Cesar Castillo, Ethan Joseph, Lael Kutame, Dmitriy Litvinenk, Michael Poropatic, Alex Ryan, Jacob Sandling

Professor Noble and Si

AI and Society

December 9, 2019

Deep Fake News Block

<https://github.com/sirmammingtonham/FakeBlock>

In recent years, the term “fake news” has spread all over the media and led people to distrust one another to disastrous effect, as any individual’s words can be taken out of context or misquoted. This has led to a general distrust of the written word, leaving videos of speeches and actions as one of the only trusted sources of proven events. As a result of this, obtaining news via video has become an incredibly common occurrence, with about one in five internet users relying on YouTube to keep up with current events.  In the past, creating a fake video of someone required a team of experts, but now, with machine learning, this process has been made accessible to anyone and can even be automated. Because they make distinguishing between real and fake reporting difficult, fraudulent videos are becoming increasingly common and dangerous. Anyone with a computer has the means to create very believable rumors and conspiracies that can influence public opinion, especially in the world of politics.

Videos modified by Artificial Intelligence are commonly known as deep fakes, referencing “deep learning”, the method used to create them. Deep fakes work by taking in large libraries of videos to identify distinct patterns in a person’s facial features, expression, movement, or voice. The neural network then effectively swaps the learned face or voice in place of another’s in the video it is given. Deep fakes began to gain public attention in 2017 when they were used to swap celebrities faces and voices into explicit videos, but they have since expanded their use widely, mainly being utilized on social media platforms.

There are many videos on the internet of politicians involved in disputable or controversial events, but these very well may be a deep fake developed by someone who wishes to harm the individual’s credibility. An example can be seen in a video of Donald Trump published in May of 2018 where he offers advice to Belgium about Climate Change: “As you know, I had the balls to withdraw from the Paris climate agreement and so should you” (Schwartz). With such vast amounts of data recordings on politicians and celebrities, it has become inherently easy to create fake voices, images, and videos of anyone to create events that never occurred. It turns out that this video of Donald Trump was actually created by a Belgium Political Party and posted on their Twitter and Facebook accounts. The video created an outrage on the internet on how the U.S. President would weigh in on Belgium’s climate policy, but in reality, it was all forged to give Donald Trump bad publicity.

This video merely scratches the surface of the threat to truth posed by deep fakes, as their increased usage could begin to break down trust and reliability in truthful journalism altogether. A video such as this making rounds on social media would likely be enough to fool many of the people who watch it, but that group could include journalists who are usually reliable and begin to spread it further. This seemingly unaltered video could then be reported on by a generally trusted and well-known news source, further propagating the falsities it was intended to spread (Alibašić). Alongside this, the existence of such indiscernible fakes could indirectly interfere with elections or even pose a threat to national security through faked propaganda (Westerlund). A faked video of a politician going viral could, even if proven fake, cause irreparable damage to their public image, thus putting the power to sway elections in the hands of anyone with access to this technology.

         The potential threats of deep fakes have not gone unnoticed, as lawmakers in the U.S. and U.K. have been working to develop legislation to protect against what they consider harmful, and mainly political, applications of this technology (Farish). This is an important distinction to draw, as not all deep fakes are created for nefarious purposes. Many movie and television studios use deep fake technology to help recreate or update footage they could not produce otherwise. Social media has even seen beneficial uses of deep fakes, as a 2019 Malaria Awareness campaign which featured David Beckham and used a neural network to make him appear multilingual, allowing celebrity endorsements to break language barriers. The constructive applications of these neural networks prove that the type of technology should not be outright restricted, but rather should be detected to help draw attention to malicious usages and distinguish what is real.

We propose to market a service, FakeBlock, that detects whether what you see online is real or AI-generated to inform readers and reduce the spread of fake news. It would work as a browser extension similar to AdBlock but for fake news. The extension itself scans the webpage for suspicious text and pictures that include faces, then uses a neural network to predict if the site’s text or pictures were created by artificial intelligence. Once we have extracted the possible deep fakes, we run a data compression detection algorithm. Based off the results of the compression we extract faces from the images, then run them through an Xception Convolutional Neural Network (CNN) trained to classify deep fakes vs. actual people. Neural networks are made up of neurons with learnable weights and biases. Each neuron receives several inputs, takes a weighted sum over them, then passes it through an activation function and responds with an output. This allows neural networks to approximate any function after training on enough data, and in our case we create a function that maps an image to a classification (e.g. deep fake or real). CNNs are an improvement to this basic neural network, which runs a convolution operation over the 3 channel image input. Xception is a further improvement to the CNN architecture which utilizes residual layers and a depthwise separable convolution operation (Chollet). This network was pre-trained on the ImageNet dataset, then fine-tuned on Google’s FaceForensics++ dataset (Rössler). The compression detection serves to increase the accuracy of our detection, since we have different networks fine-tuned with varying levels of compression from the FaceForensics++ dataset. We then pass the results from the network to FakeBlock, which removes the images classified as deep fakes from the website you are visiting.

