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

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

<table>
<thead>
<tr>
<th>Success Rate</th>
<th>Level of Effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Low</td>
<td>High</td>
</tr>
</tbody>
</table>

- SNAP_R > 30%
- Spear Phishing ~45%
- Phishing 5 -14%

Seymour and Tully, Black Hat USA 2016
Seymour and Tully, NeurIPS 2017 Workshop on Machine Deception
Generative Impersonation - Use and Misuse Cases

- 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

Fake-porn videos are being weaponized to harass and humiliate women: ‘Everybody is a potential target’
Transfer Learning

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

Task 1

Task 2

Task 2

Fine-tuning strategies:
• Lower or freeze learning rates
• Architectural modifications
• Update specific weights/layers

Early layers
Generic Attributes

Later layers
Task-Specific Attributes

Output Layer

x_1
x_2
x_3
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

<table>
<thead>
<tr>
<th>RESOURCE</th>
<th>GPT-2</th>
<th>STYLEGAN2*</th>
<th>SV2TTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>1+ weeks</td>
<td>51 yrs / 9 days</td>
<td>~25 days</td>
</tr>
<tr>
<td>Cost</td>
<td>$43k</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Data Size</td>
<td>40 GB</td>
<td>2.56 TB</td>
<td>~500 GB</td>
</tr>
<tr>
<td>Compute</td>
<td>32 TPUv3s</td>
<td>8 v100 GPUs</td>
<td>4 GTX 1080 Ti GPUs</td>
</tr>
<tr>
<td>Energy</td>
<td>?</td>
<td>131.61 MWh / 0.68 MWh</td>
<td>?</td>
</tr>
<tr>
<td>Released</td>
<td>2019</td>
<td>2019</td>
<td>2019</td>
</tr>
</tbody>
</table>
See No Evil
Synthetic Image Generation with StyleGAN2
Generative Adversarial Neural 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)

Analyzing and Improving the Image Quality of StyleGAN, Karras et al., 2019
Pre-Trained StyleGAN2

e.g. https://thispersondoesnotexist.com/
Fine-Tuning for Custom Portraits

StyleGAN2

cropped 512x512 JPEGs

fine-tuning
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

Transfer Learning from Speaker Verification to MultiSpeaker Text-to-Speech Synthesis, Jia et al., 2018
Automatic Multispeaker Voice Cloning, Jemine, 2019
Pre-Trained SV2TTS

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 speaker’s 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/trained_models/pretrained.pt... Done (60ms).
Generating the mel spectrogram... Done (59ms).
Loading the synthesizer synthesizersaved_models/logp-pretrained/tao_overtrained/linen_model ckpt=278000... Done (197ms).
Loading the vocoder vocodersaved_models/pretrained/linen_model ckpt=147000... Done (197ms).
Waveform generation: 551000/556800 (batch size: 50, rate: 13.74kHz ~ 0.85x real time) Done!
“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.”
Fine-Tuning for Speaker Adaptation

“The leaked documents clearly show that the foreign minister is corrupt and that he has misdirected funds.”
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.”
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 Maslouli and Hassan struggled to release the broken glass, they struggled with each other to provide for others. Hassan also took several steps toward Maslouli 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. Maslouli reportedly received “six short-lived minutes of sitting in a cell with people who couldn’t possibly comprehend the situation.”

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

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

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

Ma-Ziya reports that Maslouli 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 Maslouli was right in the middle of the scene.

“I knew it, I just turned around, and I slapped his cheek.” Maslouli 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 cell windows at security.

The cafe is located at 442 Peace Valley, S88 Centre St, and is open to the public from 11 p.m. to 4 a.m. They seem to open early most days.

Fine-Tuning for Information Operations

GPT-2

open source IRA datasets

It's disgraceful that they are deciding to completely ban us! #Immigrants #Wake-UpAmerica #PatriotAct

It's disgraceful that our military has to be in Syria & Iraq

It's disgraceful that people have to waste time, energy to pay lip service to #JunkScience 😒 #fakenews

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=O87AYIIPSYI&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 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
Synthentic 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
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