

Frontiers in ML

COSC 410: Applied Machine Learning

Spring 2022

Prof. Apthorpe

Outline

Today

- Explainable/Interpretable ML
- Transfer Learning
- Adversarial ML
- Lightweight ML
- Artificial General Intelligence

Previous Classes

- Generative ML for design
- Evolving neural network architectures
- Deep Q-learning for strategy games, video games, & robotics

Disclaimer: The state of the art is rapidly changing for all above topics

Explainable/Interpretable ML

Explainable/Interpretable ML

Why is this important?

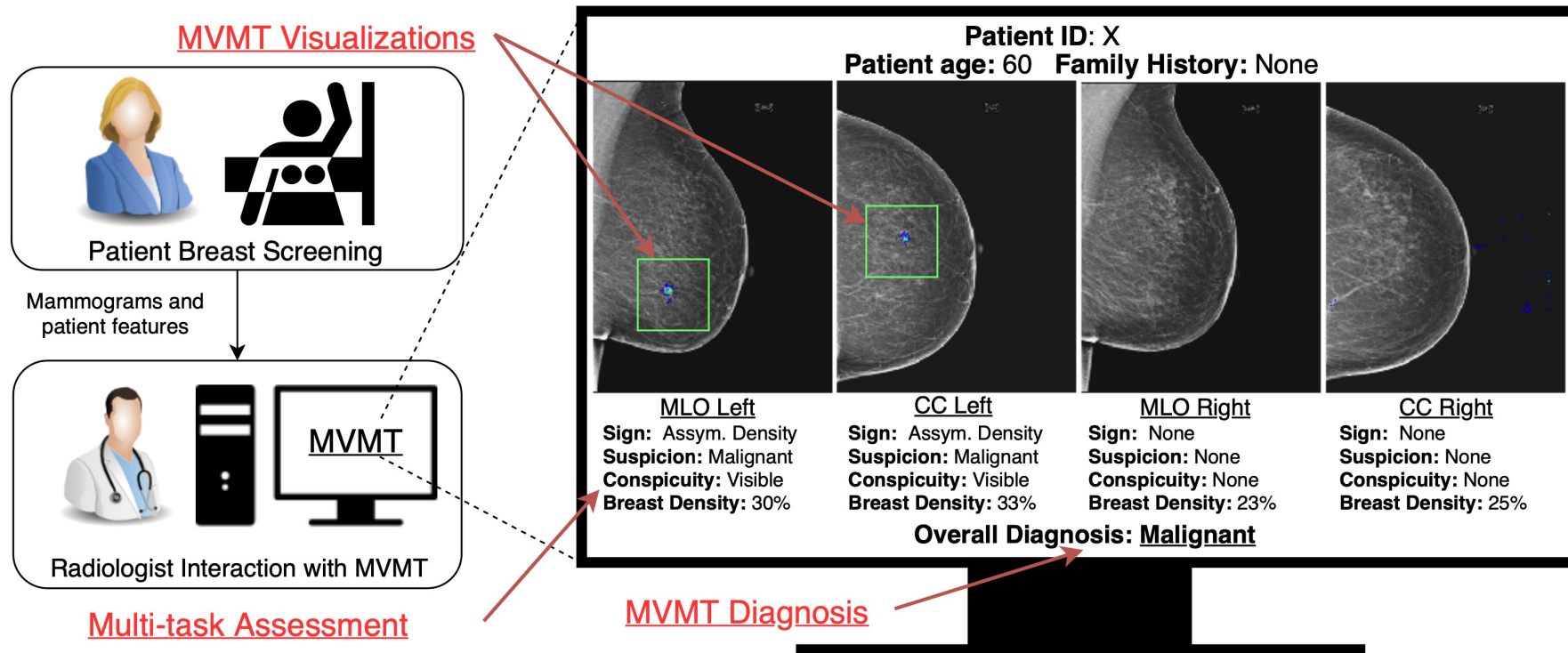
- Describing **why** models make certain decisions to experts & laypersons
- **Creating** models/architectures that are easy to explain
- *Challenge*: “Explainability” and “interpretability” are difficult to define

What is missing: the gap between correlation and causation

Most machine learning techniques, especially the statistical ones, depend highly on data correlation to make predictions and analyses. In contrast, rational humans tend to rely on clear and trustworthy causality relations obtained via logical reasoning on real and clear facts. It is one of the core goals of explainable machine learning to transition from solving problems by data correlation to solving problems by logical reasoning.

