Deepfakes: Introduction and Latest Implementations

Angel Igareta

April 29, 2020

Abstract

Deep Learning algorithms are being used in several fields nowadays, such as Self-driving cars, Healthcare or Voice-Activated Assistants. Programs developed with these algorithms are getting easier for not experienced users to use and their features are constantly growing. Despite this being positive for contributing to previous sectors, deep learning-powered applications that could cause threats in politics, privacy and security are being developed too. An example of this is the so-called "deep fakes". This paper has the goal of introducing this type of algorithms, showing state-of-the-art implementations in the area and examples of what this technology is capable of.

1 Introduction

Deepfake is a deep learning technique used to create convincing image, audio and video hoaxes. The most common application is to swap a person in an existing image or video with another one, generating a fake synthetic media. The difference with traditional techniques such as Photoshop is that deepfakes consists of powerful techniques for generating visual or audio content with high potential to deceive. Some of the popular areas where these algorithms are used are fake news, financial fraud, revenge porn and celebrity pornographic videos. As these are critical areas, most of the academic research surrounding Deepfake seeks to detect this content [10]. Later some techniques for deepfake detection will be presented. According to the legal response, some countries have been introducing legislation to tackle deepfake content. For instance, in the United States, some states such as California, Texas and New York have introduced *"The Malicious Deep Fake Prohibition Act"*, which will make a federal crime to create or distribute a deepfake when doing so will facilitate illegal conduct [4].

The technology used by deepfakes algorithms includes previous techniques such as autoencoders and generative adversarial networks. A drawback in these algorithms would be the high quantity of data needed to train models to create realistic synthetic content. For this reason, most of the initial targets for the deepfakes are public celebrities or politicians, which have a large number of videos and images available online. However, these techniques are improving significantly over time and there are some techniques available to generate deepfake content with few images, through an approach called *few shot learning*.

The first deepfake video was produced in 2017 where a popular face was swapped in pornographic content and published on Reddit. In 2018, the famous website banned this type of content classified

as *fake porn*, which was followed by Twitter or Gfycat [6]. On top of that, these algorithms can be used to threaten the national security of an entire country, for instance publishing a video of the president of the United States giving a fake speech [14], which could affect election campaigns or military actions. It could be used as part of social engineering scams through the branch of audio deepfakes, fooling people into thinking they are receiving instructions from a trusted source. In 2019, a UK energy firm CEO was scammed $\$243,000$ when he was ordered by a supposed chief executive from the parent's company to transfer the money to a Hungarian bank [5].

Despite all these dark applications of deepfake, there are many positive ones such as creating voices for those who have lost theirs, recording the same character in a film through different ages, such as in the Netflix film *'The Irishman'* [16] or updating episodes in movies without shooting them again [11].

The first section of the document has the aim of informing about the basic, as well as the state-of-the-art technologies used to generate deepfakes. In the second section, there will be presented some technologies that are being used to detect these fake content and how successful they are. Finally, there would be a deepfake demo created for showing deepfake's potential.

2 Deepfake Generation

Since the creation of the first deepfake, the algorithms used to create them have been improved significantly, at the same time as the algorithms used to detect them. This competition between the two sides has resulted in very robust deep learning techniques to generate fake content that is quite arduous to spot.

The first attempt of deepfake creation was FakeApp, developed using an autoencoder-decoder pairing structure. **Autoencoder** is an unsupervised artificial neural network that learns first how to efficiently compress data down to a selected size (encode step) and then reconstruct the data from the compressed representation to one that is as close as the input as possible (decode step). In order to measure how well the decoder is performing and how close the output is to the original input, the reconstruction loss is calculated, so the training involves using backpropagation in order to minimize this loss. Some of the most popular uses for these technologies are anomaly detection and image denoising [3].

Furthermore, as FakeApp proved by allowing to swap faces of two individuals, these neural networks can also be used to generate deepfakes. In this application, the autoencoder intention is to extract facial features from an input face (as the algorithm compresses, only the essential features of the face will remain). Then, the decoder takes those features as input and reconstruct them to produce the original input face. As the goal is to swap two faces, it does not only use a single autoencoder-decoder structure but a pair. The first one will learn the facial expressions of the subject A and make use of the compressed features to reconstruct the subject A, the second one will perform likewise with subject B. On top of that, both structures possess the same encoder network. Subsequently, in order to swap faces, it only requires to swap the decoders, as they have learnt how to reconstruct the subject they were trained with. As faces normally have similar features such as eyes, moth positions, eyebrows, the task is simple. In figure 1, an example of these approaches is displayed, it is used to copy the facial expression of subject A into subject B.

