



Repurposing Neural Networks

To Generate Synthetic Media for Information Operations

Phil Tully

Staff Data Scientist

Lee Foster

Senior Manager, Information Operations Analysis



Phil Tully

 @phtully



Machine Learning + Security Research

- Malware analysis, Threat Intelligence, Natural Language Processing

Computational Neuroscience + Neuroinformatics

- Joint PhD in Computer Science, University of Edinburgh and KTH



Lee Foster

 @LeeFosterIntel



Information Operations Intelligence Analysis

- MA Intelligence and International Security
- MA Political Science (International Relations)

Overview

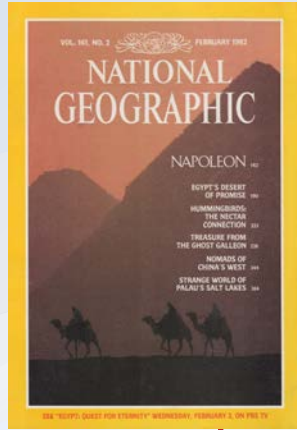
- Background
 - Synthetic Media, Generative Models, Transfer Learning
- See No Evil
 - Synthetic Image Generation with StyleGAN2
- Hear No Evil
 - Synthetic Audio Generation with SV2TTS
- Speak No Evil
 - Synthetic Text Generation with GPT-2
- Case Study
 - Social Media Information Operations
- Implications and Takeaways



Background

Synthetic Media, Generative Models, Transfer Learning

A Brief History of Synthetic Media



1930



2016

2019

time








1982




Modern Threat Environment

- Anonymity, low risk
- Immediate global reach
- Viral amplification
- Rife data disclosure
- Incentive misalignment
- Cheap content creation

Generative Models for Offensive ML

-  Select user(s) from cluster
-  Acquire timeline tweets
-  Seed LSTM / train HMM
-  Generate tailored text sequence
-  POST payload-laden tweet @target

Success Rate	High	SNAP_R > 30%	Spear Phishing ~45%
	Low	Phishing 5 -14%	
		Low	High
		Level of Effort	

Seymour and Tully, [Black Hat USA 2016](#)
Seymour and Tully, [NeurIPS 2017 Workshop on Machine Deception](#)

Generative Impersonation - Use and Misuse Cases

b NEWS WEATHER SPORTS COMMUTE FOOD CULTURE

NEWS: Coronavirus Trending Local National World Politics

A Cambridge non-profit partnered with Google to help people with ALS preserve their voice through A.I.

Project Euphon
compromised s

Mashable

VIDEO ENTERTAINMENT CULTURE TECH SCIENCE SOCIAL GOOD

Salvador Dalí deepfake brings legendary surrealist to life at Florida museum

☰

The New York Times

Deepfake Technology Enters the Documentary World

A film about persecuted gay and lesbian Chechens uses digital manipulation to guard their identities without losing their humanity. The step raises familiar questions about nonfiction movies.

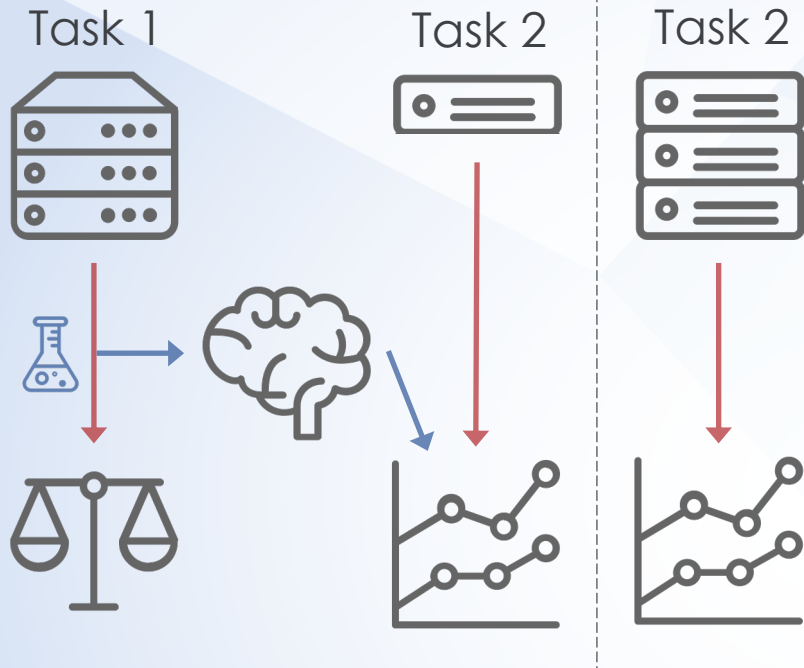
- Data Science for Good
 - Multi-lingual advertising
 - **Speech/Language Disorders**
 - **Arts & Humanities Education**
 - **Shielding Activist Identities**