Microsoft Research Lab Asia

Explainable/Interpretable ML





An example of informativeness from “Multi-view Multi-task Learning for Improving Autonomous Mammogram Diagnosis” (Kyono 2019). Here, an ML system outputs additional features for radiologists to scrutinize along with the final diagnosis, providing valuable information to assess why certain predictions were being made. As the neural network in the model learns the same radiological features that radiologists use for diagnosis, it can be much more telling than the final output of the network itself when provided to humans.

Material from <https://blog.ml.cmu.edu/2020/08/31/6-interpretability/>

Categories of Interpretability

- **Transparency**

- *Goal*: Understand the **mechanism** by which the model works
- **Simulatability**: a human can take the inputs and go through the model's calculations in a reasonable time
- **Decomposability**: each part of the model (inputs, parameters, calculation) has an intuitive explanation 
- **Algorithmic transparency**: theoretical guarantees about the convergence or behavior of the algorithm 
- *Drawback*: Limits model selection options

Categories of Interpretability

- **Post-hoc explanations**

- *Goal:* Extract information from trained models to understand what they have learned
- **Text explanations:** An explanation-generating model is trained in tandem with the prediction model

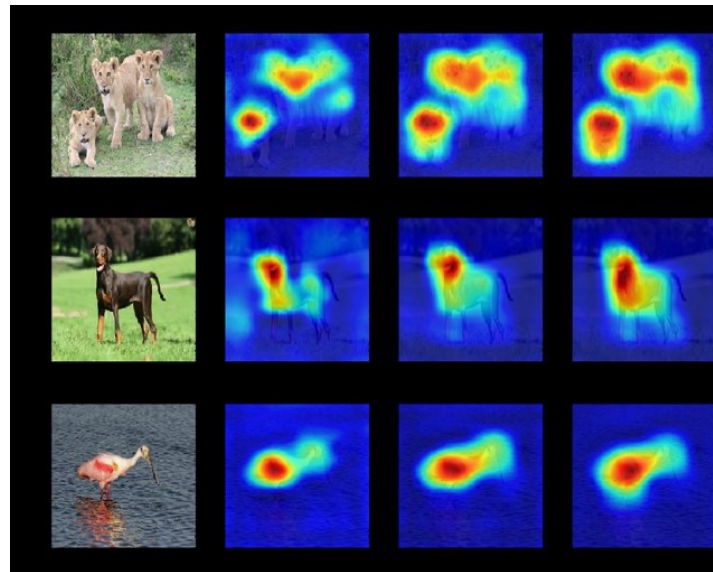


Material from <https://blog.ml.cmu.edu/2020/08/31/6-interpretability/>

Categories of Interpretability

- **Post-hoc explanations**

- *Goal:* Extract information from trained models to understand what they have learned
- **Text explanations:** An explanation-generating model is trained in tandem with the prediction model
- **Local explanations:** Visualizing "focal points" of a neural network, e.g. saliency maps



