portmanteau of "deep learning" and "fake"[1]) is a technique for human image synthesis based on artificial intelligence. It is used to combine and superimpose existing images and videos onto source images or videos using a machine learning technique known as generative adversarial network[2].

A. History of Deepfake

The phrase "deepfake" was coined in 2017, however, its story can be traced back to 1997, an early landmark project Video Rewrite program was published, which modified existing video footage of a person speaking to depict that person mouthing the words contained in a different audio track[3]. It was the first system to fully automate this kind of facial reanimation. Since then, related academic research results appear on after anther.

The Face2Face program, published in 2016, modifies video footage of a person's face to depict them mimicking the facial expressions of another person in real time.[4]The "Synthesizing Obama" program, published in 2017, modifies video footage of former president Barack Obama to depict him mouthing the words contained in a separate audio track[5]. In January 2018, a desktop application called FakeApp was launched. The app allows users to easily create and share videos with faces swapped. It uses an artificial neural network and the power of the graphics processor with three to four gigabytes of storage space to generate the fake video. In June 2018, author iperov opened source his codes of DeepFaceLab on GitHub. From that day on, people had easy access to this mysterious technology and plenty of similar software appeared continually.

Highly encapsulated software reduces user learning costs, everyone is able to create an eye-catching fake video with oven common laptop. There are three mainstream deepfake software.

1) DeepFaceLab

DeepFaceLab is the fastest updated software among theses three software. As mentioned above, it is an open source project from GitHub, everyone have free access to the source code. The greatest advantage of this project is that the installation process is so simple that no installation is required. At the same time, the operation method is not complicated. The only regret for novices is that there is no visual interface. There are five models contained in DeepFaceLab, each of them is designed for specific experimental environment. Therefore, DeepFaceLab is the most flexible software among competitors.

2) Fakeapp

Fakeapp is the most widely spread and simplest one in all face-changing software, but it has not been updated for a long time. Compared with its successors, Fakeapp is high encapsulated which enhances operability, whereas default parameters and model are not adjustable. Therefore, it cannot be

applied to every scene and the model performance is not satisfactory.

3) OpenFaceSwap

Openfaceswap is a graphical interface version customized based on the open source software Faceswap. Faceswap is the most attention-oriented face-changing project on the original website GitHub, but since there is no visual interface in the early days, the installation is very complicated. Later, even if it launched its own GUI, the use experience is not good. The emergence of Openfaceswap is a good choice for trainee players. Besides, OpenFaceSwap supports modifying parameters and models.

B. Impact of Deepfake

The advent of new technology is not always exhilarated. The abuse of synthetic fake videos destructively undermined the citizen's portrait rights and personal privacy. The chief culprit was a Reddit user named "deepfakes". He, as well as others in the Reddit community r/deepfake, shared deepfakes they created; many videos involved celebrities' faces swapped onto the bodies of actresses in pornographic videos[6], while non-pornographic content included many videos with actor Nicolas Cage's face swapped into various movies. More victims were involved since 2018. Deepfake have been used to misrepresent well-known politicians on video portals or chartooms, including Argentine President Mauricio Macri [7], speaker of the United States House of Representatives Nancy Pelosi and even former president Barack Obama.

Basically, people tend to use face of others in creating deepfake videos instead of using their own face, and this is why we can always find an icon's face on anther head in an unrelated video. Obviously, they use portrait rights of others without permission in most case, what's worse, some deepfake videos aiming celebrities tend to be more malicious and shameless.

From aspect of Aargauer Zeitung. It is not a technical problem, but rather one to be solved by trust in information and journalism. The primary pitfall is that humanity could fall into an age in which it can no longer be determined whether a medium's content corresponds to the truth.

C. Research Motivation

Through the detrimental impact of deepfake it brings, research of deepfake is still of great importance. Viewed from one perspective, fortunately, deepfake video can be detected. A new digital forensics technique promises to protect actors, world leaders, and celebrities against such deepfake, The new method uses machine learning to analyze a specific individual' s style of speech and movement, what the researchers call a "softbiometric signature.[8]" The researchers, from UC Berkeley and the University of South California, had proved the detecting accuracy of deepfake is greater than 90 percent. In this case, studying the nature of deepfake is able to boost the efficiency of fake video detection.

From anther perspective, deepfake really works in the field of journalism and movie. It is well-known that AI journalist has produced a lot of news. For example, At the Rio Olympics in 2017, the AI robot "xiaomingbot", which was developed by ByteDace laboratory in China, wrote the news manuscript in real time through the database information of the Olympic Organizing Committee[9]. Imagine the future with AI robot and deepfake together, they can probably take over job of news anchors. In film industry, deepfake is capable to help produce special effects as the traditional method is always time consuming and extremely expensive.

