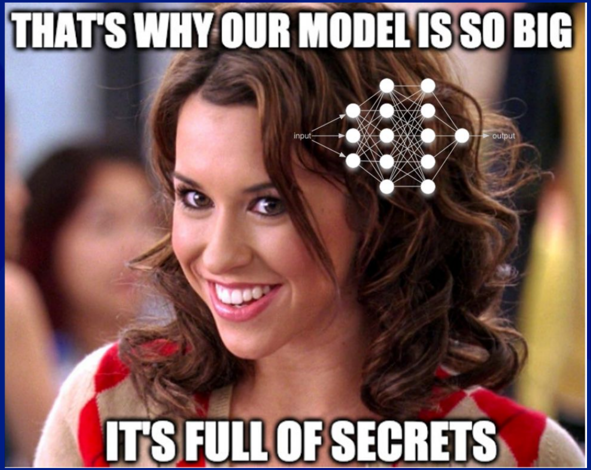
Los Alamos
NATIONAL LABORATORY

Explanations as Defense: Detecting Adversarial Inputs to Machine Learning Models

PI: **Elisabeth (Lissa) Moore, CCS-3**
Julia Nakhleh, postbac student, CCS-7
Emma Drobina, summer grad student, A-1
Subhashis Hazarika, postdoc, CCS-3 (former)

9/13/21

LA-UR-21-28972



Motivation & Goal

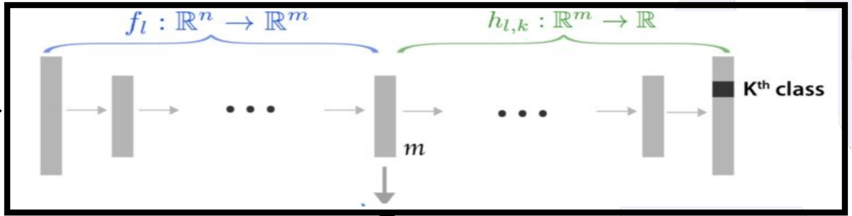
- Investigate relationship between explanations and adversarial attacks
  - Specifically, focus on concept-based explanations rather than feature attribution
- Intuition: If adversarial attacks are invisible to humans, they should not be changing concepts related to the definition of the true class
- Most recent hypotheses argue that adversarial attacks are just noise
  - i.e., not semantic

Concept-based Explanations

- Use meaningful high-level concepts rather than independent feature attribution

Input x

Trained Model (Resnet)



Output y

Lemon

Yellow Concept

$f($ $)$ $f($

$)$

Linear decision boundary

Secondary Model

Explanation

Resnet used:

Yellow Concept 95%

Experimental Setup

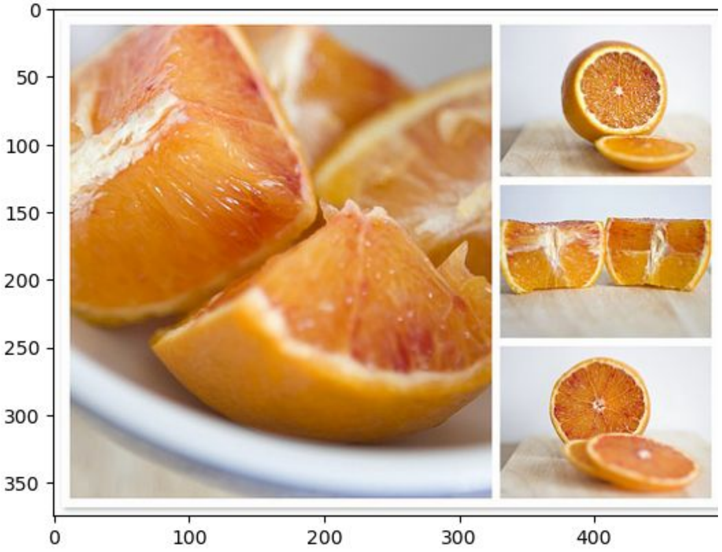
- Generate adversarial examples for Resnet with Imagenet data:
  - Fast Gradient Sign, Projected Gradient Descent, Carlini-Wagner, Momentum Iterative

