The Detectability of Fake Faces

1.0 Introduction

Seeing is no longer believing. In its most general sense, fakery runs rampant on the Internet and causes considerable inconvenience and real harm to Internet users. The faking of images is a particularly problematic issue since imagery plays a big role not only in conveying information but also in influencing our perception of trustworthiness. Worse, current advances in artificial intelligence are making the creation of fake imagery—Deepfakes—relatively easy. Deepfakes could make a person appear to say or do something they did not do, taking disinformation to a whole new level.

Our research is aimed at exploring whether a trained model detector is able to differentiate real photos from fake photos. We thus ask the question: What is the probability that a fake image detector will be deceived by fake images? And, by analyzing how the trained model identifies real and fake images, we also want to achieve a more practical contribution: how can the everyday Internet user lookout for false images (for instance, what features commonly give the falsity away)? As we answer these questions, we will also explore the ethical issues related to the use of fake imagery, especially with regards to issues of identity and trust.

2.0 Background Research

It is useful to begin with a definition of what we consider to be fake. If an image has been manipulated or modified and is presented (explicitly and/or implicitly) as something that it is not, the image is fake.

Fake images are often used to persuade the public and change their perception of a particular narrative. In this respect, some theories have been applied to understand the nature of image manipulation. During a panel on this topic, Aude Olivia, a professor on computer vision, theorized that we need categorical shifts to view an edited image as manipulation. Similarly, Carson Reynolds introduced an *Image Act Theory*, which was inspired by J. L. Austin's Speech Act Theory. Whereas Speech Act Theory concludes that speech is a performative utterance—by saying something, we do something—Image Act Theory adds that by altering "the way something appears", we perform a social act. Image manipulation is thus able to "accuse, misrepresent, persuade and entertain depending upon the audience, illustrator, and who is depicted".

In terms of the ethics of fake imagery, William J. Mitchell criticizes photo manipulation and discusses history in *How to Do Things with Pictures*. He notes the suggestion of Roland Barthes's position that press photography should not be considered as an "isolated structure." It always has a meaning "in communication with at least one other structure, namely the text—title, caption or article—accompanying every press photograph."[1] Indeed, fake images existed at the time when photoshopping wasn't as popular as today. Stalin was known for removing his enemies from pictures, and Senator Tydings could have lost his reelection campaign due to a fake picture of him talking with the head of the American Communist Party may have caused him to lose his reelection campaign.[2]

Today, with the advent of new technology, fake images (and even videos) commonly known as Deepfakes are created with the help of AI. Although high quality Deepfakes require powerful graphic cards to create high-quality, there are now many platforms and services that can help people make Deepfakes. Deepfakes are known to may cause a variety of issues, including the mapping the faces of female celebrity onto porn actresses, influencing voters by creating false narratives about candidates, shift stock prices (Tesla stock went down, because of Elon Musk's Deepfake), and, ultimately, undermine trust (or, in the words of Lilian Edwards, "The problem may not be so much the faked reality as the fact that real reality becomes plausibly deniable").

It is not easy to spot high quality Deepfakes, since the AI fixes its own weaknesses once one is identified. There are, fortunately, some attributes that can help us identify quality Deepfakes. For instance, mostly Deepfake faces don't blink normally, since people are usually photographed with open eyes; bad lip synching; patchy skin tone; flickers around the edges of the face; and badly rendered fine details such as hair, jewelry, teeth.[3]

In order to identify high quality Deepfakes, we have to utilize AI itself. To this end, many leading technology companies are trying to create an AI that could detect Deepfake images. AWS, Facebook, Microsoft, the Partnership on AI's Media Integrity Steering Committee, and

---

[1] https://web.stanford.edu/class/history34q/readings/Mitchell/MitchellHow.html

[2] https://www.businessinsider.com/fake-photos-history-2011-8#this-fake-picture-of-senator-tydings-talking-with-the-head-of-the-american-communist-party-may-have-caused-him-to-lose-his-reelection-campaign-11

[3] https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them

academics have come together to build the Deepfake Detection Challenge (DFDC).[4] They have created a competition on Kaggle with a full dataset, that consists of 124K videos, featuring eight facial modification algorithms and associated research paper[5], the dataset for which is available for download by using an AWS account with an IAM user and Access Keys setup.[6]

2.1 Similar Research:

Some research has been conducted exploring the ability of models to detect fake images. For instance, Rossler et al. (2019)[7] show that their system is able to reliably detect manipulated images and, more importantly, outperform human observers significantly.

