Deepfakes for the Good: a Beneficial Application of Contentious Artificial Intelligence Technology

Nicholas Caporusso[1]

[1] Department of Computer Science, Northern Kentucky University,
Louie B Nunn Dr, 41099 Highland Heights, United States
caporusson1@nku.edu

Abstract. Deepfake algorithms are one of the most recent albeit controversial developments in Artificial Intelligence, because they use Machine Learning to generate fake yet realistic content (e.g., images, videos, audio, and text) based on an input dataset. For instance, they can accurately superimpose the face of an individual over the body an actor in a destination video (i.e., face swap), or exactly reproduce the voice of a person and speak a given text. As a result, many are concerned with the potential risks in terms of cybersecurity. Although most focused on the malicious applications of this technology, in this paper we propose a system for using deepfakes for beneficial purposes. We describe the potential use and benefits of our proposal and we discuss its implications in terms of human factors, security risks, and ethical aspects.

Keywords: Machine Learning · Deepfakes · Cybersecurity · Digital Twin

1 Introduction

In the last decade, research in the field of Artificial Intelligence (AI) advanced at an unprecedented pace. In addition to the availability of more powerful hardware resources, novel approaches to the development of Neural Networks (NN) resulted in very efficient Machine Learning (ML) toolkits and frameworks that are effectively being used in applications in several different fields, from healthcare to business, from security to education [1] [2]. Also, meta-languages enable using NNs and other algorithms with very little programming knowledge [3]. Moreover, easier and affordable access to distributed computational power enables implementing sophisticated Deep Learning (DL) algorithms that would not otherwise execute on a single computer. As multiple Graphics Processing Units working in parallel make it possible to obtain results faster and ML architectures become more mature, a new era is being introduced in the field of AI: in addition to recognizing and clustering existing information, nowadays software can generate new content (e.g., text, images, and video) that perfectly mimics an input dataset. For instance, given a gallery of head shots of different people, Generative Adversarial Networks (GAN) can learn the features of a human face and create a series of fake pictures of individuals who are not real, which studies found to be realistic enough for them not to be identified as fake [4].

Using a similar principle, deepfake algorithms can be used to learn the features of two different individuals from a source and destination datasets of head shots and use them to output images that combine the face of the individual in the source and the expressions of the individual in the destination dataset (i.e., face swap). In addition to standard face swap techniques based on landmark detection, accurately mimicking facial expressions and movements renders the resulting material very real to the casual viewer. As a result, there is a rising concern with the risks associated with the increasing quality and level of realism of the content created with this technology. Particularly, several research groups highlighted the potential danger resulting from malicious uses of deepfakes in terms of making it difficult to confirm the authenticity of information. The authors of [5] detailed the extent of the problem and described potential countermeasures, though the novelty of this technology itself makes it difficult to predict its future directions, threats, and impact. For instance, deepfakes might facilitate cyberattacks that leverage biometric traits, such as facial features, voice, or writing style, for impersonation, identity theft, and revenge cybercrime.

2 Related Work

Given the potential of their harmful applications, deepfakes are described by many as an example that demonstrates that some aspects of ML are becoming a looming challenge for privacy, democracy, and national security. In [6], the authors analyze the current social, technological, political, economic, and legislative scenario and discuss the implications in terms of digital impersonation as forged content becomes increasingly realistic and convincing. Several groups are especially concerned with images, videos, and speeches featuring political leaders and prominent figures [7], because they could be utilized to create fake news [8] aimed at initiating national scandals or international crises. Particularly, images and videos raise the most concern, because the widespread use of social networks exposes entire datasets, current image processing algorithms have greater performances with images than with audio, and videos result in major impact in a society in which communication is increasingly visual. For instance, the circulation of examples of deepfakes that involve sexual activity demonstrate that this technology could open the door to new and more aggressive types of bullying, revenge porn, and blackmailing [9]. Consequently, most research in the field is focusing on solutions that can detect and flag fake videos. To this end, several approaches can be utilized, such as evaluating the authenticity of images using algorithms that work at the pixel level to identify discrepancies that are not visible to the human viewer [5], or introducing digital fingerprints and watermarks in the source material that prevent deepfake algorithms from learning and using the facial features of the individual.