To sum up, it is still meaningful to analyze the principle of deepfake. In the later experiment, DeepFaceLab is chosen to produce a demo video, and source code as well as detailed experimental steps will be also covered.

II. Methodology

Entire process of deepfake consists of three steps, one is to extract data, the other is training, and the third is conversion. The first and third steps need to use data preprocessing techniques, and the third step also uses image fusion. So we will illustrate it from three aspects: image preprocessing, network model, and image fusion.

A. Image Preprocessing

Entire process of deepfake consists of three steps, one is to extract data, the other is training, and the third is conversion. The first and third steps need to use data preprocessing techniques, and the third step also uses image fusion. So we will illustrate it from three aspects: image preprocessing, network model, and image fusion.

B. Network Model

The overall network structure is still in the form of encoder-decoder, but unlike the autoencoder, it consists of one Encoder and two Decoders, and the two Decoders correspond to the decoding of source image and target image respectively.

```
def Decoder():
    if not lighter_ae:
        input_ = Input(shape=(16, 16, 512))
        x = input_
        x = upscale(512)(x)
        x = upscale(256)(x)
        x = upscale(128)(x)

        y = input_  #mask decoder
        y = upscale(512)(y)
        y = upscale(256)(y)
        y = upscale(128)(y)
    else:
        input_ = Input(shape=(16, 16, 256))
        x = input_
        x = upscale(256)(x)
        x = upscale(128)(x)
        x = upscale(64)(x)

        y = input_  #mask decoder
        y = upscale(256)(y)
        y = upscale(128)(y)
        y = upscale(64)(y)

    x = Conv2D(3, kernel_size=5, padding='same', activation='sigmoid')(x)
    y = Conv2D(1, kernel_size=5, padding='same', activation='sigmoid')(y)

    return Model(input_, [x,y])
```

```
def Encoder(input_shape):
    input_layer = Input(input_shape)
    x = input_layer
    if not lighter_ae:
        x = downscale(128)(x)
        x = downscale(256)(x)
        x = downscale(512)(x)
        x = downscale(1024)(x)
        x = Dense(512)(Flatten()(x))
        x = Dense(8 * 8 * 512)(x)
        x = Reshape((8, 8, 512))(x)
        x = upscale(512)(x)
    else:
        x = downscale(128)(x)
        x = downscale(256)(x)
        x = downscale(512)(x)
        x = downscale(1024)(x)
        x = Dense(256)(Flatten()(x))
        x = Dense(8 * 8 * 256)(x)
        x = Reshape((8, 8, 256))(x)
        x = upscale(256)(x)

    return Model(input_layer, x)
```

Figure 2-1 Source code of Decoder and Encoder [10]

C. Image Fusion

Since the generated face picture is a square, how to merge with the original image is a problem. The original project has many fusion methods, including direct coverage, mask coverage, and Poisson clone "Seamless cloning". The effect of mask covering is the best with Poisson clones. The two have their own advantages. The cover covers the edges and the blush is soft. The single-picture effect is better than the mask cover, but the Poisson clone will make the picture A slight shift occurs, so there is some jitter in the video synthesis. The idea of the improvement of the picture fusion problem, the author believes that it is still necessary to start from the generation of the picture itself, the introduction of the mask in the second project, is a very good idea, but also a thought that is easier to think of, is the final generation of a RAGB with a clear Degree of pictures. I have tried many methods, the better of which is to add very small self-recovering L1Loss while introducing Gan, so that the picture is "harmony and different". After testing, this method can make the edge of the picture and the original picture basically merge, but this method also has drawbacks, that is, the relatively large changes like the face type, the network is not willing to try, the network tends to minor repair Small supplements, only change the characteristics of the five senses[11].

III. EXPERIMENT AND RESULTS

According to document of DeepFaceLab, the recommended system requirements are Window 7 and higher, processor with support fort AVX instructions, more than 8GB RAM and most importantly, the GPU processor should be NVDIA with 6GB video memory. As for software support, Visual Studio 2015, CuDNN 7.0.5, CUDA 9.0 and source file of DeepFaceLab are essential. In practical experiment, I have only a NVDIA 1050Ti graphics card with 4GB video memory, which is capable for one or two models but needs more time for training. Besides, CPU can be substitute for GPU in model training process whereas the training efficiency is disappointing.

A. Dataset

There are two necessary videos, one is source video and another is target video. The former is required to contain only one person whose face is supposed to replace faces in the target video. Specifically, we use Robert Downey jr. 's face to replace Shia La Beouf's face.