Generative Impersonation - Use and Misuse Cases



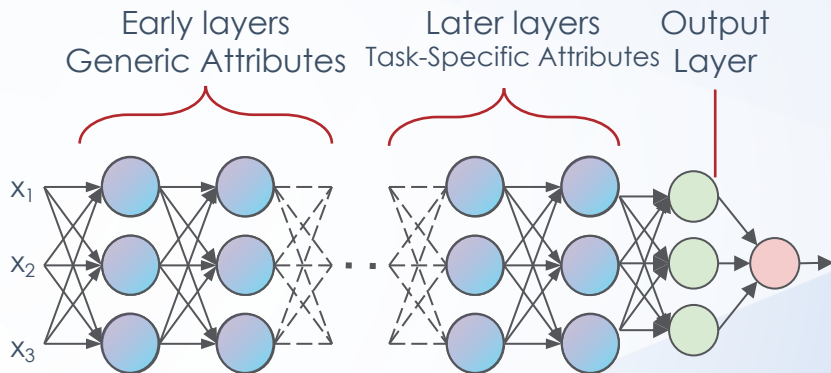
- Data Science for Good
 - Multi-lingual advertising
 - Speech/Language Disorders
 - Arts & Humanities Education
 - Shielding Identity of Activists
- Adversary Adoption
 - **Vishing and fraud**
 - News fabrication
 - Defamation, libel
 - **Revenge porn**
 - Extremist propaganda
 - Harassment, trolling, fake reviews
 - **Espionage**
 - Authentication subversion

Transfer Learning



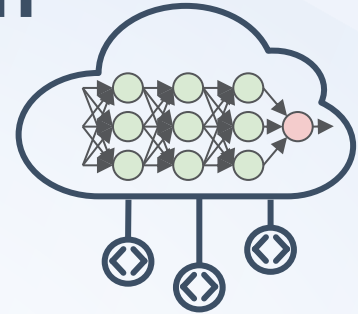
Fine-tuning strategies:

- Lower or freeze learning rates
- Architectural modifications
- Update specific weights/layers



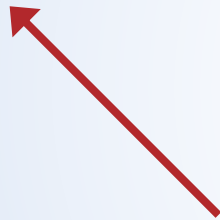
Transfer Learning = less **data, time, money, FLOPs, and energy**

The Open Source Model Ecosystem



Well-resourced industry researchers train neural networks to attain state of the art results on various tasks

They release large, pre-trained model checkpoints via open source code repositories for reproducibility



Other researchers, students, anyone anonymously downloads off-the-shelf weights for their own custom tasks



Releasing Pre-Trained Models Lowers the Barrier to Entry

- Adversaries use open source tools
- Cloud GPU Services/Notebooks
 - Authors, follow-on contributors release more code and tutorials
- Fine Tuning is not brain surgery
 - Figuratively, at least

RESOURCE	GPT-2	STYLEGAN2*	SV2TTS
Time	1+ weeks	51 yrs / 9 days	~25 days
Cost	\$43k	?	?
Data Size	40 GB	2.56 TB	~500 GB
Compute	32 TPUv3s	8 v100 GPUs	4 GTX 1080 Ti GPUs
Energy	?	131.61 MWh / 0.68 MWh	?
Released	2019	2019	2019

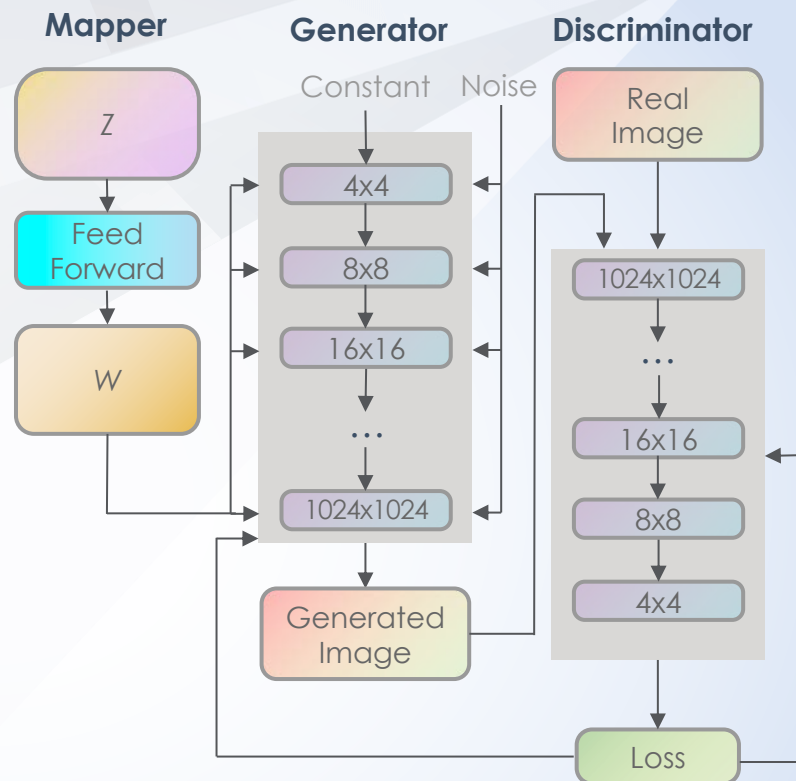
See No Evil

Synthetic Image Generation with StyleGAN2



Generative Adversarial Networks

- Synthesize indistinguishably fake images
- GAN task and architecture
 - **Mapper:** embed inputs as visual features
 - **Generator:** synthesize images from scratch
 - **Discriminator:** predict whether real images and generated images are real or fake
- Flickr-Faces HQ (FFHQ) human faces
 - 70k 1024x1024 images, ~2.56 TB
 - diverse (age, ethnicity, image background)