Material from <https://blog.ml.cmu.edu/2020/08/31/6-interpretability/>


Categories of Interpretability

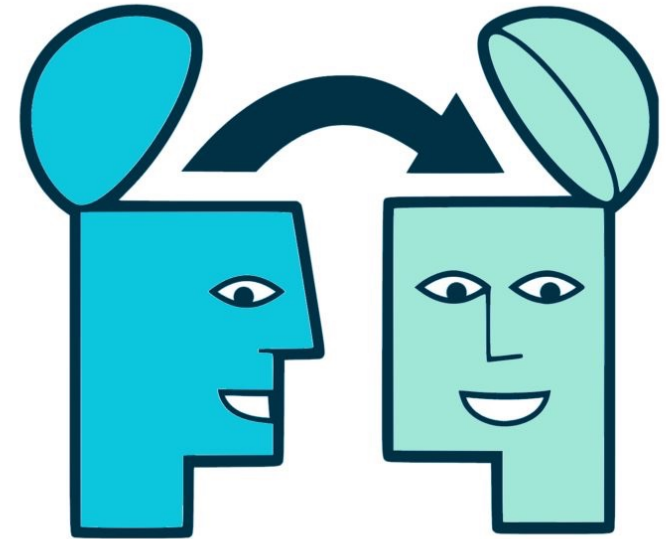
- **Post-hoc explanations**

- *Goal:* Extract information from trained models to understand what they have learned
- **Text explanations:** An explanation-generating model is trained in tandem with the prediction model
- **Local explanations:** Visualizing "focal points" of a neural network, e.g. saliency maps
- **Visualization:** Visual description of how the model has learned, e.g. using t-SNE
- **Explanation by example.** Model outputs nearest training examples in the latent representation, e.g. k-nearest neighbors.
- *Drawback:* Not guaranteed to explain the **true** underlying mechanism by which a model works → May instill false trust in the model

Transfer Learning

Transfer Learning

- Problem
 - Sufficient training data is expensive or impossible to collect for some tasks
- Solution
 - Train a model with data from a **related task**
 - **Transfer** the model (partially or entirely) to the target task and **fine-tune** with available targeted data 



Transfer Learning: NLP

- Why might transfer learning be effective for NLP tasks?
 - Structure (grammar, spelling, etc.) and content (words, punctuation) of natural language remains consistent across many tasks
- 1. Train a model to “understand” English (or other language) from vast quantities of public online documents
- 2. Fine-tune the model with relatively few examples from your specific task