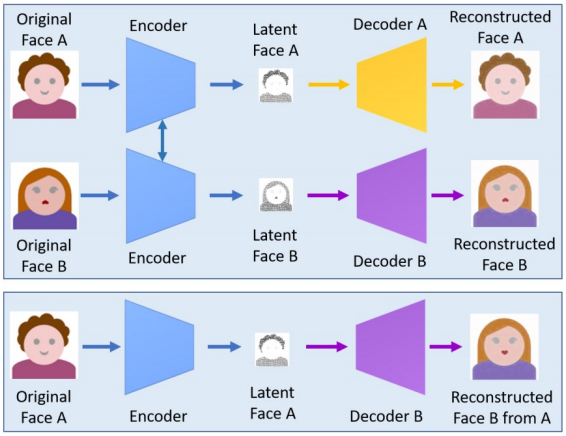

Figure 1: Example of autoencoding-decoder pairs structure to create a deepfake [12]

A major improvement in the deepfake development, as well as in the overall Artificial Intelligence advances were the generative adversarial networks (GANs), related to the field Adversarial Learning. These are algorithmic architectures constituted by two neural networks, competing one against the other, these two are called the generator and the discriminator. The generator is a neural network which aim is to produce new data instances, while the discriminator evaluates them to see if they are real or fake, in other words, it determines if the received image belongs to an actual training set. If we apply this concept to deepfakes, our generator will be generating images that mimic the target object, for instance, a human face. In the beginning, the generator will start by generating random noise, as it does not have any information about how a human face really is. That feedback is given by the discriminator, through its loss. Meanwhile, the discriminator is being trained with real faces from a training dataset and fake images previously generated by the generator. Hence, the generator progressively becomes better at creating human faces that look real, while the discriminator becomes better at spotting which ones are fake. Our stop condition would be when the discriminator can no longer distinguish real faces from fakes. An example is displayed in figure 2.

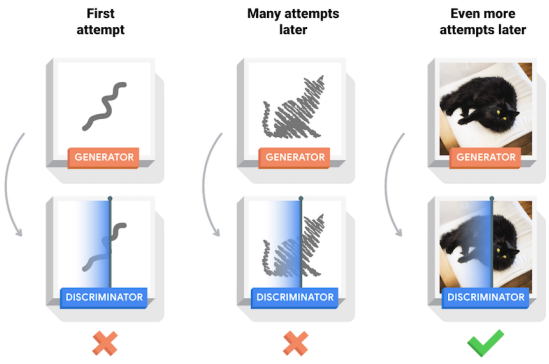

Figure 2: Generative Neural Network example trained with images of cats [1]

The addition of adversarial loss and GANs have significantly improved deepfakes creation. However, the requirement of a huge amount of input data has been a big limitation for their creation. Recently, some studies have been working on generating a video sequence of a person by only having a single or few images of the target person. Despite not obtaining the same result, it produces quite good results. This is a new branch in Artificial Intelligence called 'Few-shot learning'.

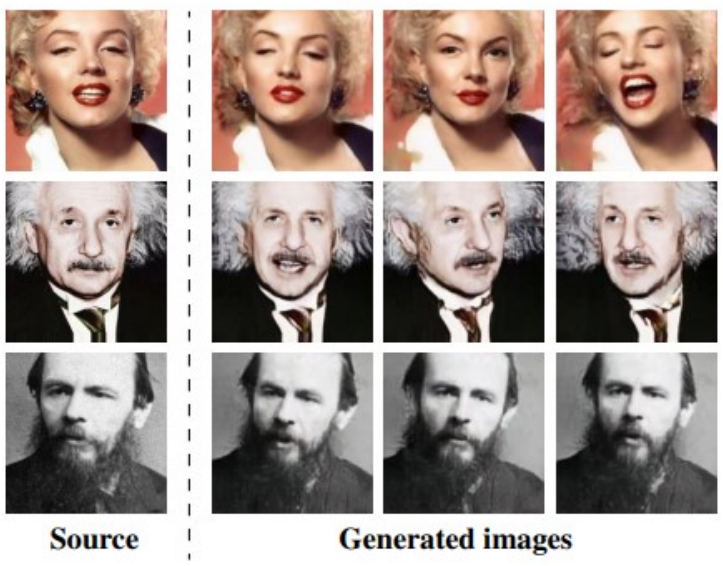

Figure 3: Results of publication Few-Shot Adversarial Learning of Realistic Neural Talking Head Models [18]