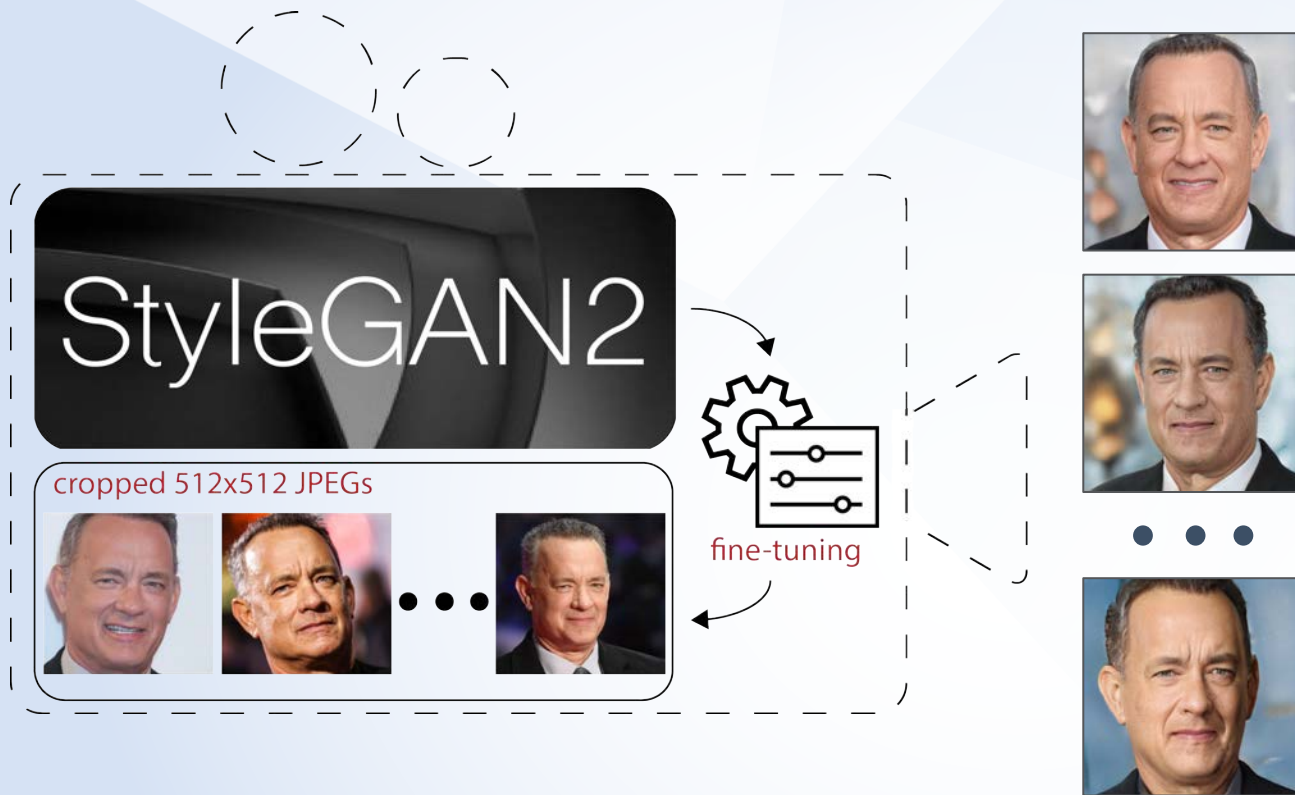


Pre-Trained StyleGAN2



e.g. <https://thispersondoesnotexist.com/>

Fine-Tuning for Custom Portraits



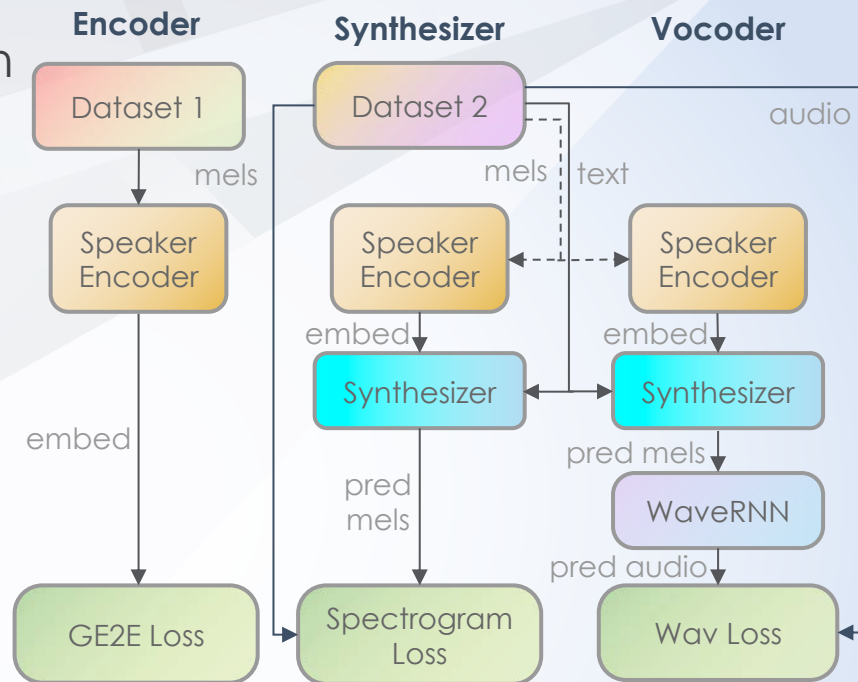
Hear No Evil

Synthetic Audio Generation with SV2TTS



Neural Voice Cloning

- Real-time text-to-speech on arbitrary voices from captured reference speech
- Sequential, 3-stage pipeline
 - Encoder** – embeds a speaker's utterance, trained on the speaker verification task
 - Synthesizer** – Tacotron2 generates spectrogram from text conditioned on Encoder's embedding
 - Vocoder** – WaveRNN infers audio waveform from Synthesizer's spectrograms
- LibriSpeech, VoxCeleb1 & 2, VCTK
 - 2,500+ hours of audio from 8,500+ speakers



Pre-Trained SV2TTS

SV2TTS toolbox

Dataset **Speaker** **Utterance**

LibriSpeech/train-cle 3112 9555/3112-9555-00

Auto select next

Use embedding from:

LibriSpeech/train-clean-100/3112/9555/3112-9555-0019.flac

Audio Input **Audio Output**

Background Music Background Music

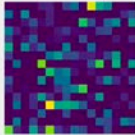
Encoder **Synthesizer** **Vocoder**

pretrained pretrained pretrained_orig

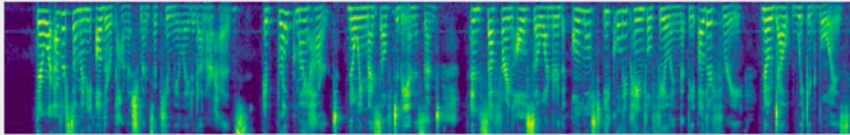
Welcome to a demonstration of neural voice cloning using an open source model
We are using a pre trained speaker encoder and providing custom input text in the upper right white box
The pre trained speaker encoder creates a vector representation of the speakers voice
These embeddings can be seen in the lower left heat maps
The voice embeddings and text embeddings are combined by a synthesizer to produce the mel spectrograms in the lower right hand corner