- Build concept discrimination models for concepts relating to common adversarial output classes

- Compare concept activations between pre-attack images, attacked images, and true images of the targeted class

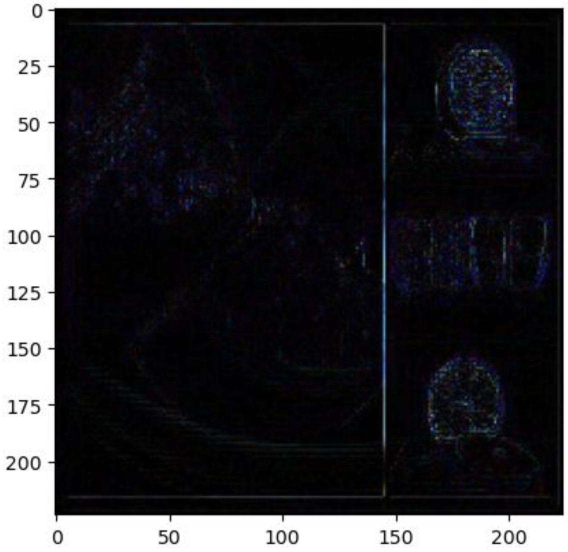
Example attacks

Pre-attack Image

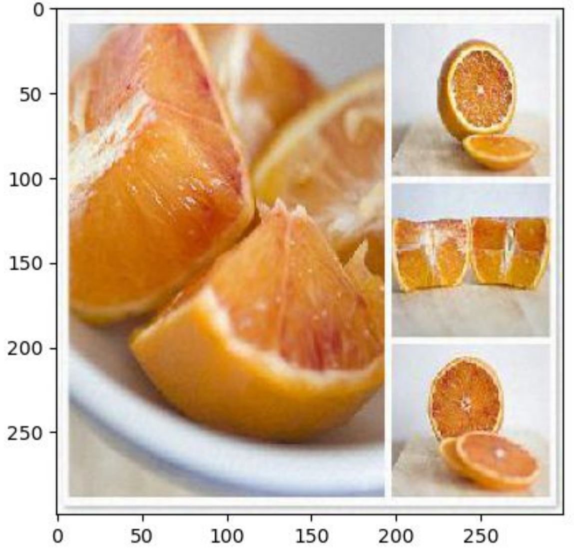


True label: Orange
Predicted: Orange

Attack



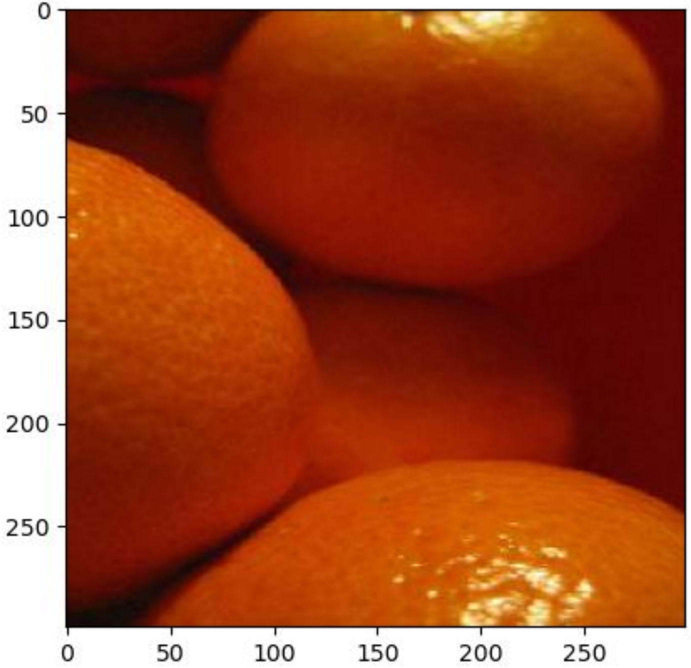
Attacked Image



True label: Orange