Transfer Learning: GPT-3

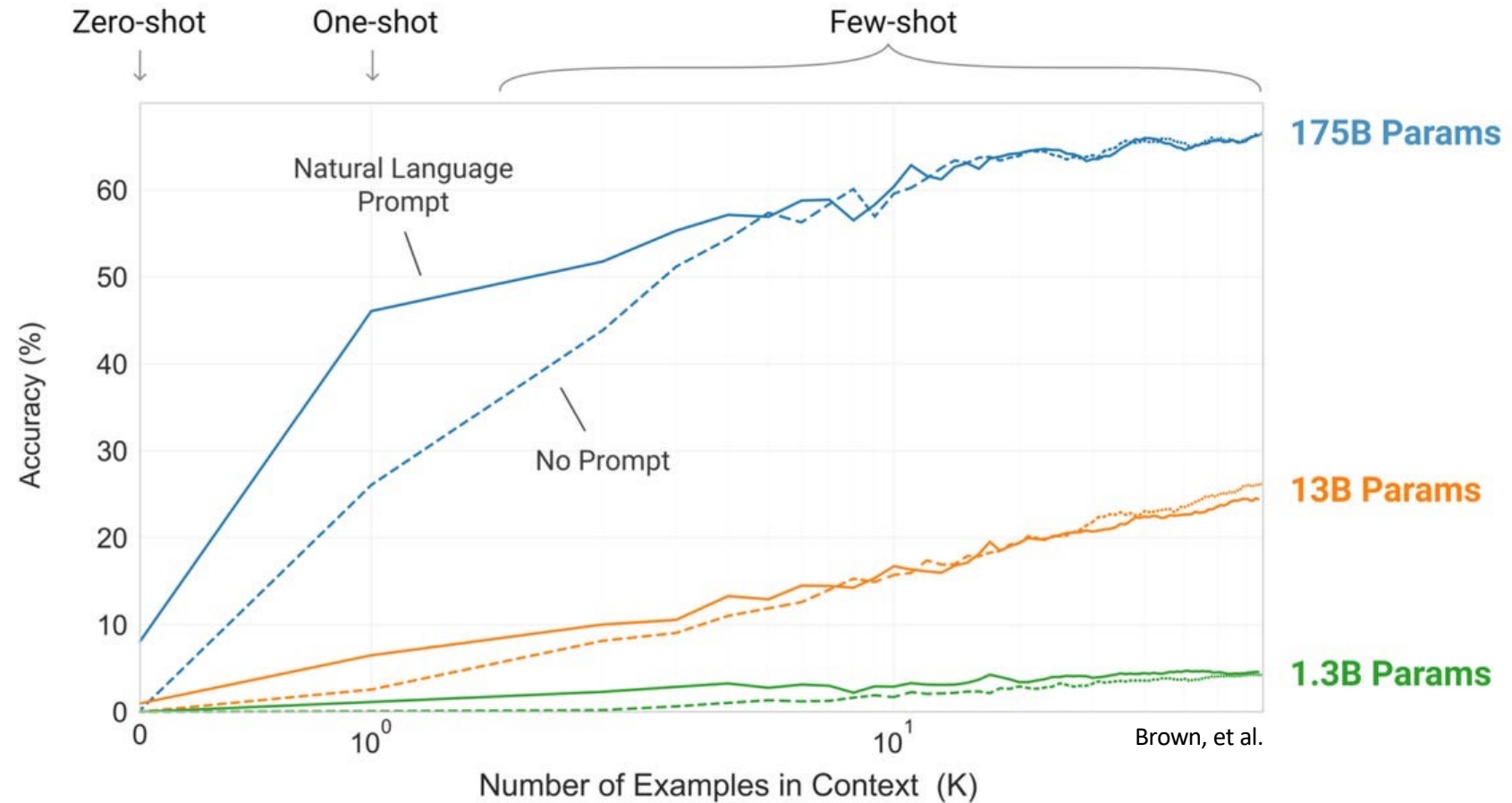


GPT-3 Training Data

Dataset	# Tokens	Weight in Training Mix
Common Crawl	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

