SECURITY ISSUES IN ARTIFICIAL INTELLIGENCE

Ajith Kumar Natarajan (UBIT Name: ajithkum)
December 8, 2019

Abstract

Artificial Intelligence (AI) has been the buzz word for the last few years. Computer vision in combination with machine/deep learning techniques has proven to be a very powerful tool. As always, there are its own demerits. The ethics of AI has been called into question and has led to the demand for an explainable AI by DARPA. The recent surge in the negative usage of AI techniques have also led to demand from different personalities and organizations to form some consortium that standardizes and manages the ethics involved. In this work, I have elucidated the necessity, risks associated and offered some solutions to the threat posed to security and specifically to fake media content generated using AIs.

1 Introduction

The growth of AI technology has led to a situation wherein they can pose multiple security threats. Development of techniques such as bots planted fictitious information on social media sites like Facebook and Twitter; swaying public opinion through AI fabricated audios & videos of political figures that look and sound like the actual people and by running *denial-of-information attacks* that generate convincing fake news stories thereby making it impossible to discern legitimate information from the noise, to weaponizing *drone swarms* with self-driving capability by fitting them with small explosives and setting them loose to carry out untraceable assassinations[1] has fueled the growth for multiple demands such as that the researchers need to consider potential misuse of AI far earlier in the course of their studies than they do at present; need for consortium consisting of people from academia and industry to define and obey standards and that the governments have to work to create appropriate regulatory frameworks to prevent malicious uses of AI.

Considering the case study of *deepfakes* in this write-up, I will offer some solutions that can be incorporated to prevent and identify such synthetic images/videos, their associated risks/disadvantages, and cite analogous cases as justification.

2 Deepfakes

Deep generative models, have enabled the creation of fake images, audios and videos in ways that have not been possible before. Such fake videos, commonly known as the *DeepFakes*[2], are eroding our trust to digital media and causing serious ethical, legal, social, and financial consequences. Initially when deepfake media came into the picture, they were distinguishable by careful observation. Minor features like blinking of eyes could be used to identify them. It is not the case anymore. The research and technology has enabled the creation of such media contents in a manner that it is now almost impossible identify them. The rapid growth of the techniques warrants some steps to counter this problem.

2.1 Risks posed by deepfakes

Synthetic videos generated by deepfakes for funny, comical purposes do not pose any problem. It is the ones that are generated with the intent to manipulate people's opinions or to take revenge that push the need for a solution. There have been cases where deepfake voice have been used to launder money by impersonating high level officials to their secretaries. Incidents of deepfake videos being spread in the internet to degrade public opinion of a celebrity is also increasing.

[1] **Slaughterbots: https://www.youtube.com/watch?v=9CO6M2HsoIA**

[2] Example deepfake video of Barack Obama

2.2 Proposed solutions

- **Image/Video fingerprinting:** Similar to how hashing is used to check for corruptions in files, images and video (which itself is continuous frame of images) can have hashes generated. These can be used to verify the data integrity.
  Constraint: Though this methodology provides a way to check the integrity of the video, the number of hashes for a 3.5 minute length video would be close to 10000 (considering a frame rate of 30fps) and hence would provide a huge overhead. Adoption of blockchain as a solution would have its own difficulties such as necessity of internet to verify the authenticity of the video.
- **Addition of perturbations:** Another solution that can be thought of the addition of high frequency noise (perturbations) to each frame of the video. This noise confuses the current AI systems making it difficult for them to detect face location in the set of images thereby making it not possible to morph the face image.
  Constraint: High-pass filtering of the images from the video might alleviate the imposed security even if it comes at the cost of losing small details in the image.
- **Open-sourcing the algorithms** This idea is inspired from the context of Linux development. Since the source code of Linux is available online, anyone who finds a bug is able to report it and anyone is able to fix the issue. This has enabled Linux to safeguard itself from viruses in comparison to other operating systems. An analogous system of incorporating wide base of opinion would provide multiple ideas to secure media from the menace of deepfakes.
- **Prevent deepfake related research publications:** While publications matter a lot in academia, it is in-turn those publications that reveal the latest security measures being taken. Revealing this information helps the violators to design counter-security techniques. This idea is analogous to what Google did. While Google was initially publishing papers about its search algorithms, competitors were able to implement them as well and hence it led Google to stop publishing them.
  Constraint: Prevention of scientific publications is a step in backward direction in the sense that it would lead to slower advancement of neighboring research fields.
- **Implementing detection of the distortion of facial structure in video sharing sites:** At this point of time, the deepfake videos created have mild distortion in the facial structure due to the warping done while creating the media. Enforced implementation of AI models that can detect and discard such videos containing distorted face structure can provide us with a solution at least until improved deepfake media creation algorithms are developed.

3 Conclusion

Artificial intelligence has vast potential, and its responsible implementation is up to us. It is an exciting new frontier with lots of positive potential — as long as the human intelligence behind it keeps thinking ahead and the developed models are built with good deed.

4 References

[1] https://www.darpa.mil/program/explainable-artificial-intelligence

[2] https://imaginenext.ingrammicro.com/networking-and-security/the-top-3-ai-security-threats

[3] https://www.theguardian.com/technology/2018/feb/21/ai-security-threats-cybercrime-political-disruption-physical-attacks-report

[4] https://stackoverflow.com/questions/596262/image-fingerprint-to-compare-similarity-of-many-images

[5] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi and P. Frossard, ”Universal Adversarial Perturbations,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 86-94. doi: 10.1109/CVPR.2017.17