Predicted: Lemon

Example attacks

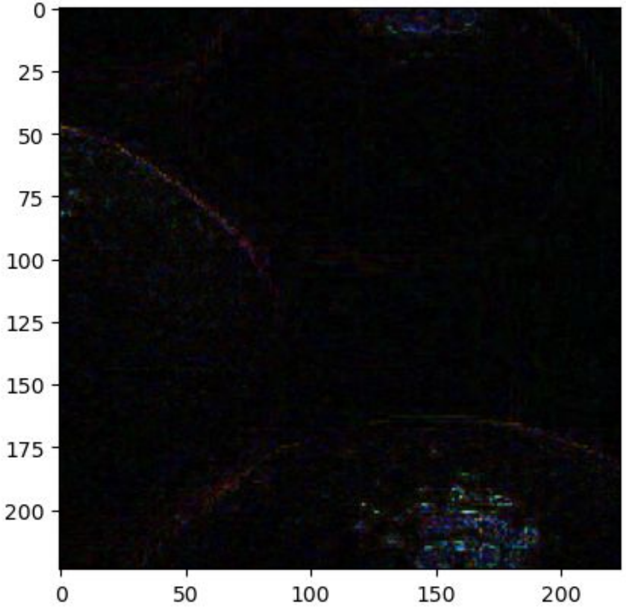
Pre-attack Image



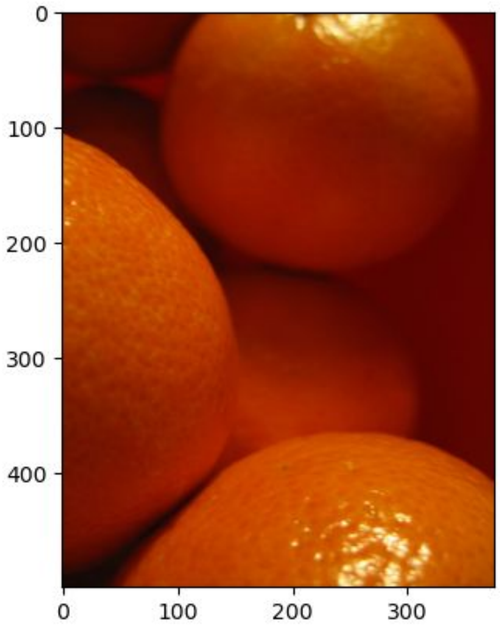
True label: Orange

Predicted: Orange

Attack



Attacked Image

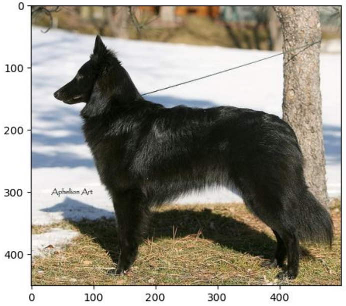


True label: Orange

Predicted: Lemon

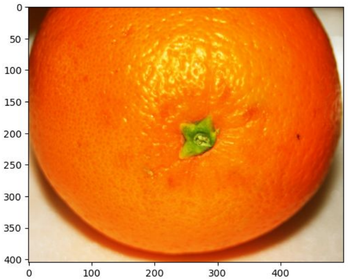
Example Attack Directions

Groenendael

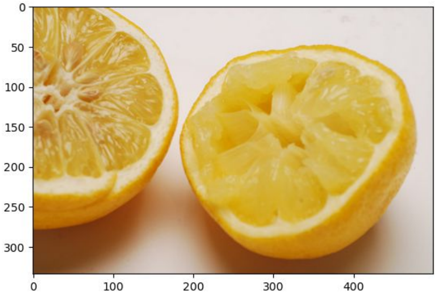


Schipperke

Orange