Nevertheless, the recent scientific literature started taking into consideration the potential benefits of this technology: the authors of [6] and [10] present examples that can be applied to improve education and to deliver a more personalized learning experience by creating instructional material that features characters students are more familiar with. Given the novelty of deepfake technology it is especially relevant to keep suggesting novel ideas that highlight the beneficial aspects for its use while the debate about its future directions is still ongoing.

3 A Beneficial Application of Deepfakes

In this paper, we propose an application that leverages the power of deepfake algorithms to extract an accurate model of an individual and generate new content especially designed for benign use. Specifically, in our work, we aim at using this technology to create an interactive Digital Twin of a subject that can serve as a replacement for in-person or virtual presence in the context of Cyber-Physical systems. The purpose of the proposed application is to provide users with easy-to-use tools that enable them to produce their own digital replica for future use, so that it can be featured in re-enactments, interactive stories, memorials, and simulations. For instance, deepfakes could be utilized to reinforce or surrogate an individual's physical and synchronous presence. This is especially useful in distance relationships: in contrast to conferencing tools or recorded footage, the use of video templates enables producing material without requiring the user to actually be in it. This can be particularly beneficial for families in which members are remote, as it could be utilized for creating digital storytelling books for children where grandparents are the readers. Moreover, deepfakes could be utilized to generate content for obituaries, to celebrate an individual who passed away and help their milieu cope with the loss of their beloved one. Also, deepfake videos could enable interaction with prominent contemporary or historical figures (e.g., scientists, politicians, and artists) to consolidate their legacies and keep them interactive in the form of their digital replicas.

To this end, we suggest a modification to the standard architecture of deepfake algorithms to enhance their features and make them available to individual users. Moreover, we propose an extension of deepfakes that supports generating videos programmatically, so that individuals can produce content on demand based on their media archive and on the desired type of output. The proposed application consists in (1) a material acquisition system, (2) a content processing component, (3) a deepfake generator that comprises (4) a system for programmatically producing output videos.

The material acquisition system has the purpose of enabling the user to add their images and videos: this can be realized either with a web-based interface where pre-existing footage and source files can be uploaded or with a dedicated app that with audio and video recording features especially designed to collect an input dataset with given specifications; also, the content acquisition component can be integrated with third-party systems (e.g., social networks), so that the material can be automatically imported from an external archive via a set of Application Programming Interfaces. The advantage of a dedicated app is in the possibility of prompting the user to capture input videos and images based on specific requirements in terms of light conditions, subject posture, facial expressions, and content. Conversely, importing content from external systems does not involve any additional overhead for the user, though it might produce a sparse and inaccurate dataset. Indeed, these acquisition techniques can be mixed to obtain improved results. One of the key tasks of the material acquisition process consists in categorizing the acquired videos along a timeline, to support generating multiple models that represent the subject in different stages of life.

The content processing component and the deepfake generator are the subsystems that wrap the encoder and the decoder, respectively. The former has the purpose of

extracting a model from the features of the subject based on the source and destination material. To this end, it realizes the following preliminary steps:

1. frame extraction, that is, separating the video into a sequence of images;
2. face detection, image cropping and rescaling, in order to obtain clear head shots of the subject; in addition, this can involve algorithms for improving the quality of the output, such as background removal;
3. image filtering, correction, and evaluation, which analyzes the source and classifies it into categories and descriptors that are associated with its ambient color, pose, angle, blurriness, similarity with other images.