Wikipedia

Crawl date	Size in TiB	Billions of pages
April 2021	320	3.1



Adversarial ML

Adversarial ML

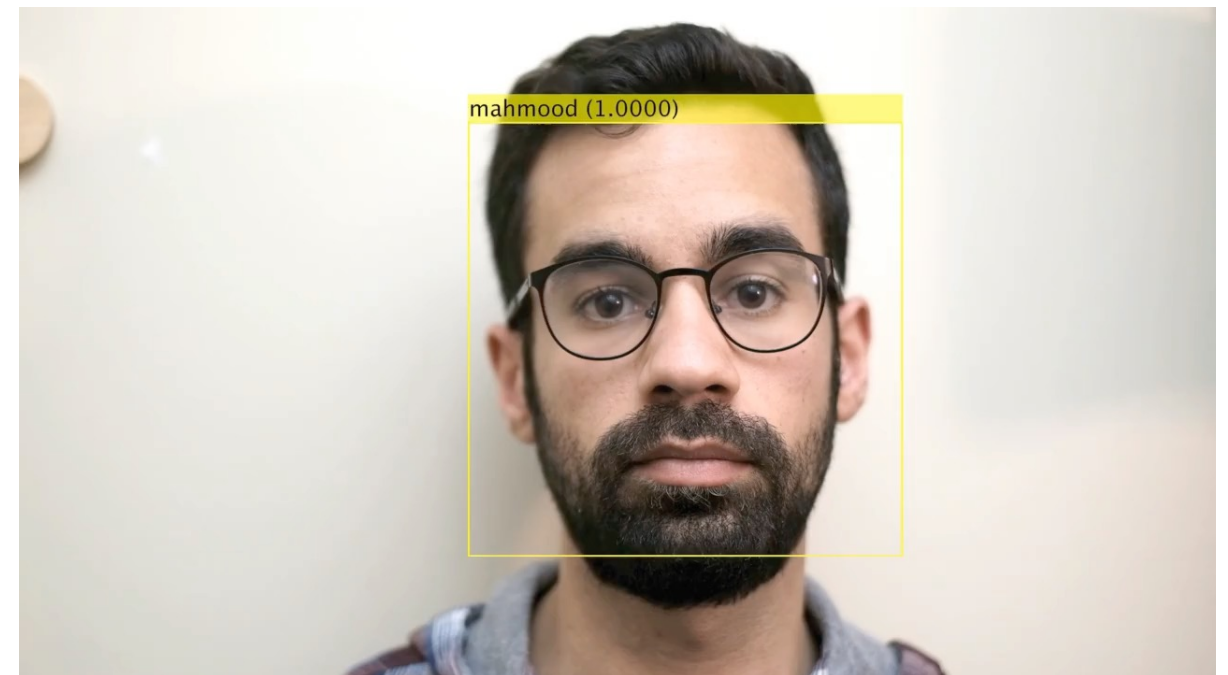
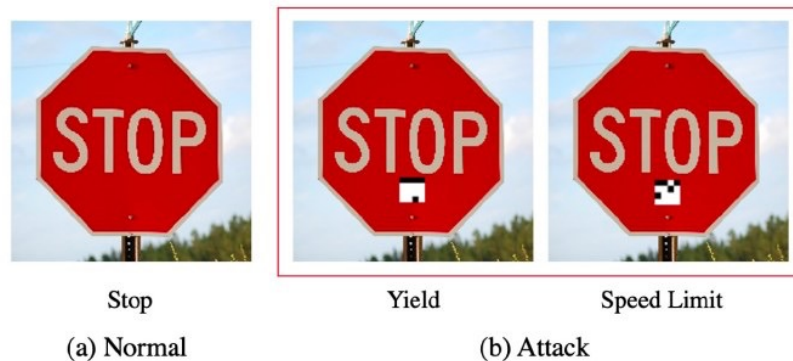
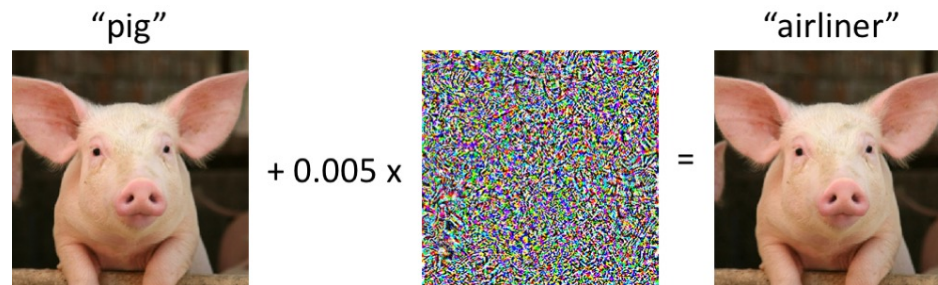
Identify/exploit facets of ML models that enable
malicious behavior

Types of Adversarial ML

- **Evasion (adversarial examples)**



- Generate **malicious inputs** that are **incorrectly classified** by target model

