The generalization challenge in deepfake image detection on human faces using generative adversarial networks (GANs)

Anonymous Author(s)

Abstract

Over the past few years, alongside the rising popularity of AI technology, deepfakes did become prevalent in our lives. Images that are generated by Generative adversarial networks can now easily fool a human due to their high quality. This comes with multiple issues, when the technologies are misused for the creation of fake news or false identities. These are problems that no one quite seemed to be able to push back. Many researchers all over the world have approached this topic but failed to build applicable solutions for our everyday lives, while deep fakes on the other hand are still improving. We propose a project to support the research in developing a well generalizing GAN that will be able to not only produce and classify training images of human faces, but also adapt to various settings using large scale datasets.

1 Introduction.

1.1 Problem definition

As technologies become more advanced, humans more often lose the competition against the machines. In some fields, we have now reached a point where machines can easily do tasks humans cannot. Looking at the two pictures (Figure 1 and 2), the question can be asked, whether we can distinguish which image is AI-generated and which one is real.

Figure 1: Human face[1]

Figure 2: Human face

[1] Both images: https://www.nytimes.com/interactive/2024/01/19/technology/artificial-intelligence-image-generators-faces-quiz.html

Most people will not guess the right solution. Both pictures above are generated by AI.

AI filters and deepfake have been developed to the point where we can no longer identify them with our own eyes [1]. Many of these technologies are very popular on social media these days because they are easily and widely available and make it simple to edit or generate more popular posts [2]. Several factors are increasing the widespread of deepfakes. For example, open source deepfake generation tools make it easier to use deep fake technology through mobile phone apps by lowering barriers to creating manipulated media. In addition, various online communities and companies have provided a platform to share technology with deep fake media, fueling the spread of deepfake [1, 3].

However, many issues that exploit false voices, images, and deepfakes tend to be dangerous in many areas of our lives. There are some examples that show that it can affect everybody. In some cases, even US presidents have been victims of deep fakes reaching from funny videos to serious political matters [4]. To name another celebrity that became a victim of deepfakes, in March 2023, a video was released that was manipulated using deepfake where Bill Gates ended the interview, embarrassed by a question about COVID vaccine during a news show interview [5]. Furthermore, most of the current legal and ethical concerns regarding synthetic media have focused on the issue of “deepfake pornography” or the non-consensual distribution of synthetically generated images, which is usually associated with an individual's face being mapped onto pornography content without the individual's knowledge [6]. Additionally, individuals and criminal organizations have already started utilizing the widely available systems to commit different crimes that involve identity theft, such as financial fraud [2, 3, 6].

As such, deepfake is increasing several risks, such as political disinformation, identity and financial fraud as well as deepfake pornography. Although most deepfake is still associated with funny celebrity videos, it opens opportunities for dangerous new fraud and slander because it evolves too quickly to be caught by authorities missing advanced protection systems [1-13]. As Generating software is widely available there is a lack of protection methods not only for individuals but also organizations with good financial resources, showing there has not yet been found a way to effectively fight the issues arising from the fast-going development in this field.

1.2 Literature review

During the past few years, deepfake generating systems have been developed many attempts to improve the detection of synthetic images. Most popular nowadays are so-called GANs, or Generative Adversarial Networks, that consist of a generator and a discriminator [2, 6-13]. During training processes, while the generation of synthetic images improves, the detection of those between natural images also improves. The improvement of these systems has become a hot topic not only in society but also in research.

Preeti et al. explained that not only manipulation of faces, but full identity swaps can nowadays be done with open-source software, allowing the public to generate high-quality and realistic synthetic images of human faces in a timespan of only a few seconds. They explain that GANs, like the so-called StyleGAN, make this real-world application possible. Furthermore, they argue that this rapid development and easy access to the public are enabling criminal misuse, such as financial fraud. They continue by stating that the generated images from modern GANs are so advanced that even humans cannot distinguish between real and fake images, leaving room for opportunities to come up with ways to computationally distinguish between them. They emphasize that among the many well-working generating models, there are only a few that bring up a discriminator that is able to perform well on new unseen data and that it is important to now build generalizing solutions to fight deepfake-based crimes [2].