Lastly a vocoder takes these spectrograms and generates an audio waveform which you are hearing right now

Loading the encoder encoder/saved_models/pretrained.pt... Done (69ms).
Generating the mel spectrogram...
Loading the synthesizer synthesizer/saved_models/logs-pretrained/taco_pretrained/tacotron_model.ckpt-278000
Loading the vocoder vocoder/saved_models/pretrained/pretrained_orig.pt... Done (107ms).
Waveform generation: 551000/556800 (batch size: 58, rate: 13.7kHz - 0.85x real time) Done!

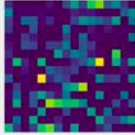
embedding



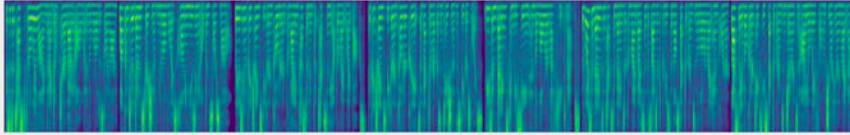
LibriSpeech/train-clean-100/3112/9555/3112-9555-0019.flac
mel spectrogram



embedding



LibriSpeech/train-clean-100_3112_gen_34277
mel spectrogram



Fine-Tuning for Speaker Adaptation

"It demonstrates that we have a common enemy but I would not count on this relationship to go beyond that. This regime has shown it will not hesitate to burn good relations for its own financial gain."

SVT2TTS

recorded ~30 second audio snippets



fine-tuning



Fine-Tuning for Speaker Adaptation

"The leaked documents clearly show that the foreign minister is corrupt and that he has misdirected funds."

SVT2TTS

recorded ~30 second audio snippets



fine-tuning



Fine-Tuning for Speaker Adaptation

"The intelligence services have indicated that these anti government protests have been organized by foreign entities. They are bent on stirring up trouble and causing harm to the people of our country."