The latter step is especially important for obtaining a dataset having adequate quantity of quality material labeled appropriately (e.g., discarding images that have high similarity). By doing so, the system can create multiple models of the subject in different stages of life and environment conditions. This, in turn, supports realizing a more accurate matchmaking with the destination videos based on their similarity in terms of aspects, such as lighting, ethnicity, and body and face shape. Subsequently, the content processing component utilizes the dataset to train the network and updates the model.

Finally, the third component of the proposed application enables the user to generate and output a deepfake by selecting a video among the templates available in the library. To this end, the decoder in the generator can swap the latent faces of the source and destination models. Alternatively, it can use generative techniques, such as GANs, to programmatically create a new scene based on a configuration provided by the user.

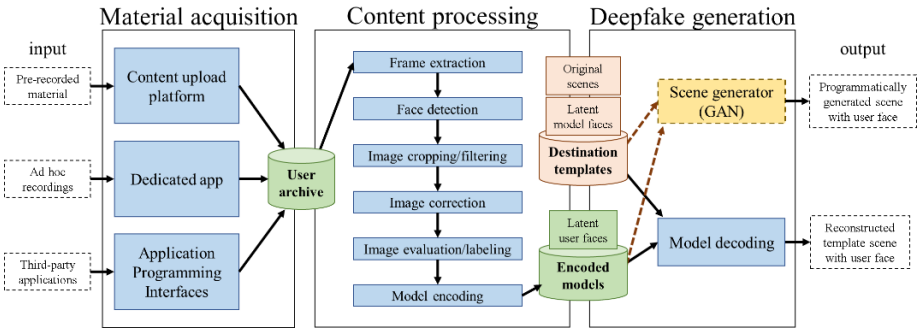

Fig. 1. The components of the proposed application and the deepfake production workflow.

4 Security, Ethical, Legal, and Human Factors

As the proposed application is deliberately designed to enable users to facilitate creating fake content in the form of images, videos, and videos, several aspects have to be taken into consideration in regard to the potential risks associated with its use. From a cybersecurity standpoint, the materials produced by the proposed application could be utilized by an attacker for malicious purposes. In this regard, the concern is three-fold and related to (1) the input content (i.e., the source material acquired from the subject), (2) the system itself, including its processes, and (3) the deepfakes generated by the system

as an output. As for the former point, any archives of personal information, and especially image galleries, involve the risk of breaches resulting in the potential misuse of the leaked material for perpetrating crimes, such as impersonation and forgery. However, nowadays individuals are accustomed with sharing their images and videos on Social Network websites with acquaintances and with the larger public. On the contrary, the proposed system is not designed as a distribution platform: for security reasons it operates as a safe box that keeps the source material private after it is recorded and uploaded by the user. In compliance with privacy and data protection regulations, the owner could be provided with the option of downloading their personal information, which could be protected by a secondary password to prevent account breaches from causing leaks of videos and images. Furthermore, the system could use predefined scripts to record interview-like sessions, as this would increase the quality and quantity of the source dataset by eliciting different facial expressions. Simultaneously, answering to pre-defined questions could prevent subjects from sharing sensitive information. Moreover, additional measures can be taken by the system to protect the visibility of the content collected from users. To this end, the source videos, images, audio, and text could be deleted or stored in encrypted archives after the model extraction algorithm has extracted and learned user's features. Secondly, the system could restrict destination videos by forcing the use of templates that are reviewed and approved by an editorial board. By doing this, in case of an account breach, the attacker has limited freedom with respect to the content of the deepfakes. The third risk element is represented by the content of the output itself: in the absence of original images and videos of the victim (or in case they are insufficient to train a model), an attacker could utilize the material produced by the proposed system and cut and assemble its parts to produce the desired message. Alternatively, output deepfakes could be exploited as a source and fed into a Deep Learning algorithm using a destination video intentionally created for malicious purposes (e.g., revenge porn or impersonation). To this end, the system could apply visible watermarks or digital fingerprints to the resulting deepfake to mark it, affect image extraction, or apply a digital tracker to any secondary material.