Their dataset consisted of generated manipulations using computer graphics-based methods, such as Face2Face,[8] FaceSwap,[9] and learning based technologies including Deepfakes[10] and NeuralTextures.[11] Using this data, the study was able to construct an automated benchmark for facial manipulation:

> "As uploaded videos (e.g., to social networks) will be post processed in various ways, we obscure all selected videos multiple times (e.g., by unknown re-sizing, compression method and bit-rate) to ensure realistic conditions. This processing is directly applied on raw videos. Finally, we manually select a single challenging frame from each video based on visual inspection. Specifically, we collect a set of 1000 images, each image randomly taken from either the manipulation methods or the original footage. Note that we do not necessarily have an equal split of pristine and fake images nor an equal split of the used manipulation methods. The ground truth labels are hidden and are used on our host server to evaluate the classification accuracy of the submitted models. The automated benchmark allows submissions every two weeks from a single submitter to prevent overfitting."[12]

---

[4] https://www.kaggle.com/c/deepfake-detection-challenge
[5] https://arxiv.org/abs/2006.07397
[6] https://ai.facebook.com/datasets/dfdc/?fbclid=IwAR2x9go1HtD9BXEnVECjaI0yhpR20QLz3qUkFkdOv9mRQ9JKUDsxzBnEiIg
[7] https://arxiv.org/pdf/1901.08971v3.pdf
[8] Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In IEEE Conference on Computer Vision and Pattern Recognition, pages 2387–2395, June 2016.
[9] Faceswap. https://github.com/ MarekKowalski/FaceSwap/.
[10] Deepfakes github. https://github.com/ deepfakes/faceswap.
[11] Image synthesis using neural textures. ACM Transactions on Graphics 2019 (TOG), 2019.
[12] Ibid., 8

3.0 Deepfake Detection Model

Files and code used in the assignment are located in an online repository and is linked in this following footnote.[13]

3.1 Model Structure

For the purpose of this assignment, we created our own deepfake image detector by training a Convolutional Neural Network (CNN).

Layer Number	Deepfake Image Detector
1	Input Layer
2	Conv2D. layer, kernel size = 5, stride = 2, #filters=32
3	MaxPool2D, pool size = 2
4	Conv2D. layer, kernel size = 3, stride = 2, #filters=64
5	Flatten Layer
6	Sigmoid Layer

3.2 Data Collection

3.2.1 Training and Testing Data

The dataset used for the initial training and testing of the model is from a Kaggle Deepfake Detection Challenge[14], an event sponsored by AWS, Facebook, Microsoft and the Partnership on AI's Media Integrity Steering Committee. The dataset was constructed using a General Adverserial Network (GAN), a popular neural network used in the construction of deepfake images. All fake images in the dataset are labeled by '1', and all real images in the dataset are labeled by '0' when the image is loaded into the model.

3.2.2 Ethical Issues

During the course of exploring and analyzing this dataset, we found there was a lack of information regarding who constructed the dataset, which prevents from us performing an assessment of their biases. This raises the question of whether using such a dataset without specific knowledge on its acquisition is ethical. Is blind trust towards dominant technology

[13] https://github.com/ma5638/CNN-Deepfake-Detector
[14] https://www.kaggle.com/robikscube/kaggle-deepfake-detection-introduction

companies enough to validate that the dataset of real and fake images was acquired ethically and with consent? Furthermore, should data science teams be criticised for using datasets without knowledge of its acquisition? The lack of dataset transparency conducted by the challenge should be questioned, but it is unclear to what extent should data scientists who use these datasets be held accountable if there arises any ethical issues within the dataset or its acquisition.

3.3 Metrics

To evaluate the performance of the model, two key metrics will be used.

3.3.1 Precision

Precision, in the context of Deepfake images, refers to the success in accuracy the model has achieved of correctly predicting fake images i.e. given that a model has classified a certain image as a fake image, how likely is it to be correct?