SVT2TTS

recorded ~30 second audio snippets



fine-tuning



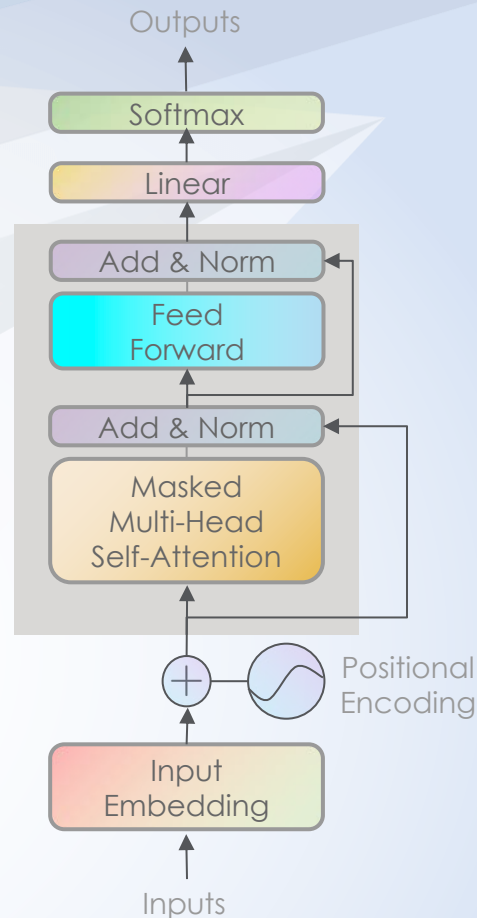
Speak No Evil

Synthetic Text Generation with GPT-2



Neural Language Models

- The quick brown fox jumps over the lazy ____
 - Successively sample from candidate word distributions
- Transformer Architecture
 - Autoregressive: stacked decoder blocks
 - Attention mechanism: masks future tokens
- Web scrape of 3+ karma Reddit URLs
 - Byte-pair encoded input words
 - > 8 million documents, ~40 GB English text



Pre-Trained GPT-2



It's disgraceful that they are not safe!"

As Masloul and Hassan struggled to release the broken glass, they struggled with each other to provide for others. Hassan also took several steps toward Masloul before becoming startled.

Earlier, security guards had shown up at the bar. They were seen entering and leaving the building on Monday night, without making a statement. Masloul reportedly received "six short-lived minutes of sitting in a cell with people who couldn't possibly comprehend the situation."

Unassailably the incident struck at home. According to the news portal Ma-Ziya, Hassan had started to touch Masloul while standing in the doorway to the cafe. Hassan then allegedly lunged with his right hand toward Masloul's shoulder. Hassan then allegedly leaned forward and grabbed Masloul's chin, mouth and throat.

Masloul claimed that when Hassan pulled his face down, his arms twisted.

He said that he fought back, and slapped Hassan in the face. When Masloul turned away to smoke a cigarette, Hassan responded by brandishing a knife.

Ma-Ziya reports that Masloul said that Hassan repeatedly grabbed his neck before running away. He claims that by that point Hassan had pulled out a knife, and wanted to strike his head on the ground, as Masloul was right in the middle of the scene.

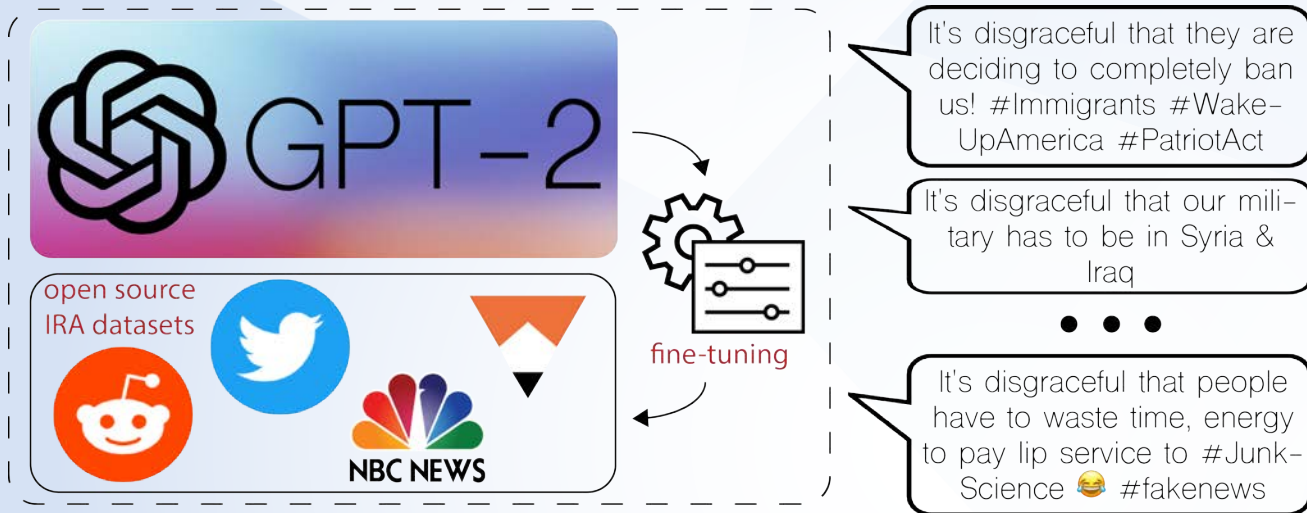
"I knew it, I just turned around, and I slapped his cheek," Masloul said. "I was thinking 'that's what I will do to him.' ... But I didn't have to do anything, he grabbed my neck, did it, and did it without realizing what was going on."

