Using Model Fingerprinting for Outlier Detection

Eunji Lee, Stanford University
Mentors: Wei-Han Lee and Mudhakar Srivatsa

Model Fingerprinting: An Overview

- Machine learning models help us make predictions in many areas today
- Problem: data they are trained on will drift over time, resulting in introduction of outlier data

There is no way to collect outlier data for the purpose to train an "outlier detection model"

- Proposed Solution: create "general fingerprints" of model
- Determine whether an input is reasonable for the dataset it was trained on
- MNIST model fingerprints have 86% inlier detection and 99% outlier (FASHION-MNIST) detection

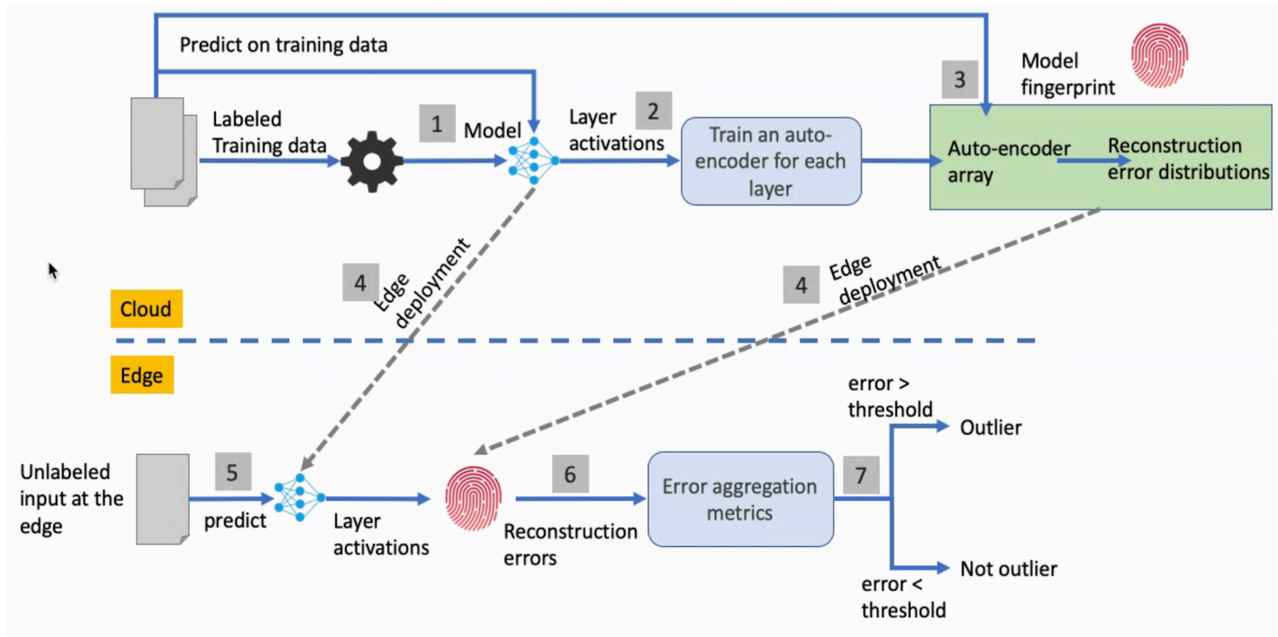

Fig. 1: Overview of the Model Fingerprinting Architecture

New Contributions: Current Progress

RGB Image Outlier Detection: CIFAR-10 and SVHN

Train model on CIFAR-10, evaluate with SVHN and vice versa:

OneOut(- log p) Accuracy	CIFAR-10	SVHN
CIFAR-10	0.94	0.97
SVHN	0.70	0.83

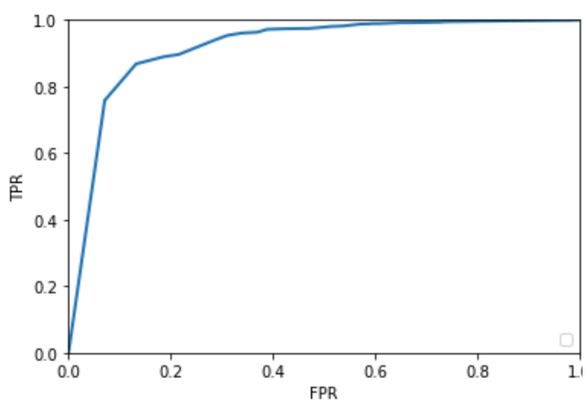

Fig. 2: AUROC for CIFAR-10 model / mixed dataset evaluation (One-Out =-0.91)

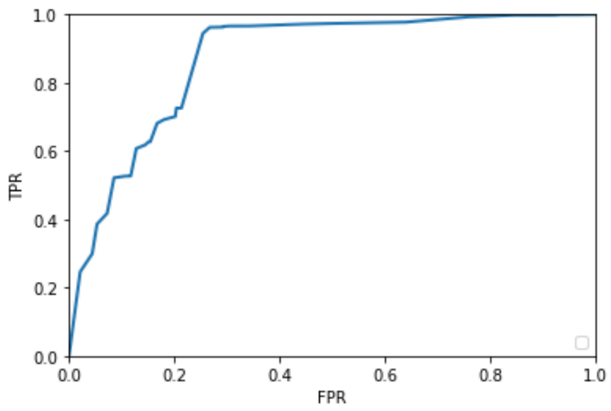

Fig. 3: AUROC for SVHN model / mixed dataset evaluation (One-Out = -0.87)

Relevant Papers
1. "A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks." Lee, K., Lee, K., Lee, H., Shin, J. 2018.
2. "Generating Adversarial Examples with Adversarial Networks." Xiao, et. al. 2019.

Adversarial Image Generation

Generate outlier images that are "similar" to inlier data, but are created using a **Deep Convolutional Generative Adversarial Network (DCGAN)**

Run model fingerprinting algorithm on adversarial dataset as outlier data and see results

Fig. 4: GAN-generated MNIST images

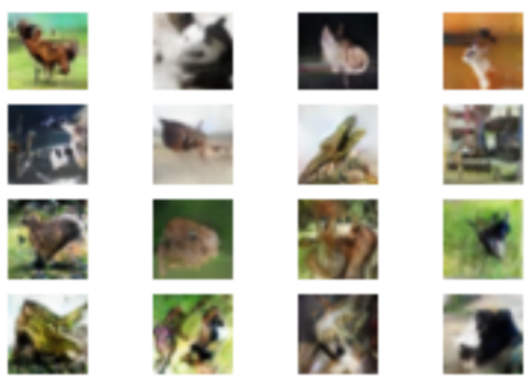

Fig. 5: GAN-generated CIFAR-10 images

AUROC for MNIST generated images is -0.65
AUROC for CIFAR-10 generated images is

Low AUROC does not necessarily indicate poor performance of model fingerprinting
Baseline performance for adversarial attacks (see **Short-Term Goals**)

Future Work

Short-Term Goals

GAN outlier dataset is not truly "malicious": has access to entire inlier dataset

- Not realistic for an adversarial attack
- Fast-Gradient Signed Method (FGSM) and DeepFool more likely to mimic malicious attacks
- Improve model fingerprinting algorithm to detect these attacks, based on performance

Long-Term Vision

Building a search engine for machine learning models

- Will help us make use of data "on the edge"
- Goal: crawl and index ML models, search and rank ML models for a certain dataset
- Outlier detection helps determine whether input data is similar enough to the dataset that a certain model was trained on

Github Link: https://github.ibm.com/IoTDL/ModelFingerprinting/tree/eunji

Research Intern Poster Session
Eunji Lee
eunji.lee@ibm.com