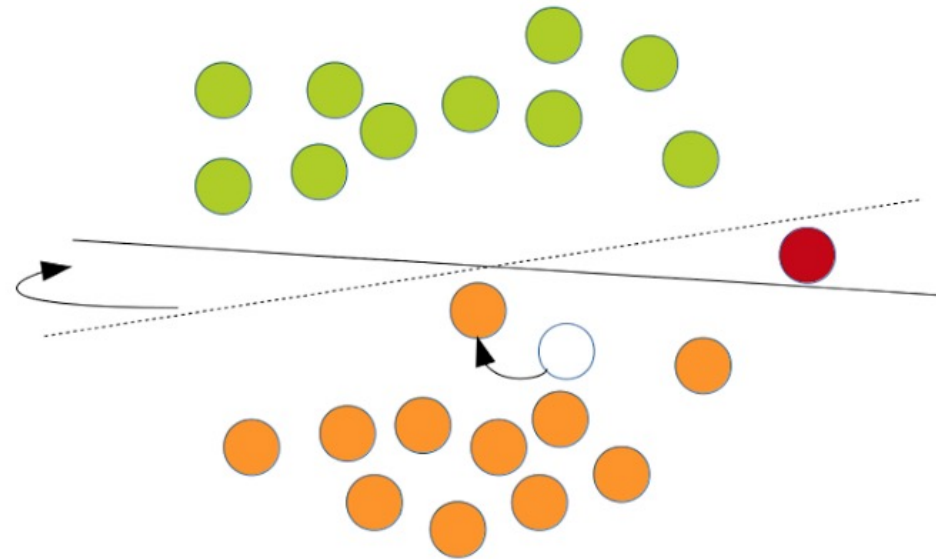


Lujo Bauer

Types of Adversarial ML

- **Poisoning**

- **Contaminate training data** to produce model with specific malicious properties