Moreover, several ethical aspects have to be taken into consideration with specific regard to the use of scenes and videos featuring individuals who have passed away: on the one hand, the proposed system could help the milieu cope with the loss and keep alive the memory of the beloved one, which might have a positive impact on their lives and on their overall psychological well-being; on the other hand, it might prevent full detachment after a loss and, thus, cause additional discomfort and trigger more serious mental health dynamics.

Furthermore, legal implications of the proposed system include issues related to the copyright of the input and output material, including enabling access to the system and transferring ownership, which are especially critical in case of individuals who are deceased or when dealing with videos of prominent figures.

5 Conclusion

In this paper, we primarily considered the positive aspects of deepfake technology with the objective of highlighting its potential benefits. To this end, we described an application that could be utilized to collect material from individuals at different stages of

their lives and feed it into a ML system to obtain their interactive models in the form of Digital Twins. These, in turn, can be utilized to generate deepfakes for a variety of purposes, such as producing re-enactments that might help recover memories or cope with a loss, rendering scenes that contain more realistic, image-based avatars than their three-dimensional counterparts, and serving movies, commercials, and shows with custom characters chosen by the user. Moreover, we highlighted some of the crucial security, ethical, and human factors involved in the production and use of deepfakes. In addition to presenting a new system that supports the argument for a positive perspective on this contentious technology, our paper primarily aimed at fostering the scientific debate on deepfakes and stimulating new ideas as well as critics. Despite having beneficial objectives and several advantages, the application presented in our paper might also result in users' discomfort and harmful psychological dynamics that we will evaluate in a follow-up work, in which we will also detail the result of further research that evaluates whether the benefits overcome the potential risks.

References

1. Bevilacqua, V., Carnimeo, L., Brunetti, A., De Pace, A., Galeandro, P., Trotta, G.F., Caporusso, N., Marino, F., Alberotanza, V. and Scardapane, A., 2016, December. Synthesis of a neural network classifier for hepatocellular carcinoma grading based on triphasic ct images. In International Conference on Recent Trends in Image Processing and Pattern Recognition (pp. 356-368). Springer, Singapore.
2. Bevilacqua, V., Trotta, G.F., Brunetti, A., Caporusso, N., Loconsole, C., Cascarano, G.D., Catino, F., Cozzoli, P., Delfine, G., Mastronardi, A. and Di Candia, A., 2017, July. A comprehensive approach for physical rehabilitation assessment in multiple sclerosis patients based on gait analysis. In International Conference on Applied Human Factors and Ergonomics (pp. 119-128). Springer, Cham.
3. Caporusso, N., Helms, T. and Zhang, P., 2019, July. A Meta-Language Approach for Machine Learning. In International Conference on Applied Human Factors and Ergonomics (pp. 192-201). Springer, Cham.
4. Caporusso, N., Zhang, K., Carlson, G., Jachetta, D., Patchin, D., Romeiser, S., Vaughn, N. and Walters, A., 2019, August. User Discrimination of Content Produced by Generative Adversarial Networks. In International Conference on Human Interaction and Emerging Technologies (pp. 725-730). Springer, Cham.
5. Maras, M.H. and Alexandrou, A., 2019. Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. The International Journal of Evidence & Proof, 23(3), pp.255-262.
6. Chesney, R. and Citron, D.K., 2018. Deep fakes: A looming challenge for privacy, democracy, and national security.
7. Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K. and Li, H., 2019, June. Protecting world leaders against deep fakes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 38-45).
8. Ajao, O., Bhowmik, D. and Zargari, S., 2018, July. Fake news identification on twitter with hybrid cnn and rnn models. In Proceedings of the 9th International Conference on Social Media and Society (pp. 226-230).
9. Harris, D., 2018. Deepfakes: False pornography is here and the law cannot protect you. Duke L. & Tech. Rev., 17, p.99.
10. Silbey, J. and Hartzog, W., 2018. The Upside of Deep Fakes. Md. L. Rev., 78, p.960.