COLLECTIVE HUMAN ACTION AGAINST DEEPFAKES

"All our knowledge begins with senses, proceeds to the understanding and ends with reason."

Introduction

For Immanuel Kant, our senses are the gate to perceive information from the environment and to generate our knowledge. Yet, in the age of advanced technology, our senses are easily becoming subject of manipulation. In such context, the fundamental question arises whether we, humans with manipulated sense, can continue relying on our own decision making. There has been an unprecedent progress in the quality of techniques for human image synthesis based on Artificial Intelligence (AI), which can manipulate our sense of sight. Deepfakes constitutes the most famous example of it. In just few years, many alarming examples of fake content have involved politicians, governments, technology leaders, and media celebrities. What does this mean for our future, the future of our societies and the future of our countries? What will this manipulation entail at the moment we exercise our rights as citizens and voters?

Perhaps instead of jumping into the complexity of these questions, it is worth focusing on how our collective efforts can help us preventing technology from manipulating our senses. This consideration served as a guiding principal for the solution developed by the Open|DSE team in response to the UNICRI challenge at the Hackathon for Peace, Justice and Security (The Hague, June 2019). Before proceeding with the description of the solution, let's have a closer look at the AI technology behind the creation of this fake content.

Deepfakes: how does it work

Deepfakes is the result of years of advancement in the field of AI. The technology behind it is known as Deep Learning and it dates back to the early '40s, when scientists developed the first artificial neuron, called the Perceptron.[1]
The Perceptron, in its function, resembles a neuron, the basic unit of the human brains. Nevertheless, it took few decades before scientists proved that networks of Perceptrons can possibly behave like the human visual system.[2]
Nowadays, scientists with access to reasonably powerful computers, have the created AI systems known, which, despite the term "neural", loosely resemble our brain. Nevertheless, these AI systems are capable to perceive the world in the same way we do. In fact, they can detect objects in pictures, recognize faces and understand human language.