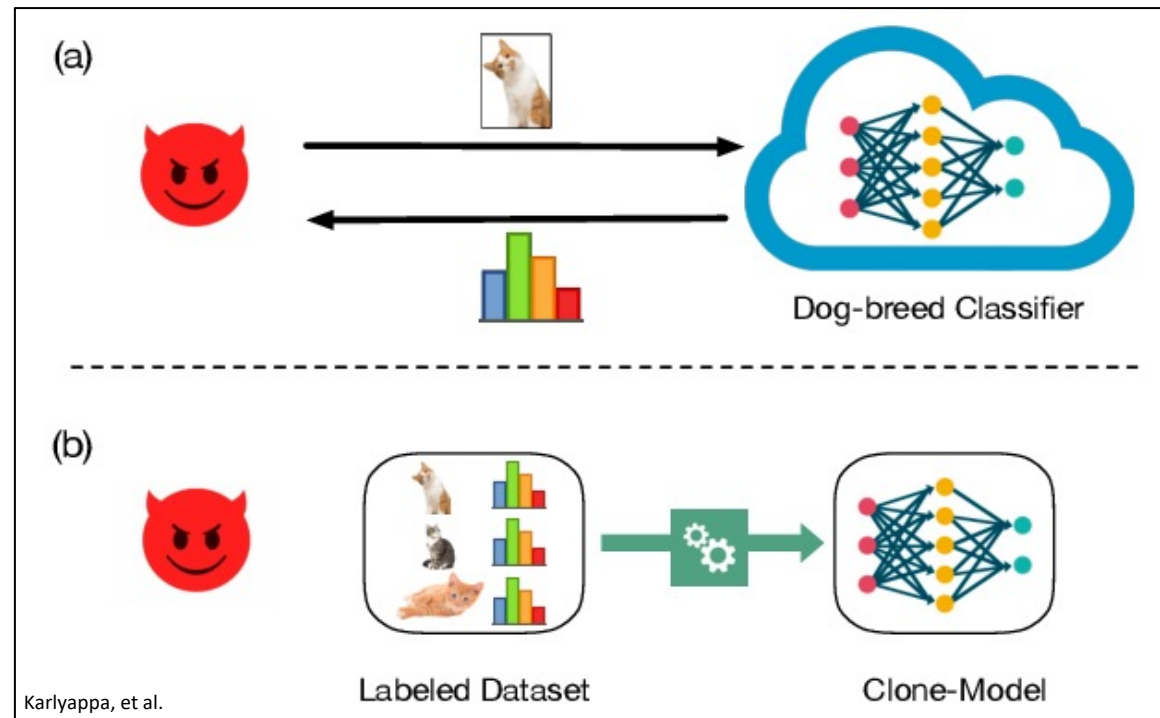


Data poisoning:
modifying training samples intelligently

Types of Adversarial ML

- **Model Theft**

- Use predictions to **reconstruct** (steal) a **proprietary model**



Types of Adversarial ML

• Membership Inference



- Use predictions to **identify private data** used to train model

2017 IEEE Symposium on Security and Privacy

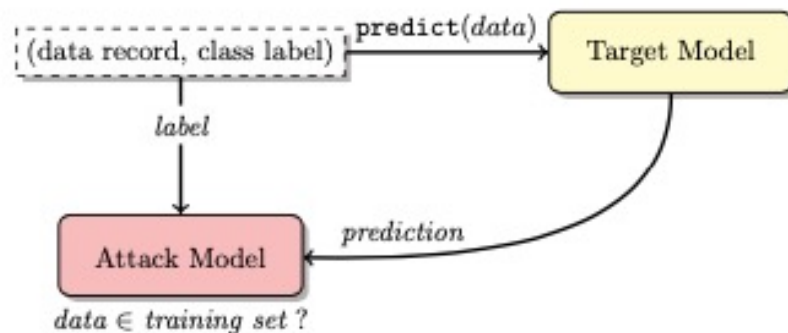
Membership Inference Attacks Against Machine Learning Models

Reza Shokri
Cornell Tech
shokri@cornell.edu

Marco Stronati*
INRIA
marco@stronati.org

Congzheng Song
Cornell
cs2296@cornell.edu

Vitaly Shmatikov
Cornell Tech
shmat@cs.cornell.edu



Abstract:

We quantitatively investigate how machine learning models leak information about the individual data records on which they were trained. We focus on the basic membership inference attack: given a data record and black-box access to a model, determine if the record was in the model's training dataset. To perform membership inference against a target model, we make adversarial use of machine learning and train our own inference model to recognize differences in the target model's predictions on the inputs that it trained on versus the inputs that it did not train on. We empirically evaluate our inference techniques on classification models trained by commercial "machine learning as a service" providers such as Google and Amazon. Using realistic datasets and classification tasks, including a hospital discharge dataset whose membership is sensitive from the privacy perspective, we show that these models can be vulnerable to membership inference attacks. We then investigate the factors that influence this leakage and evaluate mitigation strategies.

Lightweight ML

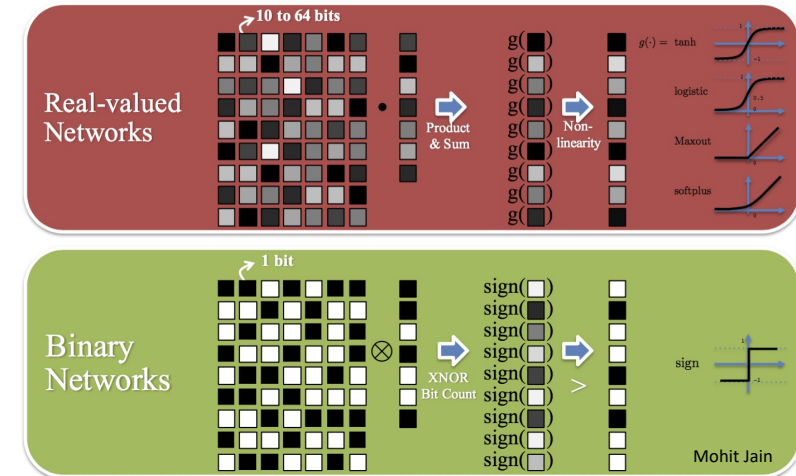
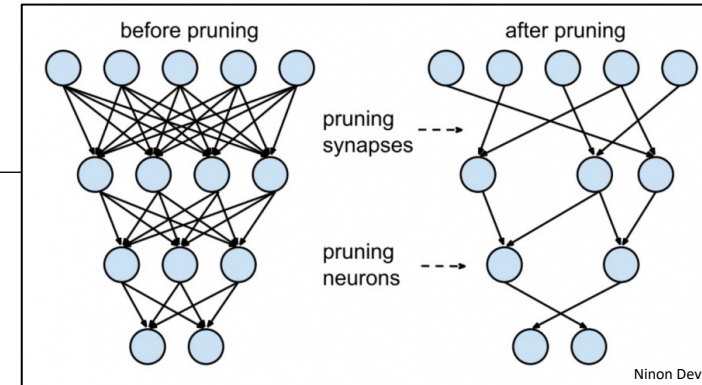
Lightweight ML

- Problem

- Training state-of-the-art ML models is too resource intensive for low-power, low-compute embedded systems

- Solution(?)