Almar's study presents diverse deep learning approaches for detecting GAN-generated content, where deep learning has notably excelled. Techniques include CNN, RNN, long short-term memory (LSTM), a type of artificial recurrent neural network, and methods for analyzing statistical features of images using image preprocessing techniques and improving the detection of fake face images. He explains that each of these implementations of GAN works well on either images or videos during training and validation, while suffering from a lack of good performance on new data. Additionally, he introduces Forensic CNN, which employs Gaussian Blur and Gaussian Noise for better model generalization. However, this might bring small improvements; generalization remains challenging. He states that with the mass of new generated images every day, there is still an issue finding well-curated datasets, which might be another reason for not much visible improvement regarding the generalization of the introduced models and that this as well requires further development [7].

A team from Catania recently mentioned that GANs leave identifiable traces determined by their architecture and specific parameters when generating synthetic images. They examined whether methods that utilize frequency domain analysis to detect these traces perform well compared to augmentation methods. For instance, some detectors utilize the discrete cosine transform (DCT) to analyze images or extract features from DCT blocks to identify unique traces. In the challenge, the focus was on evaluating the detectors' resilience to alterations in images, with analytical approaches based on DCT analysis demonstrating the highest accuracy. After comparing the method to a synergy of CNN and Vision Transformers to allow large-scale classifications with both local and global vision, they found out that the application of DCT achieved high accuracy while employing an analytical approach, even with complex datasets. An important point they mentioned aside from the generalization issue, which was reduced by regularization, was that the analytical method also improved the explainability of the classification result. They explained that explainability is another important part, since humans can no longer identify fake images and are more likely to trust a system they understand [8].

In comparison to previous ones, Zha et al. introduced a different approach to solve the issues that appeared alongside the development of generative models. They explained that due to the nature of most internet platforms, videos and images get compressed while being uploaded, which leads to fading or loss of image attributes that are often the ones GAN discriminators focus on for their detection, leading to them not being useful in a real-life setting. The idea they proposed is to implement a model that focuses on the robust representations of each class using a real-centric hard feature fusion method (footnote that says: important for pattern classification using fine-grain features) to distinguish inter- and intra-class features. Even though they were not able to achieve results using this rather experimental approach further away from the state-of-the-art method, studies like this show there is still demand for new and different ways of tackling the problem. Even though they proposed a different method, they still emphasized the importance of improvement of generalization throughout the field [9].

As we can see throughout the field, most researchers agree that the main problem that needs to be solved is the lack of generalizing models to build useful systems to reliably detect deepfakes [1-3, 6-13].

1.3 Project objectives

Combining the earlier statements, our goal is to build a well-generalizing classifier for synthetic and real images. As we have seen, it is broadly agreed upon that with the development of models that generate images so close to real images that humans cannot detect the differences anymore, there is a demand for lasting solutions for systems that are able to consistently detect these deepfakes in order to prevent negative consequences of the achieved development [1-11]. There is, therefore, a need for solutions that outperform humans more and more. Our goal is, therefore, to build a system that not only is able to generate excellent fake images but also detect these, while humans cannot. Additionally, it will not be enough to build another GAN that generates images well. We need to find a way to train the discriminator inside the GAN so that it will not only detect images generated by the corresponding generator but also those created through different systems and of different quality.

2 Scope of work

The scope of work in this project will be limited to the discussed issue of creating well generalizing solutions. As the title already states, the project includes the task to set up a convolutional GAN model. The GAN architecture includes a Generator that generates new synthetic images from a random input and typically gets better during training while it tries to fool the second part of a GAN, the discrimination model. To build the model, whose performance on new data you want to later improve the discrimination between well-generated synthetic images and the natural training data, the discriminator should be a convolutional architecture that works well on classifying images, such as ResNet or other advanced solutions [10]. Due to limited time and the availability of many tutorials and pretrained models, it is suggested to use pretrained models and then working on top, making adjustments to the architecture and trying techniques to improve generalization. However, if someone wants to implement it on their own, that is acceptable as well. Whatever way the model is implemented, whether it is pretrained or build and trained from scratch, it needs to be runnable in Colab.

