TOWARDS SECURE & INTERPRETABLE AI

SCALABLE METHODS, INTERACTIVE VISUALIZATIONS, PRACTICAL TOOLS

Polo Chau
Associate Professor
Associate Director, MS Analytics
Georgia Tech
Polo Club of Data Science

**AI** + **HI**

**ARTIFICIAL INTELLIGENCE** + **HUMAN INTELLIGENCE**

Scalable interactive tools to make sense of complex large-scale datasets and models
Human-Centered AI
- ActiVis: Visual Exploration of Facebook Deep Neural Network Models
- GAN Lab: Playing with Generative Adversarial Networks in Browser

Cyber Security
- Cyber MoneyBall: Predicting Cyber Threats with Vi Security Products
- MARCO: Fake Review Detection
  - SDM’14 Best Student Paper

Adversarial ML
- SHIELD: Fast, practical defense for deep learning
- ShapeShifter: 1st Targeted Physical Attack on Faster R-CNN Object Detector

Large Graph Mining & Visualization
- MMap: Easy billion-scale graph computation on a PC using virtual memory
- Apolo: Explore million-node graphs in real time

Social Good & Health
- DeepPop: Deep Learning on Satellite Imagery for Population Estimation
  - KDD’16 Best Student Paper, runner-up
- Firebird: Predicting Fire Risk in Atlanta
  - Deployed, Atlanta Fire Rescue Department
Today’s Main Topics

Secure AI  
Interpretable AI

Why focus on them?  
How are they related?
AI now used in safety-critical applications. Important to study threats & countermeasures.

Secure AI

How a Self-Driving Uber Killed a Pedestrian in Arizona

New York Times, 2018

The self-driving Uber was traveling north at about 40 m.p.h.
AI Security Problems Are Everywhere

"The toaster has been hacked into thinking it's a blender."

Smart toaster does exist!
AI Security is becoming increasingly important

Increased > 10-fold

# incidents reported by U.S. federal agencies

Source: Cisco

Source: US Department of Homeland Security
How do we know if a defense for AI is working?
AI models often used as black-box
Interpretable AI
Interpretable AI

Via scalable, interactive, usable interfaces to help people understand complex, large-scale ML systems.
Secure AI

Attack & Defense (DNN)
- ShapeShifter
- SHIELD

Do-it-yourself Adversarial ML
- ADAGIO
- MLsploit

Interpretable AI

Understand Industry Models
- ActiVis

Interactive Learning (Education)
- GAN Lab

Research landscape
- Survey
Our Goal

Study ML vulnerabilities and develop secure AI for high-stakes problems
Secure AI

Attack & Defense of Deep Neural Networks
- **ShapeShifter** - Physical Adversarial Attack
- **SHIELD** - Real-time Defense for *Images*

Do-it-yourself Adversarial ML
- **ADAGIO** - Experimentation with Real-time Defense for *Audio*
- **MLsploit** - Interactive Experimentation with Adversarial ML
ShapeShifter

First Targeted Physical Adversarial Attack for Object Detection

Shang-Tse Chen
Georgia Tech

Cory Cornelius
Intel

Jason Martin
Intel

Polo Chau
Georgia Tech

ECML-PKDD 2018
Image Classification

output a single label, e.g., “car”
Object Detection

recognize and localize multiple objects!
Deep Neural Networks are vulnerable
Deep Neural Networks are **vulnerable**

Classified as **Stop Sign**
Deep Neural Networks are vulnerable

Benign Image

Classified as Stop Sign

Misclassified as Max Speed 100

But most attacks have impractical threat model
Physically Realizable **Adversarial Attack**

Manipulate Physical Environment = More Realistic, *Targeted* Attack

Attacker has **no** access to internal pipeline

Digital Attack
Stop Sign $\rightarrow$ Person

Real Stop Sign

Printed Adversarial Stop Sign

car: 89%
car: 89%

stop sign: 60%
Prior Work on Physical Attacks

Glasses that fool a face classifier
[Sharif et al. CCS’16]

3D objects that fool an image classifier
[Athalye et al. ICML’18]

Stickers that fool a traffic sign classifier
[Evtimov et al. CVPR’18]

They all focus on attacking image classifiers
Attack Object Detectors: Naïve Approach

Lu et al. [1] show the current technique cannot fool state-of-the-art object detectors like Faster R-CNN and YOLO.