- Develop new models, training algorithms, and efficiency optimizations that enable high-accuracy ML with less overhead

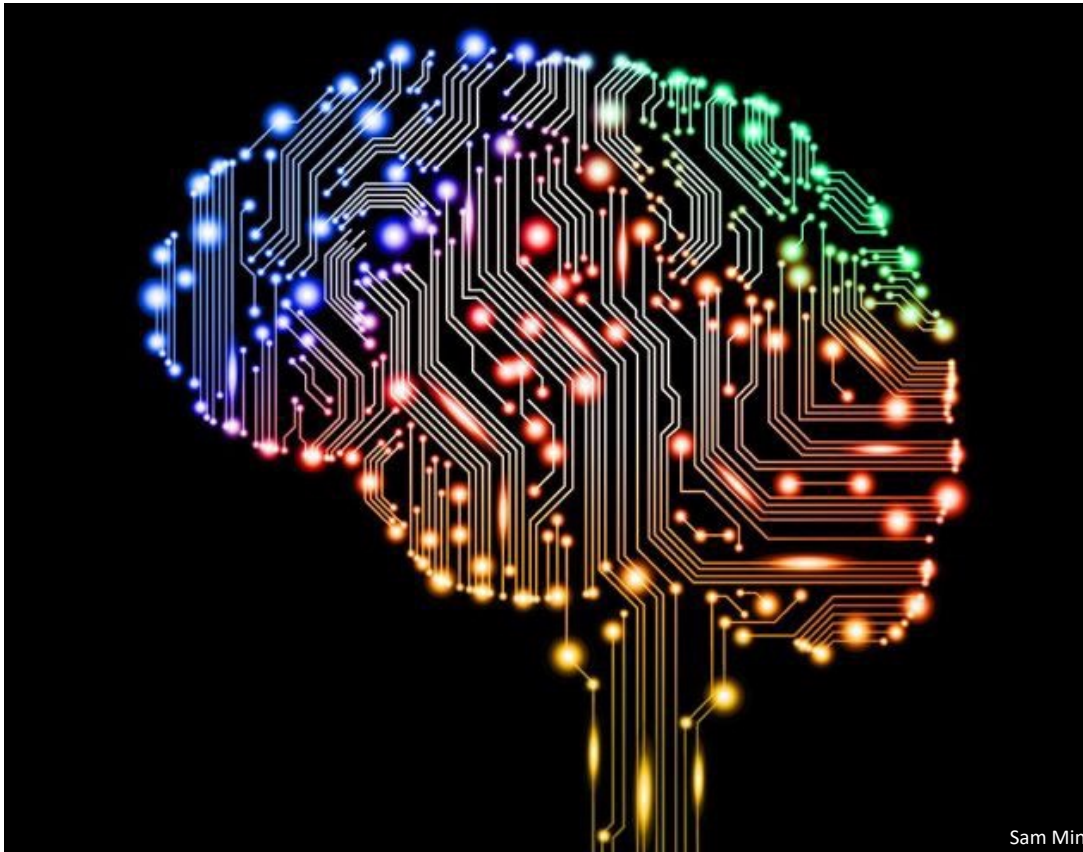


PVANET: Deep but Lightweight Neural Networks for Real-time Object Detection

Kye-Hyeon Kim,*Sanghoon Hong,*Byungseok Roh,*Yeongjae Cheon, and Minje Park
Intel Imaging and Camera Technology
21 Teheran-ro 52-gil, Gangnam-gu, Seoul 06212, Korea

Artificial General Intelligence

Artificial General Intelligence



“If we are ever to make a machine that will speak, understand or translate human languages, solve mathematical problems with imagination, practice a profession or direct an organization, either we must reduce these activities to a science so exact that we can tell a machine precisely how to go about doing them or we must develop a machine that can do things without being told precisely how” — Richard M. Friedberg, 1958