Two different witnesses to the scene reported that Hassan, who was in a cell in a security check room, was throwing letters out of the car windows at security.

The cafe is located at 442 Place Vallé, 589 Centre St. and is open to the public from 11 p.m. to 4 a.m. They seem to open early most days.

<https://www.fireeye.com/blog/threat-research/2019/11/combating-social-media-information-operations-neural-language-models.html>

Fine-Tuning for Information Operations



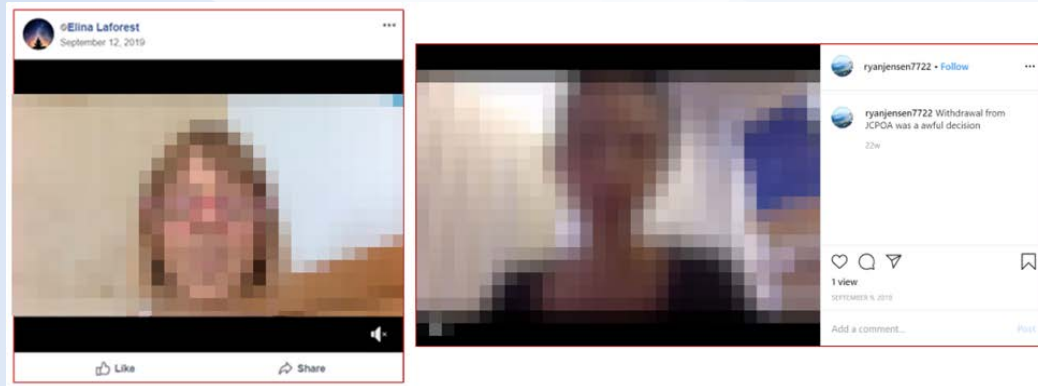
<https://www.fireeye.com/blog/threat-research/2019/11/combating-social-media-information-operations-neural-language-models.html>

Case Study

Social Media Information Operations

Some Recent IO Tactics

- Solicitation and dissemination of audio/video interviews with real experts (e.g. "Distinguished Impersonator")



- Well-developed, cross-platform personas designed to infiltrate online communities and/or disseminate fabricated content (e.g. "Ghostwriter")

<https://www.fireeye.com/blog/threat-research/2020/02/information-operations-fabricated-personas-to-promote-iranian-interests.html>

Some Recent IO Tactics

- Networks of inauthentic SM accounts amplify political narratives (e.g. Pro-China networks targeting Hong Kong protestors, pushing COVID-19 narratives)



- Personas and accounts often leverage appropriated photos of real individuals, recycle text/content

<https://www.youtube.com/watch?v=O87AYlIPsYl&t=1029s>

How Could Synthetic Media Exacerbate

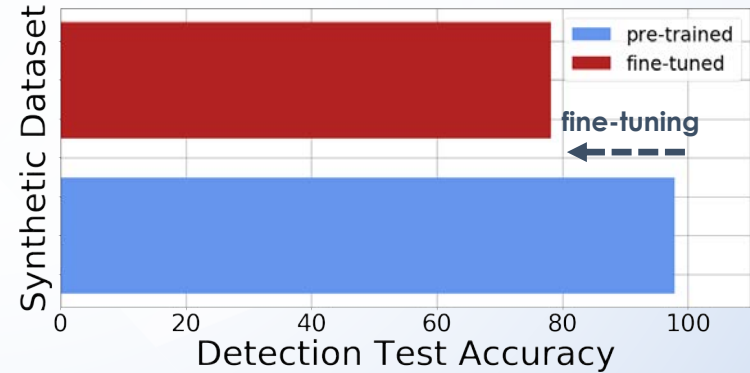
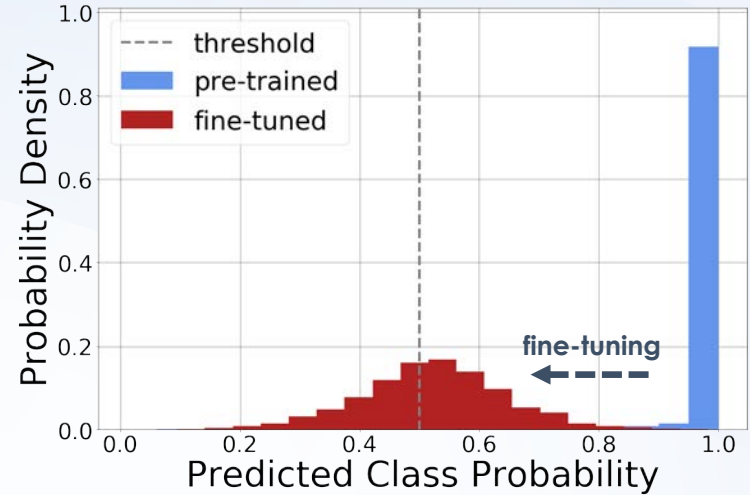
- Synthetically generated persona photos (already happening!)
 - Create convincing personas corresponding to a particular minority group to instigate political conflict, incite animosity or violence (trained on images of real people from target group or geography)
- Synthetically generated or altered audio interviews would lower actor burden, make attribution more difficult
- Synthetic text lowers barriers to creating diverse content at scale