To train the model on classifying synthetic and real images, the synthetic images should be generated by the generator. Therefore, for the training only natural non synthetic images need to be used to make sure the model is not trained to identify generated images as natural faces. There is a dataset of natural faces collected by Nvidia that is accessible on Kaggle via https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces?select=test.csv . This dataset should be used as your training dataset. In case of using additional data, please refer from using the one provided on https://www.chicagofaces.org/download/ , since it will be used for the evaluation of the generalization and therefore should not be used for training purposes.

3 Evaluation criteria

The main evaluation criteria for this project will be the accuracy on the provided test dataset. Since we are trying to tackle the issue of the lack of generalization in current methods, the performance on new, unseen data is what we are interested in when evaluating our results. This is further the case, because deepfake generation itself is advancing at a fast pace, and a model that does not generalize well on existing unseen data will most likely not be able to detect even more advanced fakes, therefore having no chance of making an impact in the field of deepfake detection [2,11-13]. Furthermore, computational cost and test runtime are important to manage, not only because we are working with limited Colab resources and limited time, but also because if we want to apply our solution to a social media platform with millions of new images uploaded every day, the system needs to take as little time and computation as possible to be useful in the real world [2,9]. For many other issues like fraud detection, fast and cost-effective computation is also mandatory to catch the criminals before damage has been done to the victim. In conclusion, the evaluation should be done on the test accuracy with respect to runtime on the test dataset.

References

[1] Dr.A.Shaji George, A.S.Hovan George, “Deepfakes: The Evolution of Hyper Realistic Media Manipulation”, Partners Universal Innovative Research Publication (PUIRP), vol 01, 02, PU Publications, 2023, pp 58–74

[2] Preeti et al., “A GAN-Based Model of Deepfake Detection in Social Media”, Procedia Computer Science, 2023

[3] Bateman, Jon, “Deepfakes and synthetic media in the financial system: Assessing threat scenarios”, Carnegie Endowment for International Peace, 2022

[4] Helmus, Todd C., “Artificial Intelligence, Deepfakes, and Disinformation: A Primer”, *Jstor*, 2022.

[5] Jack Goodman, “Coronavirus: The Fake Bill Gates Post and Other Claims to Ignore”, *BBC News*, BBC, 2020, www.bbc.com/news/52039642.

[6] Raza, Ali et al., “A Novel Deep Learning Approach for Deepfake Image Detection”, Applied Sciences, 2022.

[7] Almars, Abdulqader M, “Deepfakes Detection Techniques Using Deep Learning: A Survey”, *Journal of Computer and Communications*, 2021

[8] Guarnera, Luca et al., “The Face Deepfake Detection Challenge”, *Journal of Imaging 8*, 2022

[9] Zha, Ruiqi et al., “*Real-centric Consistency Learning for Deepfake Detection*”, *ArXiv*, 2022

[10] Guarnera, Luca et al., “Level Up the Deepfake Detection: a Method to Effectively Discriminate Images Generated by GAN Architectures and Diffusion Model.” ArXiv, 2023

[11] Nguyen, Thanh Thi et al., “Deep learning for deepfakes creation and detection: A survey”, Comput. Vis. Image Underst. Vol. 223, 2022

[12] Chen, Liang et al., “Self-supervised Learning of Adversarial Example: Towards Good Generalizations for Deepfake Detection”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022

[13] Xiao, Shuai et al., “MCS-GAN: A Different Understanding for Generalization of Deep Forgery Detection”, IEEE Transactions on Multimedia 26, 2024

Training Dataset

140k Real and Fake Faces [Data set,] Kaggle, 2020

Test Dataset

Synthetic Faces High Quality (SFHQ) part 1 [Data set], Kaggle, 2022

Chicago Face Database, CFD, 2023