· Source Video

Facial information is a key to a perfect performance of deepfake video. It is supposed to contain facial expressions under a proper light condition as much as possible. Length of the source video depends on the frame rate of video. As normal video has only 30 frame rate, a 50 seconds video is qualified because 1,500 frames of human face are adequate for present model. Specifically, American actor Robert Downey jr. 's video is used in this part.

· Target Video

There is no strict restrictions for target video, however, if an Asian face is extracted to cover the face of an European, it is likely to produce some obvious defects in the video as their distribution of facial features varies greatly and present model is not capable to solve the problem. For faces with different color, specific algorithms are used to blur the boundaries of different patches but the product is not always satisfactory. In this experiment, a video of American actor Shia La Beouf is used as target video.

Figure 3-1 Target face of Shia La Beoufand and source face of Robert Downey jr.

B. Experimental Protocols

Although DeepFaceLab does not have a visual interface, it is very clear that the whole process is divided into eight steps. Each step can be executed by simply clicking the BAT file. The BAT file is a batch file under dos. A batch file is an unformatted text file that contains one or more commands. Its file extension is .bat or .cmd[12].

1) Clear Workspace: It is used to reset the workspace (delete all model files, extracted pictures, and videos). There is no need to clear workspace for the first time using.

2) Extract PNG from Source Video: Extract each picture from the source video frame by frame.

3) Extract PNG from Target Video: Extract each picture from the target video frame by frame.

4) Extract Faces from Source PNG: Use python package dlib to detect and extract human face pictures from source pictures, the pixel of each face picture is 128*128.

5) Extract Faces from Target PNG: Use python package dlib to detect and extract human face pictures from target pictures, the pixel of each face picture is 128*128.

6) *Train Model H128*: H128 is a model for device with 4GB video memory, it costs really long time. In my experiment, I spent totally 48 hours training the model, training record will be illustrated later in chapter 3.3.

7) *Convert Faces:* Use the trained H128 model to change the face of each picture from target video.

8) Convert PNG to mp4 Format: Convert pictures that has been changed faces into a video.

C. Experimental Results

The same source video and target video are tested with different experimental configurations. The performances diverse obviously, some of them are fake to the naked eye while some cannot be distinguished at a glance.

1) Model Analysis

There are two models involved, H64 and H128. H64 indicates that the pixel of extracted human face picture is 64*64 while H128 means a pixel of 128*128. Therefore, the face got from H64 is supposed to be more blurred. Another difference is that the H64 is designed for device with 2GB video memory while the other is for a higher video memory of 4GB. In order to ensure the preciseness and reliability of the experiment, two groups of experiment are set to default parameter and trained with same time (90 minutes) on the same device.

It can be seen from Figure 3-1 and Figure 3-2, the output of H64 model after 90 minutes of training is disappointing, it looks like a picture is attached to a face. The boundary between two different faces is clear and distinct, and the skin color difference between the two faces is obvious.

By contrast, the output of H128 model looks better, it is basically a complete human face. However, if you zoom in on this picture, the central area of the face is more blurred. Compared with the original picture, their facial expressions are also different, eyes watching different direction and mouth opening different wideness.

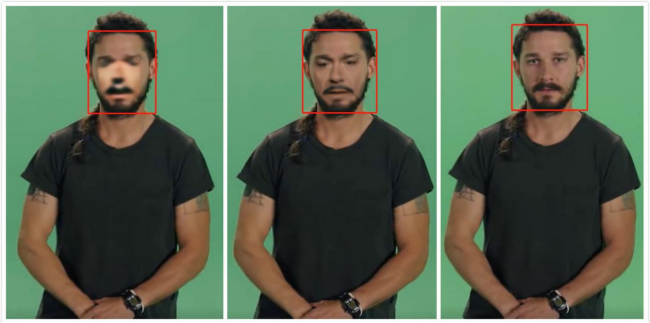

Figure 3-2 Output of H64, H128 and original human face

2) Training Time Analysis

Training time is an essential considerations in terms of deep learning process. In general, more training time brings higher validation accuracy and lower loss. So there are three groups of experiment, the first is trained for 90 minutes, the second group is trained for 7 hours and the last group is trained for 24

hours. The performance can be evaluated by both loss function and visual observation.

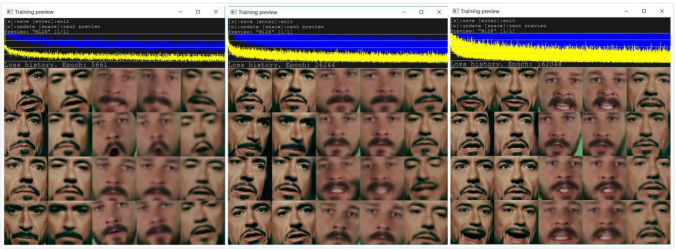

Figure 3-3 Epoch 5661, 26244 and 162408

Figure 3-3 Output after training for 90 minutes, 7hours and 24hours. They respectively correspond to training epoch of 5661, 26244 and 162408.

