Deepfakes

Deceptive Machine Learning

Presented by: Bob Wright
Deepfakes

What are they?

**Definition:** Media altered or generated synthetically by a Deep Neural Network with deceptive intention
Spot the Fakes
Who is real?

https://thispersondoesnotexist.com/
Made by StyleGAN2 - NVIDIA
How are they made?
Deep Neural Networks

A Neural Network is a compute system that is loosely based on the biological neural networks that constitute animal brains.

Key Idea: Automatic Feature Extraction
- Algorithms discover hierarchical features and patterns directly from data
- No Handcrafted Features
- Practical/useful machine learning!
Deep Classification

ImageNet Performance

Source: https://paperswithcode.com/sota/image-classification-on-imagenet

- SIFT + FVs
- AlexNet
- SPPNet
- ResNet-152 (60M)
- NASNET-A(6)
- FixResNeXt-101 32x48d

Other models

Models with highest Top 5 Accuracy

Human-level Performance
Early Deepfakes

Pros:
- Convincing stills and short clips
- Simple and easy to perform

Cons:
- Video artifacts
- Limited pose and range of movements

Generative Models

Next Generation Deepfakes

**Discriminative Models**
- Learns what is and is not
- Supervised learning
- e.g. classification, regression

**Generative Models**
- Learn how to describe what is
- e.g. decoders
- Show me this face
Generative Adversarial Networks

Adversarial Training

A Style-Based Generator Architecture for Generative Adversarial Networks
Tero Karras (NVIDIA), Samuli Laine (NVIDIA), Timo Aila (NVIDIA)

http://stylegan.xyz/paper

StyleGAN2 Distillation for Feed-forward Image Manipulation
Yuri Viazovetskyi, Vladimir Ivashkin, Evgeny Kashin
What can be faked?
Images and Video

https://www.buzzfeednews.com/article/davidmack/obama-fake-news-jordan-peepe-psa-video-buzzfeed

DeepFake - Jim Carrey as Jack Torrance from The Shining.
https://www.youtube.com/watch?v=HG_NZpkttXE&i
Audio Generation

It is possible to download all the require materials and duplicate anyone's voice in under two hours.

Jukebox OpenAI

We're introducing Jukebox, a neural net that generates music, including rudimentary singing, as raw audio in a variety of genres and artist styles. We're releasing the model weights and code, along with a tool to explore the generated samples.

Lyrebird

We create the most realistic artificial voices in the world.

Cybercrooks conned man with AI fake of boss's voice

Criminals used "deepfake" technology to give orders to the energy ... about the calls and emails I had received," he told The Mail on Sunday.

2 weeks ago

Language Models

OpenAI’s new language generator GPT-3 is shockingly good—and completely mindless

The AI is the largest language model ever created and can generate amazing human-like text on demand but won’t bring us closer to true intelligence.


Figure 1: The Transformer - model architecture.

The Transformer architecture as present in the Attention is all you need paper by Google.
Language Models

MIT Technology Review

Manuel: Albert Einstein. Nice to meet you. I'm Manuel Aráoz, and I've come from the future to meet you. *smiles*
Albert: *laughs* Oh, really. How is the future?
Manuel: It's amazing, actually. What year is this?
Albert: It's 1947. It's January. This is Princeton, NJ. You are in America.
Manuel: Your name is Albert Einstein, right?
Albert: That's correct.
Manuel: Do you mind if I ask you a few questions for a book I'm writing? It's about great minds of our times.
Albert: I guess I can spare a few minutes.
Manuel: Thanks so much. *sits down and asks the waiter for a coffee*
Albert: Go ahead. Ask me whatever you want.
Manuel: What are you doing here in Princeton, in the US. You were born in Germany, right?
Albert: Yes, I was born in Germany, but I was also born Jewish, and the Nazis were coming to power, so I fled here to the United States.
Manuel: How do you like it here?
Albert: It's amazing. I'm so proud of being able to contribute to science and help humanity.

OpenAI GPT-3 generates faux interview with Einstein

https://machinelearningknowledge.ai/openai-gpt-3-demos-to-convince-you-that-ai-threat-is-real-or-is-it

Surya Raj
@_suryaraj_

AI-generated faces using natural language powered by GPT-3. All these generated faces do NOT exist in real life. They are machine-generated. It'll be handy if you want to use models in your mock designs.

#ArtificialIntelligence #gpt3 #ML

7:46 AM - Sep 5, 2020
Rapid Improvements

Doom Guy – StyleGAN2 2019
Generated Media is Believable

Figure 4: Human evaluation. For each article, three annotators evaluated style, content, and the overall trustworthiness; 100 articles of each category were used. The results show that propaganda generated by Grover is rated more plausible than the original human-written propaganda.


Figure 4: Forgery detection results of our user study with 204 participants. The accuracy is dependent on the video quality and results in a decreasing accuracy rate that is 68.69% in average on raw videos, 66.57% on high quality, and 58.73% on low quality videos.
Applications & Impact
Fun
Content Creation

- AI assistants – Google Duplex
- Virtual worlds - Games and Movies
- Special effects
- Software Development
- Artificial Intelligence
- Architecture
- Art
- Music - [https://openai.com/blog/jukebox/](https://openai.com/blog/jukebox/)
Cyber Security

Art of deception

Detriments
- Targeted misclassification of detection systems
- Large-scale convincing spear phishing campaigns

Enhancements
- Elaborate Honeypots
- Hardened detection systems
Disinformation
Real Fake News

https://www.youtube.com/watch?v=cQ54Dm1elO0&feature=youtu.be
https://www.youtube.com/watch?v=xPOc9Y7niZE
Legal System Ramifications

• Is media admissible evidence anymore?
  ◦ Can media be verified?

• Existence of deepfakes → reasonable doubt?

• Should deepfakes be regulated and how?

Courts and lawyers struggle with growing prevalence of deepfakes
BY MATT REYNOLDS
JUNE 9, 2020, 9:29 AM CDT

`DEEPFAKES`: A NEW CHALLENGE FOR TRIAL COURTS

The Center for Internet and Society
Mitigations
Detection

Creating a data set and a challenge for deepfakes

https://ai.facebook.com/blog/deepfake-detection-challenge/

Fingerprinting

- Unique and prevalent artifacts in output produced by: data, model architectures, and training regimen
  - Detectable
  - Attributable
  - Reversible?

Reverse Engineering of Deceptions (RED)
Regulation and Legality

The Deepfake Report Act would require the Homeland Security Department to study the threats posed by manipulated and synthetic media—including text.


SAG-AFTRA Commends Gov. Andrew Cuomo for Signing Law Banning ‘Deep Fake’ Videos

Andrew Cuomo and the New York Legislature for signing a law protecting members from “deep fake” sexually explicit material and prohibiting ...

1 week ago

US Senate Approves New Deepfake Bill

US legislation mandating government research into deepfakes took a step closer to becoming law this week after it passed the Senate by ...

3 weeks ago
Awareness

Healthy Skepticism
Questions

Contact:
Bob Wright
wrightr@ainfosec.com

https://makemeazombie.com/#