This branch aims to feed a learning model with a very small amount of training data, contrary to the traditional approach of using large amount of data. For instance, if we have a problem of categorizing birds from pictures, the input data might not have enough pictures of some rare species to be used as training images, so that would be treated as few-shot machine learning problem, looking for an approach to reduce that limitation. One of the latest papers applying few-shot adversarial learning in deepfakes developed a system that can generate frame few and one-shot neural talking head models of previously unseen people [18]. Some of the results are represented in figure 3. A latest publication in the area achieved a similar result that given an input image it could generate a fake animation. The most interesting thing about this latest project is that it was not only able to animate faces, but any specific object (faces, human bodies, cartoons), where the system have already been trained with a set of videos of that object [13].

3 Deepfake Detection

With all these advances in the area of deepfakes, it has become easier to generate them and the input data size constraint has been lowered. Therefore, the detection tasks have become more important to recognise these fake synthetic media. Early techniques were based in obtaining handcrafted

features from artifacts and inconsistencies of the fake videos. On the other hand, recent methods are based on deep learning to automatically extract this features.

The task of deepfake detection is a classification problem which goal is to distinguish between tampered images and original ones. To accomplish that, most of the techniques require a large amount of input data, meaning a large amount of tampered or deepfake images. To tackle this issue, Korshunov and Marce [7] produced a significantly big amount of deepfake data consisting of 620 videos based on the GAN model previously explained.

In deepfake detection two major categories can be distinguish: fake image detection and fake video detection. In the latter, apart from analyzing the features, it is possible to analyze the temporal features between frames based on deep learning recurrent classification models.

3.1 Fake Image Detection

One of the most dangerous techniques in Fake Image Detection is *face swapping*, as cyber-attackers could use this technique to penetrate in identification or authentication systems and gain illegitimate access. The application of GANs to enhance deepfakes have made challenging the job of detecting face swapped images, as they can imitate the pose and lighting conditions of the target image. One of the approaches with better generalization capability was introduced by Xuan et al. [17], which applied an image preprocessing step, for instance, Gaussian blur, to remove low-level clues of GAN images. This forces the forensic classifier to learn deeper and meaningful features of the images.

3.2 Fake Video Detection

Most image detection methods are not suitable for video detection because of the strong degradation of each frame after video compression. Besides, videos possess temporal features that vary among the set of frames and that can be used for the detection. Recurrent Convolutional Networks (RCNs) could be applied to exploit these temporal discrepancies across frames. For instance, a study used the physiological signal of eye blinking to expose deepfakes, based on the evidence that a person in deepfakes has much less frequent blinking than in real videos [9]. To extract this information from the tampered videos a CNN based on long short term memory (LTSM) was used. On top of that, recent research funded by Google and DARPA (research wing of the Pentagon) presented a digital forensic technique able to detect world leaders or celebrities deepfakes. The method uses machine learning to analyze specific individual's *style of speech and movement*, what researchers reference as a "soft biometric signature" [2].

4 Deepfake Tools and Demo

Nowadays there exist multiple open source projects to generate deepfakes that any user with no additional experience in the field could use. After researching the state-of-the-art ones, they have been represented in table 1. The purpose of this list is to enhance efforts into research and development of deepfakes and their detection and not to support the creation of nefarious content.

Name	Description	Link
DeepFaceLab	- Leading software for generating deepfakes. - It contains advance techniques and workflow so that the users can achieve more professional results.	https://github.com/iperov/DeepFaceLab
FaceSwap	- Tool that utilizes deep learning to recognize and swap faces in pictures and videos.	https://github.com/deepfakes/faceswap
DFaker	- A generator of larger resolution faces masked, weirdly warped, deepfakes.	https://github.com/dfaker/df
First Order Motion Model for Image Animation	- Applicate the animation of a driving video to a unique source image. - Included Demo with Google Colab and Python	https://github.com/AliaksandrSiarohin/first-order-model
Motion Supervised co-part Segmentation	- Self-supervised deep learning method for co-part segmentation. - Included Demo with Google Colab and Python	https://github.com/AliaksandrSiarohin/motion-cosegmentation
Disrupting Deepfakes	- Tool to disrupt deep fake creation on conditional image translation networks.	https://github.com/natanielruiz/disrupting-deepfakes

Table 1: Deep Fakes Tools

Regarding face swapping, by using the Google Colab demo from the repository Motion Supervised co-part Segmentation [8], the results represented in the figure 4 were generated. The first result was produced through unsupervised segmentation, while the second one was formed using supervised part-swaps.