These applications materially help threat actors scale campaigns AND evade detection



Generative Fine-Tuning for Detection Evasion

- Diversity at Scale is Problematic
 - Fine-tuning advantages attacker, who benefits from internet data availability
 - It shifts positive class probabilities towards chance, decreasing detection accuracy
- Training Data Availability Correlates w/ Target Value
 - Politicians, candidates, staffers, gov officials
 - Journalists, media personalities
 - Academics, influencers, celebrities



Synthetic Media In the Wild

Network of Inauthentic Social Media Accounts Previously

Network of Inauthentic Pro-PRC Accounts Leveraging

Suspected Deepfake Profile Photos Promotes Claim of U.S. as

Network of Spanish-Language Social Media Accounts Leveraging Suspected

Kuwait News Agency Twitter Defacement Claimed U.S. Forces support of

Are to Withdraw from Kuwait; Suspect Twitter Accounts

Promote Fabricated News from the Defacement and Accounts

Updated: Network of Inauthentic Social Media Accounts Uses

Deepfake Images as Profile Pictures, Promotes Pro-Cuban

Government and Anti-U.S. Narratives

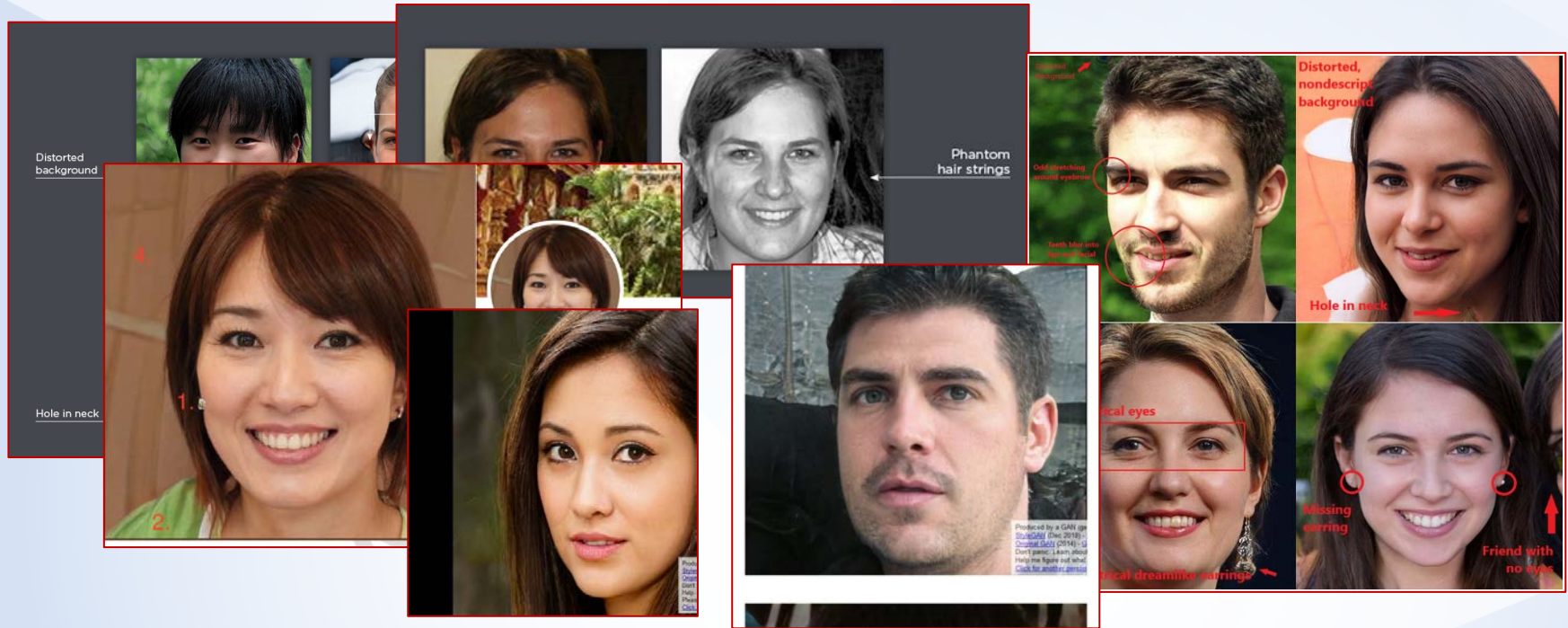
Jan 09, 2020

Jan 21, 2020

19-00016683, Version: [2]

Synthetic media being repurposed for profile pics on social media platforms for several IO campaigns

Synthetic Media In the Wild



Synthetic media being repurposed for profile pics on social media platforms for several IO campaigns

Synthetic Generation for Fun and Profit!