[1] Standard detectors aren’t (currently) fooled by physical adversarial stop signs. Lu et al., arXiv ‘17
Brief Overview of Faster R-CNN

A state-of-the-art Object Detector Model
Brief Overview of Faster R-CNN

Stage 1: Generate region proposals

Stage 2: Refined localization and classification

feature map sharing
Challenges of **Physically Attacking Faster R-CNN**

1. Multiple region proposals
2. Distances, angles, lightings
Our Solution: **Fool Multiple Region Proposals**

Minimize: \( \text{sum of classification losses} + \text{deviation loss} \)

Only perturb **RED** area

Human eye is less sensitive to changes in darker color
Our Solution: **Robust to Real-World Distortions**

Adapt **Expectation over Transformation** [Athalye et al, ICML’18]

Optimize over different backgrounds, scales, rotations, lightings
Untargeted Attack
ShapeShifter Motivates DARPA Program GARD (Defense for AI)

Highlights ShapeShifter as the state-of-the-art physical attack

Fast, Practical Defense for Image Classification

Nilaksh Das
Madhuri Shangbogue
Shang-Tse Chen
Fred Hohman
Siwei Li
Cory Cornelius
Li Chen
Michael Kounavis
Polo Chau

SHIELD

KDD’18 Audience Appreciation Award (runner-up)
KDD’19 LEMINCS

[Open-sourced]
Adversarial Machine Learning Landscape

Our Focus: Fast & Practical (digital)
ShiELD
Secure Heterogeneous Image Ensemble with Localized Denoising

"Chain Mail" (Attacked)

Labrador Retriever

Real-time Compression Preprocessing

Vaccinated Deep Neural Network Ensemble

Correctly Classified

Correctly Classified
SHIELD leverages JPEG compression

SHIELD’s SLQ applies JPEG compression of a random quality to each 8 x 8 block of the image

* larger blocks shown for presentation
**SHIELD** is a multi-pronged approach that incorporates:

- Stochastic Local Quantization
- Model Vaccination (re-training)
- Ensembling

...to mitigate adversarial attacks.
Results with ResNet-50 v2 (on ImageNet validation set)
tested on 50,000 images from the ImageNet validation set
Adversarial Attack on Speech-to-Text

**ADAGIO** incorporates compression as defense, which blocks the gradient to the attacker.
ADAGIO: Interactive Experimentation with Adversarial Attack & Defense for Audio

- Upload your own audio sample
- Perform audio adversarial attack
- Apply compression to defend
- Play audio, listen for differences

ADAGIO = Attack & Defense for Audio in a Gadget with Interactive Operations

[PKDD18]
Secure AI

Attack & Defense of Deep Neural Networks

ShapeShifter - Physical Adversarial Attack
SHIELD - Real-time Defense for Images

Do-it-yourself Adversarial ML

ADAGIO - Experimentation with Real-time Defense for Audio
MLsploit - Interactive Experimentation with Adversarial ML
A Framework for Interactive Experimentation with Adversarial Machine Learning Research


[BlackHat Asia ’19, KDD’19 Showcase]
MLsploit

- **Research modules** for adversarial ML
  - Enables *comparison* of attacks and defenses
- **Interactive experimentation** with ML research
- Researchers can **easily integrate** novel research into an intuitive and seamless *user interface*
MLsploit

★ **AVPass** (leaking and bypassing Android malware detection systems)

★ **ELF** (bypassing Linux malware detection with API perturbation)

★ **PE** (create and attack ML models for detecting Windows PE malware)

★ **Intel®-Software Guard Extensions**
  (privacy preserving adversarial ML as a service)

★ **SHIELD** (attack and defend state-of-the-art image classification models)
  * Attacks: FGSM, DeepFool, Carlini-Wagner
  * Defenses: SLQ, JPEG, Median Filter, TV-Bregman
MLsploit ARCHITECTURE
Intel® AI Courses

Learn AI theory and follow hands-on exercises with our free courses for software developers, data scientists, and students. These lessons cover AI topics and explore tools and optimized libraries that take advantage of Intel® processors in personal computers and server workstations.

- Machine Learning
- Deep Learning
- Introduction to AI
Secure AI

Attack & Defense (DNN)
- ShapeShifter
- SHIELD

Do-it-yourself Adversarial ML
- ADAGIO
- MLsploit

Interpretable AI

Understand Industry Models
- ActiVis

Interactive Learning (Education)
- GAN Lab

Research landscape
- Survey
Practitioners very interested in **why** and **how** AI works
OUR KEY IDEA

Scalable Interactive visualization as a medium for connecting users with ML models
Why interactive visualization?