The formula for precision is given as:

$$Precision = \frac{TP}{TP+FP}$$

TP=True Positives and FN=False Positives

The number of true positives is the number of fake images that have been correctly identified as fake, and the number of false negatives is the number of real images that have been incorrectly identified as fake.

3.3.2 Recall

In the context of Deepfake images, recall refers to the ability of the model to detect fake images overall i.e. given a fake image, how likely is it that the fake image be detected.

The formula for recall is given as:

$$Recall = \frac{TP}{TP+FN}$$

TP=True Positives and FN=False Negatives

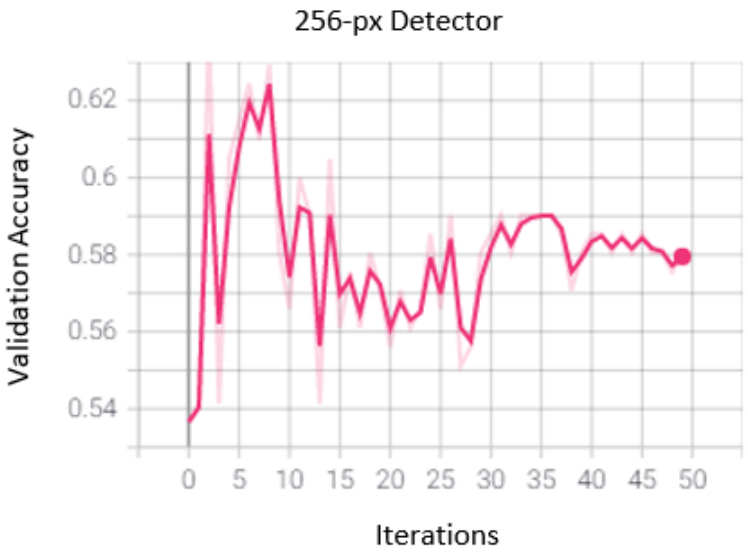
The number of true positives is the number of fake images that have been correctly identified as fake. The number of false negatives is the number of fake images that have been incorrectly identified as real.

3.4 Feature Selection

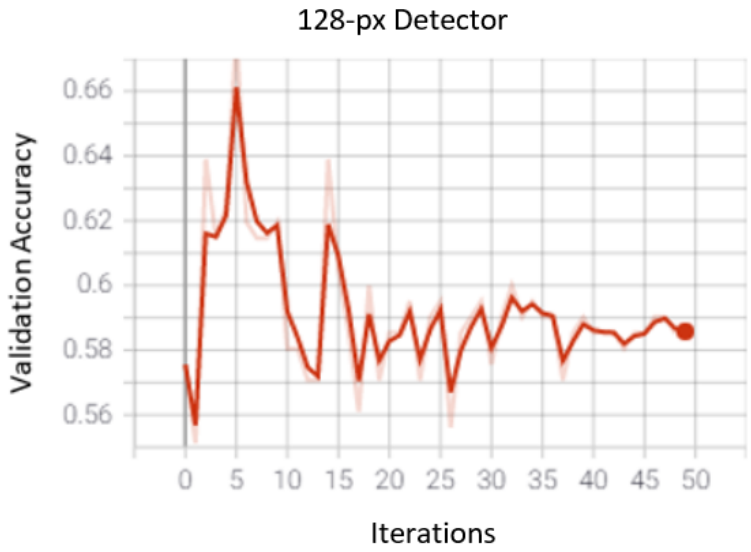
In the convolutional neural network, there were 2 variables that had to be selected and varied in an attempt to optimize the results: image size and epochs/iterations. The images loaded in the model were square in shape, and 4 different sizes for the length of the square were chosen: 32, 64, 128, and 256 pixels.

