About the deepfake classifier

An introduction to deepfakes and our solution to the deepfake classification problem

DuckDuckGoose

Abstract

Deepfake videos are videos altered by a neural network. Deepfakes can be used to generate a video of a person saying things he or she never actually said, or can be used to faceswap a person into a video. This opens the door for mall-intended purposes. State of the art deepfake generation techniques are extremely good at fooling humans and virtually indistinguishable from real videos. At DuckDuckGoose we have developed a system that can detect if a video has been altered using deepfake techniques. This system obtains up to 92% accuracy on deepfake detection datasets. Next to that, we have added a module that highlights the location of artefacts introduced by deepfake generation, giving the user an idea what are the fake parts of the image.

Introduction

Modern techniques for fake video generation pose a new threat for digital communication. Especially deepfakes, which is a technique for altering images or videos with a neural network are becoming more and more realistic. Deepfake technology allows replacing a person in an image or video with the likeliness of someone else. This concept of video alteration is not new, face swap technology has been around for over a decade. However, deepfake technology uses recent improvements in artificial intelligence and deep learning techniques to increase the realism of the altered video. State of the art deepfake techniques are extremely good in fooling people that the altered video is a real one.

The number of possible use cases for deepfakes is very large, ranging from applications for sheer entertainment to educational purposes. However, just like most technologies it can be used for mall-intended purposes too. Deepfake technology makes it very easy to generate a video of a person saying things he or she never actually said. Doing this is not difficult, the newest models only need one image and they are able to make the image move and talk as if it was a video.

Furthermore, deepfakes have been used for extortion and slander. Peoples faces are being swapped with pornagraphic actors to create non-concentual pornography, which could be used for blackmail. Additionally, it is currently possible to deepfake video in real time. This can be used to create a false identity, for example in a video conference. If left unchecked, bad actors could infiltrate digital meetings, commit corporate espionage or even sabotage.

Problem statement

Deepfake generation has gotten so good that deepfake videos are virtually undistinguishable from real with the naked eye. Deepfake analysis requires an expert, or a team of experts. These experts investigate the video frame by frame, looking for artifacts created by the deepfake. These artifacts often consist of inconsistencies between pairs of frames or very subtle but unrealistic facial features.

This is a very time consuming and expensive job, considering that video analysts are not cheap. Besides, due to the monotone nature of this job the video analyst can become prone to making errors.

Solution

The solution to this problem is a system that can detect if an image or a video has been altered with deepfake techniques. This system can be used as a tool for video analysts or allow journalists to verify the authenticity of footage. At DuckDuckGoose we have developed such a system, called the DeepDetector.

The DeepDetector is a deep learning neural network designed to track down deepfake imagery. This DeepDetector can filter through large datasets to spot false imagery before it spreads. Our model is unique in the fact that it not only finds deepfakes, but it can explain its actions and reveal why a certain image is a deepfake. This makes the DeepDetector not just a simple filter tool, but a machine that can account for its choices.

The DeepDetector is trained on open source datasets of altered videos. By feeding the DeepDetector large quantities (in the order of 100.000s) of real and fake images, it learns to distinguish real images from fake images. On the test test of these datasets we achieve the following performance:

- 92% accuracy
- 92% True Positive Rate
- 9% False Positive Rate

Next to the binary classification we added a module that pinpoints the location of the artifacts introduced by the deepfake. This is done in the form of a heatmap of the input, giving a per pixel score allowing the user of the system to have an idea where the deepfake left artifacts.

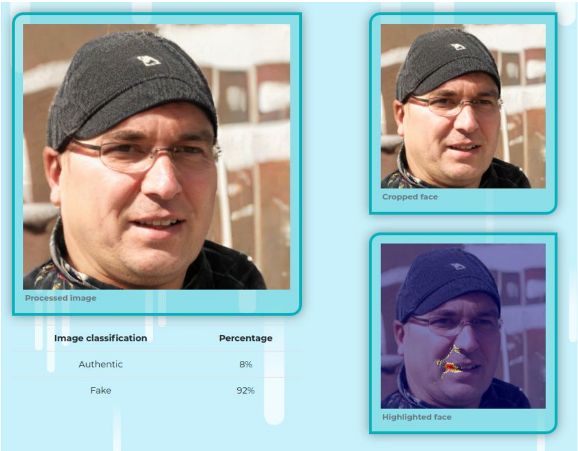

Image classification	Percentage
Authentic	8%
Fake	92%

Figure 1: Example image (image is from www.thispersondoesnotexist.com and is synthesized by an AI) and the result. It has located a suspicious region around the corner of the mouth.

Limitations

We perform five different types of experiments to measure the robustness of our classifier. The different experiments dive into face detection techniques, brightness, resolution, rotation and scale differences. In general, the results show that small changes in the images for all methods do not result in major changes in the model predictions. Furthermore, most of the major differences in outcomes do result in bigger changes in prediction, yet these can be explained by loss of information in the difference between the images. Since this is beside the scope of the current investigation this is not a major issue. Next to these general insights, some smaller insights have become apparent in this research. Since, these all have practical implications and possibilities, they are discussed in the section below. First, a quick conclusion per experiment is given.

Brightness

For the current results, we can say that all models proposed can sufficiently deal with brightness variations. However, for the extreme cases the differences grow. For real images this is more extreme for dark images whereas the brighter images the differences are larger for fake images.

Face detection

For the results of the experiment focused on face detection, the main conclusion is that in the system we should not use the MTCNN detector for now. Furthermore, we should never use a detector with scale < 1.6. It would be fruit-full to investigate if more effective detectors are available.

Resolution

For this experiment, the results show that for changes that change less than half of the quality of the original image, the network performs robustly. For lower quality, the performance drops, but this is a tricky range as it is also easier to create deep-fakes of that quality and therefore I think the task of detection these images is for now beyond the scope. For higher quality images, the results seem fine, it would be interesting to see how the network deals with high quality images created with super-resolution methods.

Rotation

For the rotation experiments, the results look good. If images are rotated over a small angle, the result remains rather equal. Cropping the image creates a loss in information, yet, it seems to result in smaller differences in prediction.

Scale

The results for small differences in scale seem good, probably since about the same information can be found in the image. For more extreme scale differences the results are poor, probably due to this change/loss in information. Before further experiments can be performed, it would be good to investigate what differences in scale will be found in practical use of the DeepDetector.

Conclusion

Deepfakes are an emerging threat for online communication. State of the art deepfake generation techniques make it very hard to judge the authenticity of a video. Analysing a video to check if it has been altered with deepfake techniques is an expensive and time consuming process done by video analysts. At DuckDuckGoose we have developed a system, the DeepDetector, that performs this job utilizing deep neural networks. It yields impressive results on publicly available deepfake detection datasets. De DeepDetector also includes a module that pinpoints the location of the artefacts introduced by deepfake generation, giving the user an idea what are the fake parts of the image.