Machine learning aims to find patterns from data. Visualization amplifies human cognition to find patterns.
Why interactive visualization?

By **interacting** with visualization, users can incrementally make sense of AI models.
Interpretable AI via Visual Analytics

Understanding Industry-Scale Models
- ActiVis - Activation analysis by subsets

Interactive Learning of Complex Models
- GAN Lab - Experimentation with GANs

Research Landscape
- Survey
Facebook data scientists need visualization tools to interpret complex models.
Practical Design Challenges

DEEP/WIDE MODELS
1,000+ operations/layers

LARGE DATASETS
1 billion+ instances

DIVERSE FEATURES
image, text, numerical, categorical, ...

Enjoying nice weather with kiki❤️

tags: #mycat, #cute
date: 10/1/2017
location: 33.7, 88.4
UNDERSTANDING USERS’ NEEDS

Participatory design sessions with 15+ researchers, engineers & data scientists at Facebook over 11 months
ActiVis

Visualizing activation of industry-scale deep neural nets, **deployed** by Facebook
Challenge #1

How to visualize many model parameters?

Observation: No need to show everything

Where is Mercedes-Benz Stadium located?

Number 11%
Person 8%
Location 81%

many layers
particularly useful
ActiVis Key Ideas #1

Model Overview to Activation Details

ActiVis: Visualization of Deep Neural Networks #15782570
Challenge #2
How to analyze many data instances?

Observation: Two Analytics Patterns

<table>
<thead>
<tr>
<th>INSTANCE-LEVEL</th>
<th>SUBSET-LEVEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>How model responds to individual instances?</td>
<td>How model behaves at higher-level categorization (e.g., by topic)?</td>
</tr>
<tr>
<td>Useful for debugging</td>
<td>Useful for large datasets</td>
</tr>
</tbody>
</table>
ActiVis Key Ideas (2)

Unified Analysis for Instances & Subsets
ActiVis Key Ideas (3)

Scaling Up ActiVis for Facebook

1. User-guided Instance Sampling
2. Selective Pre-computation of Layers
3. Matrix Computation for Billion-Scale Instances
Deployed on **FBLearner**
Facebook’s ML platform used by >25% of engineering team
Scalably summarize and interactively visualize neural network feature representations for millions of images.
Summit

Scalably summarize and interactively visualize neural network feature representations for millions of images

white wolf
Interprettable AI via Visual Analytics

Understanding Industry-Scale Models
  ActiVis - Activation analysis by subsets

Interactive Learning of Complex Models
  GAN Lab - Experimentation with GANs

Research Landscape
  Survey
GAN Lab
Understanding Complex Deep Generative Models using Interactive Visual Experimentation

Minsuk Kahng
Georgia Tech

Nikhil Thorat
Google

Polo Chau
Georgia Tech

Fernanda Viégas
Google

Martin Wattenberg
Google

Georgia Tech

PAIR | People + AI Research Initiative
Visualization for ML Education
Modern deep models are complex
Generative Adversarial Networks (GANs)

“the most interesting idea in the last 10 years in ML”
- Yann LeCun

Face images generated by BEGAN [Berthelot et al., 2017]
Generative Adversarial Networks (GANs)

Hard to understand and train even for experts

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))).$$

**Discriminator**

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^m \left[ \log D \left( x^{(i)} \right) + \log \left( 1 - D \left( G \left( z^{(i)} \right) \right) \right) \right].$$

**Generator**

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^m \log \left( 1 - D \left( G \left( z^{(i)} \right) \right) \right).$$
Why GANs are hard?

A GAN uses two *competing* neural networks

**Generator**
- synthesizes outputs

**Discriminator**
- spots fake

**Counterfeiter**
- makes fake bills

**Police**
- spots fake bills
Can we design an interactive tool for GANs?

1. Conceptual understanding of GANs
2. Interactive model training
3. Easily accessible by anyone
What type of data to visualize?

2D distribution, instead of high-dimensional images
What type of \textit{data} to visualize?

2D distribution, instead of high-dimensional images

Why 2D data points?