3.4.1 Iteration Selection

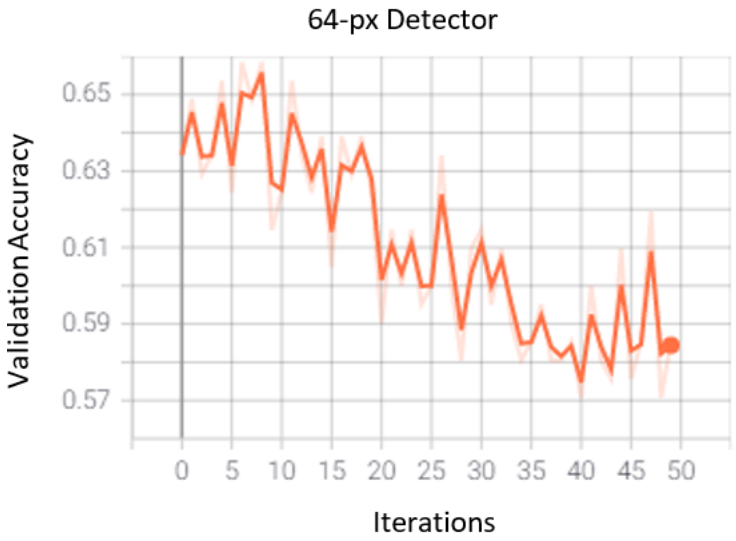
The first step was to try to optimize the 4 different detectors (of the given sizes) before a valid comparison across them. To achieve this, for the 4 detectors, 50 iterations were carried out and graphed with a smoothing factor of 0.3 to derive the number of iterations that maximizes the validation accuracy for each detector.



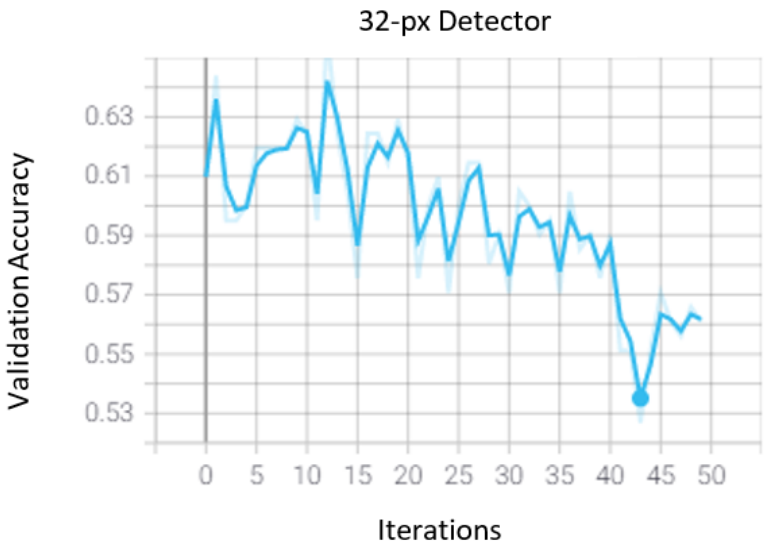
For the 256px-detector, the validation accuracy was maximized at ~8 iterations.



For the 128px-detector, the validation accuracy was maximized at 5 iterations.



For the 64px-detector, the validation accuracy was generally non-increasing after 8 iterations.



For the 32px-detector, the validation accuracy was maximal at 12 iterations.

3.4.2 Evaluation Across Image Sizes

The general hypothesis was that increasing the size of the image means that more detail was preserved and hence, led to better fake image detection.

We now began a final evaluation of the different models with their respectively set iterations. To account for randomness in training, each detector with their unique image sizes were trained 5 times and the detector with the maximum evaluation accuracy was chosen.

Image Size (px)	Accuracy	Precision	Recall
32	0.655	0.631	0.641
64	0.655	0.632	0.635
128	0.682	0.648	0.708
256	0.597	0.552	0.745

Whilst the 128px-detector did produce the highest accuracy and best balance of precision and recall, the 256-px performed poorly in accuracy and precision when compared to the other detectors. The table above shows that the accuracy of the detectors is maximized at 128 pixels. This may not mean that bigger image sizes are worse and should be avoided, but would rather

imply that having a bigger image size may require a bigger model to capture more features and details than the one used for this assignment.

However, this can mean that in order to deceive fake image detectors, much lower resolution photos may be used as performance generally decreases as the image sizes decrease. This is because whilst a larger image can be downscaled to the required image size of a fake image detector, it is often harder to upscale a smaller image to the required image size.

The results of the 128px-detector shows that given a fake image, the probability of the detector failing to detect it as a fake is $1 - Recall = 1 - 0.708 = 0.202$

3.3 Implications

As we have noted, the use of deepfake has many implications. Especially now that the world is relying on a more digital source for information. Deepfakes disseminates fake news to people, hindering their ability to do different things in different contexts.