Figure 4: Adapted Demo of Motion Supervised co-part Segmentation [8]

On top of that, in the figure 5, there are displayed some examples from the First Order Motion Model for Image Animation Demo [13]. It was made in Google Colab modifying the original version of the demo and generating deepfakes with a picture of myself as the source image. It was tested in different driving videos, two of them represented in the figure.

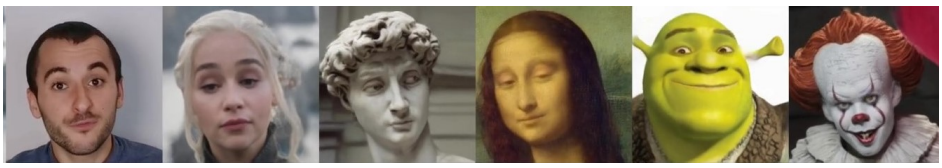

Figure 5: Adapted Demo of First Order Motion Model for Image Animation Demo [13]

5 Conclusions

The current state in deepfakes is complicated. Every time a new deepfake technology is released, users will try to develop a technology to detect the images generated with that technology, and then hackers would try to bypass that detection and ideate a new approach for generating them, and this will continue with a virus/anti-virus dynamic. On top of that, despite achieving high accuracy in deepfake detection, for instance, 99%, sometimes that 1% of no detected deepfakes could compromise users of platforms such as Instagram or Twitter.

Another idea in the area for tackling this issue is to use programs that can automatically watermark and identify images taken on cameras or implementing blockchain technology to verify content from trusted sources [15]. However, the likely future is that none of these approaches would help 'solve' this issue, it is an endless competition that can have very critical results in terms of privacy, safety and political concerns. People need to be aware of these technologies as well as verify each source they read or watch. Besides, governments should take imminent measures for people generating or sharing deepfakes for unethical uses.

References

[1] Deep convolutional generative adversarial network : Tensorflow core.

[2] Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., and Li, H. Protecting world leaders against deep fakes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2019), pp. 38–45.

[3] Badr, W. Auto-encoder: What is it? and what is it used for? (part 1), Jul 2019.

[4] Brown, N. I. Congress wants to solve deepfakes by 2020. that should worry us., Jul 2019.

[5] Damiani, J. A voice deepfake was used to scam a ceo out of $243,000, Sep 2019.

[6] Kelion, L. Reddit bans deepfake porn videos, Feb 2018.

[7] Korshunov, P., and Marcel, S. Vulnerability assessment and detection of deepfake videos. In *The 12th IAPR International Conference on Biometrics (ICB)* (2019), pp. 1–6.

[8] Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N., et al. Motion-supervised co-part segmentation. *arXiv preprint arXiv:2004.03234* (2020).

[9] Li, Y., Chang, M.-C., and Lyu, S. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)* (2018), IEEE, pp. 1–7.

[10] Manke, K., and Manke, K. Researchers use facial quirks to unmask 'deepfakes', Jun 2019.

[11] Marr, B. The best (and scariest) examples of ai-enabled deepfakes, Jul 2019.

[12] Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., and Nahavandi, S. Deep learning for deepfakes creation and detection. *arXiv preprint arXiv:1909.11573* (2019).

[13] Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., and Sebe, N. First order motion model for image animation. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 7137–7147.

[14] Suwajanakorn, S., Seitz, S. M., and Kemelmacher-Shlizerman, I. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG) 36*, 4 (2017), 1–13.

[15] Vincent, J. Deepfake detection algorithms will never be enough, Jun 2019.

[16] Wood, C. A deepfake artist's attempt to make robert de niro look younger in 'the irishman' is being hailed as superior to netflix's cgi, Jan 2020.

[17] Xuan, X., Peng, B., Wang, W., and Dong, J. On the generalization of gan image forensics. In *Chinese Conference on Biometric Recognition* (2019), Springer, pp. 134–141.

[18] Zakharov, E., Shysheya, A., Burkov, E., and Lempitsky, V. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 9459–9468.