1. To focus on GAN’s main concepts
2. To easily visualize data distribution
VER. 0.1

Real (green)

Generated (purple)
How to visually explain the *generator*?
How to visually explain the generator?

map an input point into a new position
How to visualize the *discriminator*?
How to visualize the discriminator?

2D heatmap, to represent binary classification

Data points in this region are likely real.

Data points are likely fake.
Each dot is a 2D data sample: real samples, fake samples.

Background colors of grid cells represent discriminator’s classifications. Samples in green regions are likely to be real; those in purple regions likely fake.

Manifold represents generator’s transformation results from noise space. Opacity encodes density: darker purple means more samples in smaller area.
Hard to develop mental models for GANs
Each dot is a 2D data point. 

- **Real samples**: these are the actual data points that the discriminator is trained on.
- **Fake samples**: these are generated by the generator and are trying to fool the discriminator.

**Discriminator loss** measures how well the discriminator is distinguishing between real and fake samples.

**Generator loss** measures how well the generator is creating realistic samples.

- **Gradients**: Represents the change in the loss function with respect to the model parameters.
- **Noise**: Input to the generator.

**Samples** are generated from the noise input by the generator. The discriminator then tries to classify these samples as real or fake.

**Prediction of Samples** shows the probability that the discriminator assigns to each sample being real.

**Background colors** of grid cells represent the discriminator’s classifications. Samples in **green regions** are likely to be real, while those in **purple regions** are likely fake.

**Manifold** represents the generator’s transformation results from noise space. Opacity encodes density: darker purple means more samples in smaller area.

**Pink lines** from fake samples represent gradients for generator.

- **This sample needs to move up right to decrease generator’s loss.**
Draw a distribution above, then click the apply button.

APPLY
GAN Lab broadens education access

Conventional Deep Learning Visualization

Visualization in JavaScript

Model Training in Python with GPU
GAN Lab broadens education access

Everything done in browser, powered by TensorFlow.js

Visualization in JavaScript

Model Training also in JavaScript

Accelerated by WebGL
GAN Lab is Live!  Try at bit.ly/gan-lab

30K visitors, 135 countries

❤️ 1.9K Likes  🔁 800+ Retweets
Interpretable AI via Visual Analytics

Understanding Industry-Scale Models

**ActiVis** - Activation analysis by subsets

Interactive Learning of Complex Models

**GAN Lab** - Experimentation with GANs

Research Landscape

Survey
Visual Analytics in Deep Learning
An Interrogative Survey for the Next Frontiers

Fred Hohman
Georgia Tech

Minsuk Kahng
Georgia Tech

Robert Pienta
Symantec

Polo Chau
Georgia Tech
Visual Analytics in Deep Learning

**WHY**
Why would one want to use visualization in deep learning?
- Interpretability & Explainability
- Debugging & Improving Models
- Comparing & Selecting Models
- Teaching Deep Learning Concepts

**WHAT**
What data, features, and relationships in deep learning can be visualized?
- Computational Graph & Network Architecture
- Learned Model Parameters
- Individual Computational Units
- Neurons In High-dimensional Space
- Aggregated Information

**WHEN**
When in the deep learning process is visualization used?
- During Training
- After Training

**WHO**
Who would use and benefit from visualizing deep learning?
- Model Developers & Builders
- Model Users
- Non-experts

**HOW**
How can we visualize deep learning data, features, and relationships?
- Node-link Diagrams for Network Architecture
- Dimensionality Reduction & Scatter Plots
- Line Charts for Temporal Metrics
- Instance-based Analysis & Exploration
- Interactive Experimentation

**WHERE**
Where has deep learning visualization been used?
- Application Domains & Models
- A Vibrant Research Community
Key Takeaways

1. Most tools aimed at expert users
2. Instance-based analysis
3. Inherently interdisciplinary
4. Lacks actionability
5. Evaluation is hard
6. State-of-the-art models not robust
Secure AI

- Attack & Defense (DNN)
  - ShapeShifter
  - SHIELD

- Do-it-yourself Adversarial ML
  - ADAGIO
  - MLsploit

Interpretable AI

- Understand Industry Models
  - ActiVis

- ML Education
  - GAN Lab

- Research landscape
  - Survey
TOWARDS SECURE & INTERPRETABLE AI

Scalable Methods, Interactive Visualizations, Practical Tools

Thanks!

poloclub.github.io

Polo Chau
Georgia Tech