- Hobbyism
- An open research ethos
- Synthetic Media as-a-Service
 - Micro-Targeting
 - Personalized Advertisements and General Marketing
 - Corporate Communications and Internal Learning and Development Materials
 - Assets for Consumable Media (e.g. video game characters)

Actor Benefits of Commercial Outsourcing

- Multiple avenues of deployment mask attribution, reduce direct ties back to sponsors
- Increased diversity and/or specialization of assets and content
- Lower in-house expertise and operational investment required
- Plausible deniability and anonymity

Sponsor



PR Firm



Synthetic Generation as-a-Service





Implications and Takeaways

Technical Mitigations

■ Forgery Detection

- Statistical/ML-based
- Fingerprint/Forensics (hard to scale)
 - Poor spelling, grammar, punctuation
 - Eye alignment, teeth abnormalities, ear asymmetry, no blinking, hairline artifacts

■ Content Authentication

- Verification/other reputation signals
- Watermarking, cryptographic signing
- Controlled capture, Provenance, audit trails

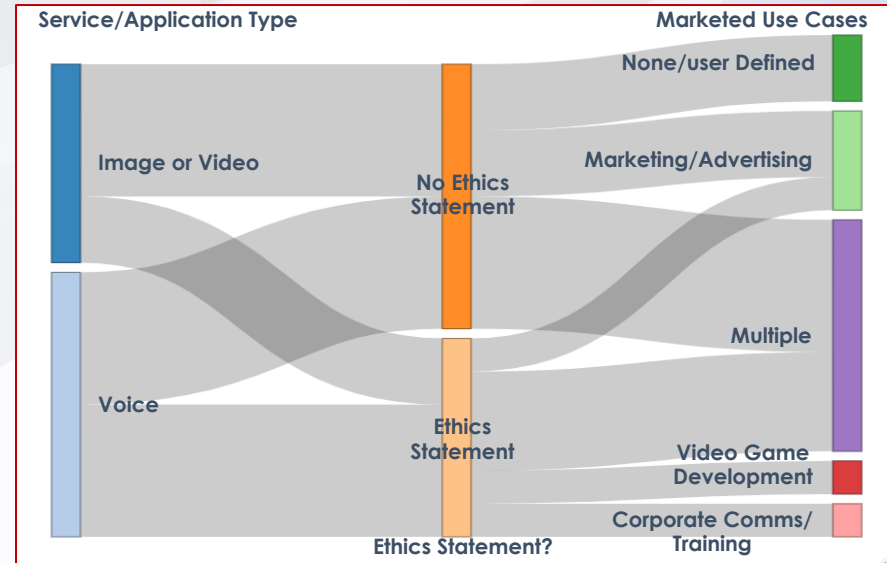
■ Platform Integrity, metadata context

- Content moderation, acct creation bottlenecks, fact-checking, policies



Patching Human Perception

- Community Efforts
 - Detection Challenges, Workshops
 - Coordination across disciplines
 - Threat modeling, red teaming
 - Acknowledgement of social Impact or Ethics Statements
- End-user Education and Awareness
 - Beware of risk hyperbole, disinformation about disinformation. Be vigilant
- Legal/Regulations (e.g. AB 730)
 - Software licensing
 - Terms of Service/Codes of Conduct



The Calm Before the Storm

- Synthetic Media tech will become cheaper, easier, more pervasive, and more credible
- New Trends Risk Further Escalation:
 - Few/One-shot learning
 - Controllability and Steerability
 - Distillation, pruning, sparsification, etc.
 - Multi-modality (text, images, *and* audio)
 - Video (deepfakes, face swap), Full body
 - Low code/no-code platforms
- User susceptibility - see what you want to see
 - Short, authoritative social media text
 - Cell-phone quality audio and video
 - Does not require high bar of credibility, only needs to be "good enough"



Black Hat Sound Bytes

- Fine tuning for generative impersonation in the text, image, and audio domains can be performed by nonexperts, can be weaponized for offensive social media-driven information operations
- Detection, attribution, and response is challenging in scenarios where actors can anonymously generate and distribute credible fake content using proprietary training datasets
- We as a community can and should help AI researchers, policy makers, and other stakeholders mitigate the harmful use of open source models



Thank you for your attention.

Acknowledgements

Sam Riddell (FireEye)
Ryan Serabian (FireEye)
Sajidur Rahman (University of Florida)
ML Visuals/dair.ai/@omarsar0/@srvmshr
Black Hat organizers and staff