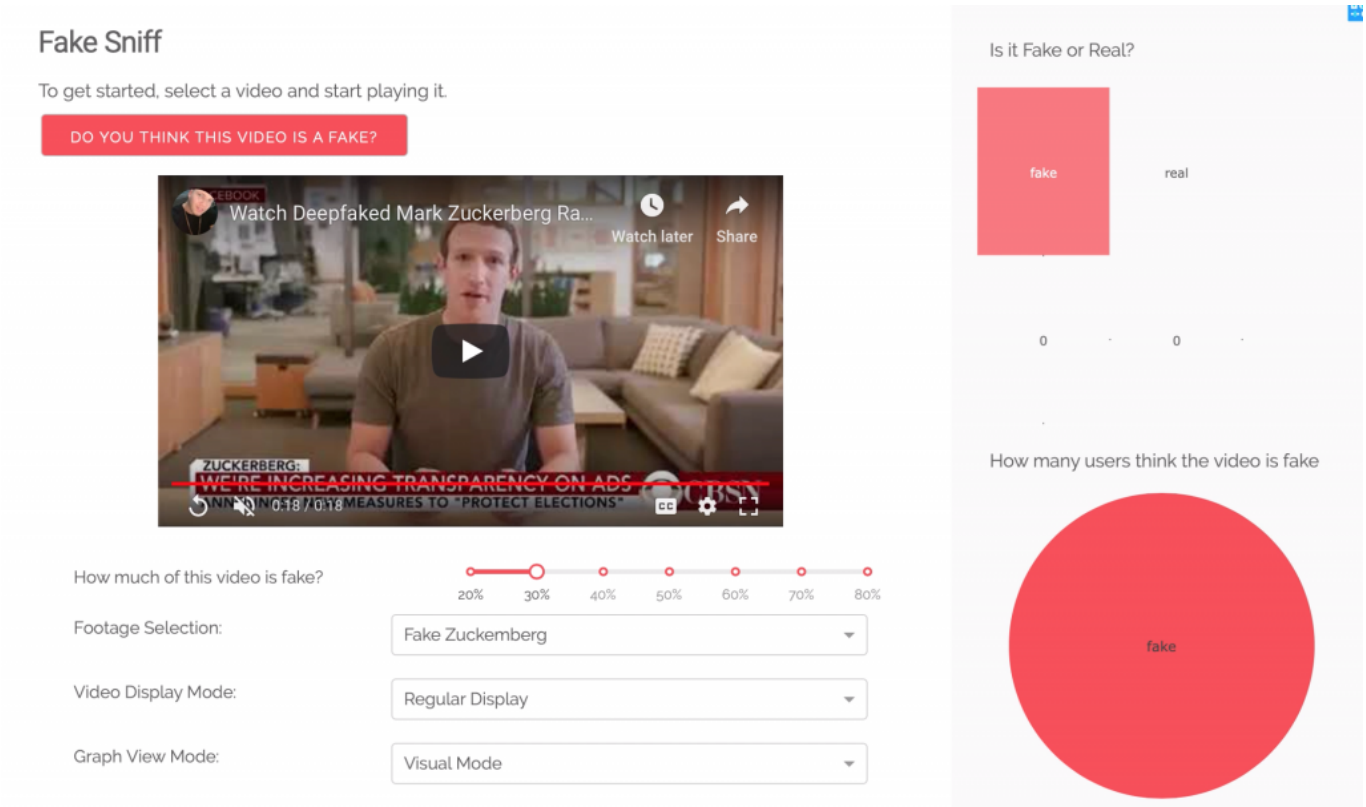
Imagine now we let

two of this AI systems play against each other. One of them tries to create fake contents, e.g., fake faces. Its name is Generator.
The opponent tries to understand if the content created by the Generator is real or fake. Its name is Discriminator. Which network wins the game? Up to now, the winner is the Generator. These systems, known as Generative Adversarial Networks (GANs), are capable to create any type of content, from images, to video, to news. The so-generated fake content is so plausible (see Figure 1) that we humans have only an average probability of 65% to recognize its falsehood.[3]

GANs are an open-source technology. It means that everyone with a good computer, such as those used for gaming, can use GANs to generate their own original content. Tech companies have promptly seen the benefits of using them for commercial purposes. For example, self-driving cars manufacturers are using them to let the cars learn how to drive autonomously using artificially generated content, e.g., weather and traffic conditions.
Unfortunately, GANs have become widely known to the general public because of their usage to spread misinformation and discreditation. Deepfakes it's the most predominant example of this usage.[4]

FakeSniff against Deepfakes

The UNICRI Centre for Artificial Intelligence has promptly recognized the need to counteract the use of Deepfakes technology for discreditation and made a call for action.[5] Data scientists, AI researchers and developers have been challenged to create innovative solutions to this rising problem. The Open|DSE team ultimately succeeded to come up with the winning solution, also gaining the grand prize of the overall hackathon. The developed solution is called FakeSniff.



FakeSniff is meant to primarily serve the general public. Single users can use FakeSniff on laptop or mobile phone to detect fake content and receive an instantaneous response in an easy and intuitive way. Nevertheless, FakeSniff contains technical possibilities to address more demanding needs of law enforcement, courts and security agencies. Recognizing these needs, we prepared the system to accommodate more complex and accurate frameworks to address the rapid change of manipulation techniques.

The formula of our solution relies on two major components. The first one is the detection module. It contains one of the most advanced Deep Neural Network available in the research community known as EfficientNet (EN). EN provides an immediate and accurate decision about fake contents, with the advantage to operate faster than other AI systems. EN has an accuracy to detect fake content of more than 97% in comparison to human judgement that stands at 65%. A companion module works alongside EN focusing on the explainability of the results. This explainability model is devoted to support law enforcement agencies in their decision-making process when using the system.

The FakeSniff interface (see Figure 2) is a web-based application with social media features. Anyone can connect to the FakeSniff website, tagging content as real or fake, providing comments about why they

think a specific content is fake, signalizing and even upload new content. The beauty of the solution is the crowdsourcing component which allows the system to be continuously updated and being able to upgrade the modules. Conscious that people judgment can be influenced by cultural, political, religious and personal biases, FakeSniff keeps currently track of the feedback to have an understanding about the public perception of a certain content. We are currently investigating how to remove the human bias and use the feedback to improve its decisions. When succeeded, FakeSniff will be able to automatically update itself and to provide more accurate decisions even in case of more advanced manipulation algorithms. We designed FakeSniff to fit multiple purposes and to be incorporated in different tools. The knowledge acquired can be increased in the future in a very efficient way and can be able to adapt the system to different contents without losing the previous knowledge.

The power of collective human action

We believe that the human factor is fundamental in the development of future technologies. FakeSniff wants to be an example of this and, hopefully, a guidance. Our bet is that the human collective intelligence will be the truly differentiator in counterbalancing the spread of untrustworthy content. Nevertheless, we, as humans, need to increase our capability of recognizing fake content: nature didn't equip us for that! FakeSniff is aimed for that goal. In this seemingly unfair GAN competition between fake and real, humans are the necessary addendum to beat the Generators. Does it mean that the war between machines and humans has started? We don't think so. Technology is only a tool and, as such, we need to drive its use in the light of a collective benefit.

The Authors

The Open|DSE team was composed by Pierluigi Casale, Vladimir Osin, Grazina Raguckaja and Giulia Violatto. The team members are a heterogeneous team of data scientists, AI researchers and project managers, working in both public and private sector organizations. Open|DSE is an open community that aims to spread knowledge and best practices in Data Science and Data Engineering. Members of Open|DSE regularly organize meetups and participate to hackathons to self-sustain the organization. For more information, visit: https://open-dse.github.io.

---

[1] "A logical calculus of the ideas immanent in nervous activity". McCulloch, W. and Pitts, W., Bulletin of Mathematical Biophysics, 5:115–133, 1943.

[2] "Neocognitron:
A hierarchical neural network capable of visual pattern recognition." Fukushima, K., Neural Networks 1:119-130, 1988.

[3] "A
Style-Based Generator Architecture for Generative Adversarial Networks", Karras, T. Laine, S. and Aila, T. , International Conference of Computer Vision and Pattern Recognition (CVPR), 2019.

[4] "The
Best (And Scariest) Examples Of AI-Enabled Deepfakes" Marr, B. Forbes, July 2019.

[5] https://www.hackathonforgood.org