To determine whether an article is AI generated, we compare that article word by word to the output of a GPT-2 Neural Network. A GPT-2 Neural Network is designed to predict the next word in a given sequence of words, based on how likely a human writer is to write that word (Radford). Text generated by a neural network often only contains words with a high probability. Human generated text, conversely, will typically utilize multiple words that receive a low probability in the GPT-2 network. We leverage this probability distribution to our advantage, and compare the words from the article to the probability that our GPT-2 network would output them. FakeBlock then classifies the article based on the number of words with a high probability of being AI generated. Once FakeBlock has determined an article is AI generated, it hides those paragraphs for the user.

The question that needs to be answered is the limit of what deep fake software can produce and at what point will we not be able to differentiate between an actual video and a deep fake. Despite how real deep fake created videos and photos can seem, it is not magic and is ultimately a neural network with massive amounts of data finding patterns in that data to then replicate. Although the internet makes it much easier for celebrities and people in the public eye to have large amounts of data collected about them, it would be much harder to create a believable deep fake with the lack of information available for a typical person, making this a significant limitation to deep fakes. The most effective method to counter deep fakes is to use another form of A.I to determine article authenticity (like our program FakeBlock) just as deep fakes use A.I. to replicate someone's facial expressions and gestures. Some crude videos made by someone inexperienced can even be distinguished by the naked eye, for fakes often appear distorted or blurry. With polished deep fakes impressive enough to convince the audience that the person appearing in the video is truly them, however, A.I. comes into play for detecting their legitimacy (Dickson). With the help of A.I., even minute differences such as facial expressions or changes in their voice can be detected through machine learning due to the large amount of data available on the celebrities often imitated in deep fakes. Deep fake audio is commonly made by someone impersonating the celebrity- This is why they often sound very similar to the individual. This was the case for a video posted in 2019 on Instagram of Mark Zuckerberg discussing how data is controlled (Cole). When compared to his actual voice, the video is clearly not Zuckerberg, but the video is very believable to someone who is unfamiliar with him. A.I. would be able to detect this change easily, but over time, the machine learning behind creating such videos will likely continue to advance until it cannot be identified by other A.I.. When that time comes, we need another way to combat deep fakes.

With the only limitations to deep fakes currently being information about the faked individual and the seamlessness of their voices or faces, deep fakes will continue to evolve and become more and more believable. The availability of big data is essential when it comes to who the program recreates, for information on how the person looks at various angles, how their voice fluctuates, specific gestures they make, and even the way they speak in terms of words and sentence structure must be available across the web for the neural network to pick up on. This limitation therefore makes it extremely difficult to create deep fakes of people who are not famous, since there is little to no information on them available on the internet. Deep fakes are a dangerous application of artificial intelligence with the capability to discredit officials and spread false information, but our program FakeBlock can be used to detect and remove fake content, preventing the user from becoming its next victim.

Bibliography

Alibašić, Haris & Johnathon Rose. “Fake News in Context: Truth and Untruths.” *Public Integrity*, Volume 21, Issue 5. 24 June, 2019. <https://www.tandfonline.com/doi/full/10.1080/10999922.2019.1622359>.

Chawla, Ronit. “Deepfakes: How a pervert shook the world.” *International Journal of Advance Research and Development,* Volume 4, Issue 6. 10 June, 2019. <https://www.ijarnd.com/manuscripts/v4i6/V4I6-1143.pdf>.

Chollet, François. “Xception: Deep Learning with Depthwise Separable Convolutions.” ArXiv:1610.02357 [Cs], Apr. 2017. arXiv.org, <http://arxiv.org/abs/1610.02357>.

Cole, Samantha. “This Deepfake of Mark Zuckerberg Tests Facebook's Fake Video Policies.”

*Vice*, 11 June 2019, [www.vice.com/en\_us/article/ywyxex/deepfake-of-mark-zuckerberg-facebook-fake-video-policy](http://www.vice.com/en_us/article/ywyxex/deepfake-of-mark-zuckerberg-facebook-fake-video-policy).

Dickson, Ben, et al. “How to Counter the Threat of AI-Based Forgery.” *TechTalks*, 20 Apr. 2018, [bdtechtalks.com/2018/04/16/artificial-intelligence-deepfakes-blockchain/](https://bdtechtalks.com/2018/04/16/artificial-intelligence-deepfakes-blockchain/).

Farish, Kelsey. “Do deepfakes pose a golden opportunity? Considering whether English law should adopt California’s publicity right in the age of deepfake.” *Journal of Intellectual Property Law & Practice*. 13 Nov. 2019. <https://academic.oup.com/jiplp/advance-article/doi/10.1093/jiplp/jpz139/5625122>.

Radford. Language Models Are Unsupervised Multitask Learners. 2019, <https://openai.com/blog/better-language-models/>.

Rössler, Andreas, et al. “FaceForensics++: Learning to Detect Manipulated Facial Images.” ArXiv:1901.08971 [Cs], Aug. 2019. [arXiv.org, http://arxiv.org/abs/1901.08971](http://arxiv.org/abs/1901.08971).

Schwartz, Oscar. “You Thought Fake News Was Bad? Deep Fakes Are Where Truth Goes to

Die.” *The Guardian*, Guardian News and Media, 12 Nov. 2018, [www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth](http://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth).

Westerlund, Mika. “The Emergence of Deepfake Technology: A Review.” *Technology Innovation Management Review*, Volume 9, Issue 11 : 39-52. Nov. 2019. <https://timreview.ca/sites/default/files/article_PDF/TIMReview_November2019%20-%20D%20-%20Final.pdf>.