With the growth of training time, the face is getting increasingly clear and vivid. After 7 hours of training, the sharpness of picture is greatly improved training but details are still not perfect (eyes are blurred). However, 24 hours later, the facial muscles and shadows of the character are more realistic. Thanks to the improvement of the fineness of the picture, even the texture details of the skin can be seen, and the boundaries between the deaf and the white part of eyes become more distinct, therefore the character's demeanor become more lifelike.

Increase of training time does works to improve video quality to some extent, but there is still a upper limit of model. In another experiment, I continued training for 48 hours, the loss keeps in a constant area and the clarity of the character in the preview box did not change much. To sum up, after the training time exceeds a certain threshold, the output will not change significantly. The proper training time is different under multiple experimental configuration, the choice of perfect timing really depends on personal experience.

Nevertheless, there are other problems that can be only seen from the video. The character's expression changes rigidly. This should not be blamed for video frames because the processed video has the same frame rate as the original video. Basically this is mainly caused by the weakness of the model, because changing faces is not equal to attaching pictures. The facial expressions of the characters correspond to the movement of the muscles. Only by precisely positioning the facial muscles and making appropriate changes can the character expressions be more natural.

In additional to this, other conspicuous defects still exists, as it can be seen from Figure 3-11, the matching of mouth is totally wrong and irregular blurring occurs at the edges when the face is blocked by an object. Besides, strange color patches appear on the face of the character sometimes.

Figure 3-3 Training 90 minutes, 7hours and 24hours

IV. Conclusion

To sum up, the H128 model performs better than H64 under a 4 GB NVIDIA 1050Ti graphic card, and the quality of training output will improve with growth of time in a short period of time. If we take out a few static frames of the video, it does look really realistic. However, even after a long period of training, there are still defects in deepfake video which can be distinguished by naked eyes.

After the experiment, we have reason to believe that most of deepfake videos can be distinguished by naked eyes. Generally these videos cannot be too long because it really takes time for transformation. By contrast, deepfake pictures seems more detrimental as it is easier to produce and looks quite realistic.

References

[1] Brandon, John (2018-02-16). "Terrifying high-tech porn: Creepy 'deepfake' videos are on the rise". Fox News. Retrieved 2018-02-20.

[2] Schwartz, Oscar (12 November 2018). "You thought fake news was bad? Deep fakes are where truth goes to die". The Guardian. Retrieved 14 November 2018.

[3] Bregler, Christoph; Covell, Michele; Slaney, Malcolm (1997). "Video Rewrite: Driving Visual Speech with Audio". Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques. 24: 353–360 – via ACM Digital Library.

[4] Thies, Justus; Zollhöfer, Michael; Stamminger, Marc; Theobalt, Christian; Nießner, Matthias

(June 2016). "Face2Face: Real-Time Face Capture and Reenactment of RGB Videos". 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE:2387–2395. doi:10 .1109/CVPR.2016.262. ISBN 9781467388511.

[5] Thies, Justus; Zollhöfer, Michael; Stamminger, Marc; Theobalt, Christian; Nießner, Matthias (June 2016). "Face2Face: Real-Time Face Capture and Reenactment of RGB Videos". 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE: 2387–2395. doi:10.1109/CVPR.2016.262. ISBN 9781467388511.

[6] Cole, Samantha (24 January 2018). "We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now". Vice. Retrieved 4 May 2019.

[7] "Wenn Merkel plötzlich Trumps Gesicht trägt: die gefährliche Manipulation von Bildern und Videos". az Aargauer Zeitung. 2018-02-03.

[8] A new deepfake detection tool should keep world leaders safe—for now. (Jun 21, 2019). https://www.technologyreview.com/s/613846/a-new-deepfake-detection-tool-should-keep-world-leaders-safefor-now/.

[9] 新闻写作机器人的应用及前景展望——以今日头条新闻机器人张小明（xiaomingbot）为例. 赵禹桥 (2017-01-11).http://media.people.com.cn/n1/2017/0111/c409691-29014245.html.

[10] DeepFaceLab. iperov.(2018-07).https://github.com/iperov/DeepFaceLab/blob/master/models/Model_H128/Model.py.

[11]DeepFake.fenneishi.(2019-01-22).https://blog.csdn.net/weixin_36673043/article/details/86593495#_deepfake_158.

[12] BAT 文件.https://baike.baidu.com/item/bat%E6%96%87%E4%BB%B6/5457821?fr=aladdin.