In a political context, for instance, Deepfakes can be misused for the unfair advantage of particular candidates. Before putting a government official into office, they go through a campaign and if the photos from his/her campaign are fake then a person's trust will be put into someone who knowingly spreads false information and puts them into power which in turn will lead to said person gaining control over a group. This has already been witnessed in history by the governor of San paulo who claims that a video surfacing of him was in fact fake and was edited.

In a developmental context, the rise of Deepfake images will hinder developing countries. Many developing nations do not have the ability or resources to fact check images and information. In an interview with the BBC about Deepfake professor Li stated " It could be even more dangerous in developing countries where digital literacy is more limited. There you could really impact how society would react. You could even spread stuff that got people killed."[9] In the human centered perspective the information that can be spread takes aways the people's power and negates all their thought process in turn basically making them believe false data blindly.

Lastly, in a personal context, Deepfakes can ruin the reputation of normal people if it becomes commonly used by a large number of people. Since in this class we have discussed that the data only does what the person using it asks for and if any individual can in turn make Deepfakes about anything then they can use it for bullying purposes or for individual gain which would mentally and maybe even physically harm a person's health.

From the information that we have gathered we view that Deepfake images are very concerning to the world. Since we are moving towards a more digital world the implications of Deepfake are on the rise rather than decreasing. We view that as a community centered around data science the implications surrounding Deepfake are too big and need to be addressed in a way that the majority of the people can distinguish Deepfake images from real images.

3.4 Limitations

However, as the digital age is progressing at a very fast rate the limitations that Deepfake faces are declining. In the near future it will be a very difficult task to spot Deepfakes. Facebook has already opted to ban Deepfake images from their posts. Although there is no guarantee that in the future Deepfake images will be very difficult to spot. As we have already mentioned the difficulties around Deepfake lie within the motions of actions such as blinking, flickering and attention to detail. Even so, AI is programmed to learn from its mistakes and fix them in the future. Since our research is dedicated towards testing how to see whether the trained model detector is able to differentiate real photos from fake photos and what is the probability of the fake image detector being deceived by fake images, our model had a success rate of 70% to correctly classify a fake image, indicating that Deepfakes are not easy to spot.

Given the implantation that Deepfake has, its limitations are low and are only showing a decline in future years. From a data perspective that may be viewed as a positive since the software is doing all that it is supposed to and learning from its mistakes. On the other hand a human centered perspective might view this as an issue since the data can cause a lot of harm if used for malicious purposes. In class we often viewed ethics in a very grey area when it came to data science since the data itself isn't inherently a bad thing but if Deepfake continues to grow at this rate and be shown to have no limitations than from a ethical standpoint that can be very upsetting and in turn cause a lot of harm.

Whilst our model may indicate a 70% rate of correctly identifying fake images, there was only one single dataset used for its training and evaluation. Hence, its actual performance may be worse when a more diverse and complex dataset is included. Hence, an improvement for the model would be to use a more diverse dataset with multiple data sources.

4.0 Conclusion/Reflection

After spending the whole semester studying the principles of human-centered data science, our team was highly concerned about the issue of digital data transparency. In the age of technological advancements, people's lives become more centered around collecting information from online sources. However, the web cannot always guarantee the credibility of provided data. To take a case in point, Deepfake face images are becoming more widespread for manipulating internet users for commercial or other illegal purposes. More specifically, the (malicious) use of make faces misappropriates another's identity or creates false ones all together. Because identity is core to the human experience and accountability is one of the core tenets of sustainable data usage, it makes natural sense that we try to understand how fake faces are created, why they are created, and how we might identify the fake from the truth.

Whilst it may be easier to detect some Deepfake images right now, in the future, it is very likely that more complex technologies will arise that will increase the difficulty and complexity of detecting fake images. Eventually, there will be a really high risk of being deceived on the Internet by Deepfake pictures of people who do not exist. In this case, AI can either create biases or facilitate humans' to make smart decisions on the web. It follows then that in the future, when AI will be implemented in all the fields of our lives, it will be critically important for computer scientists and program developers to provide a smooth transition for humans in order to ensure a comfortable and transparent user adoption.