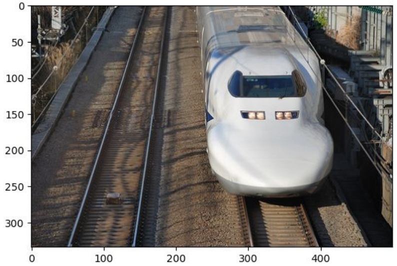
Lemon



Studio_couch

Quilt

Bullet_train

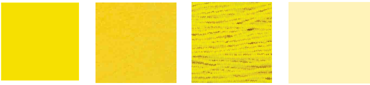


Sports_Car

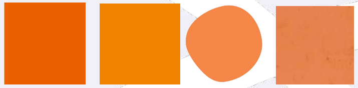
Los Alamos
NATIONAL LABORATORY

Per-layer Concept Discrimination Models

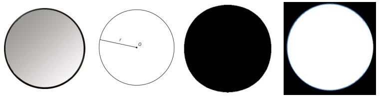
Yellow

network_activations()

Orange

network_activations()

Circle

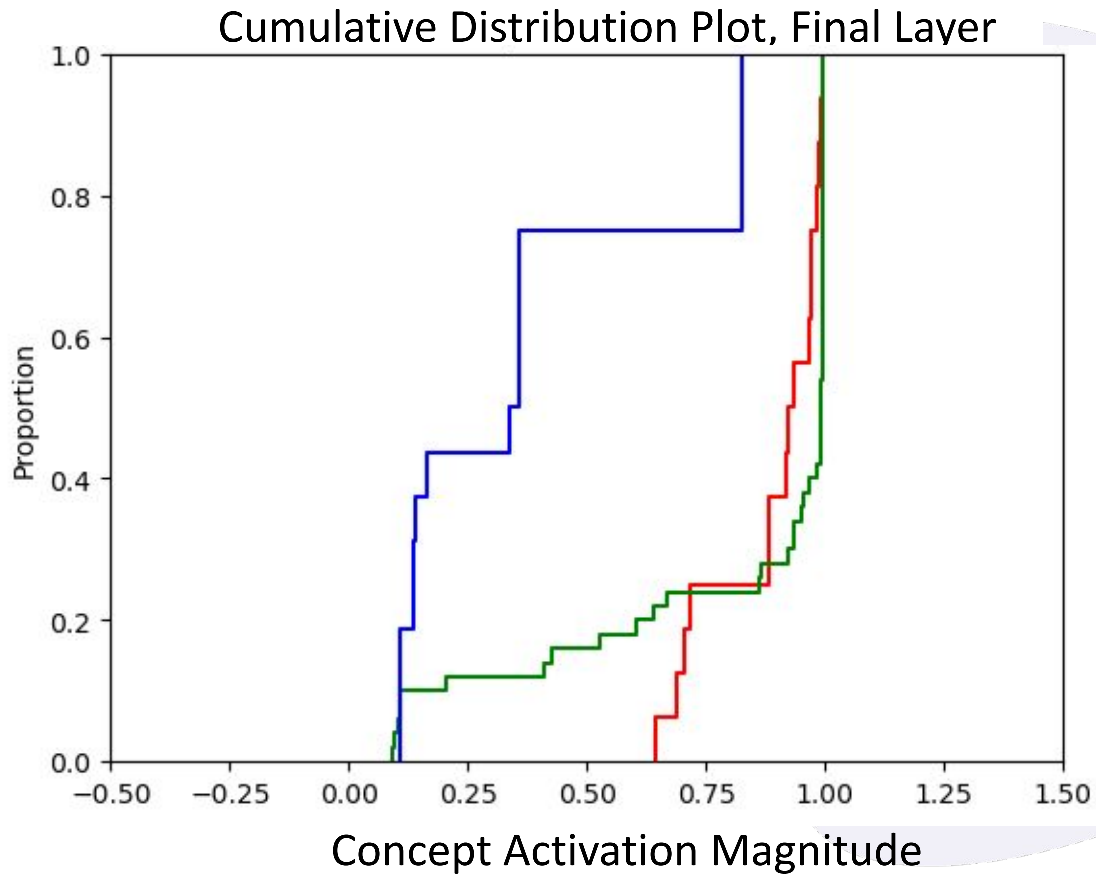
network_activations()

Spheroid

network_activations()

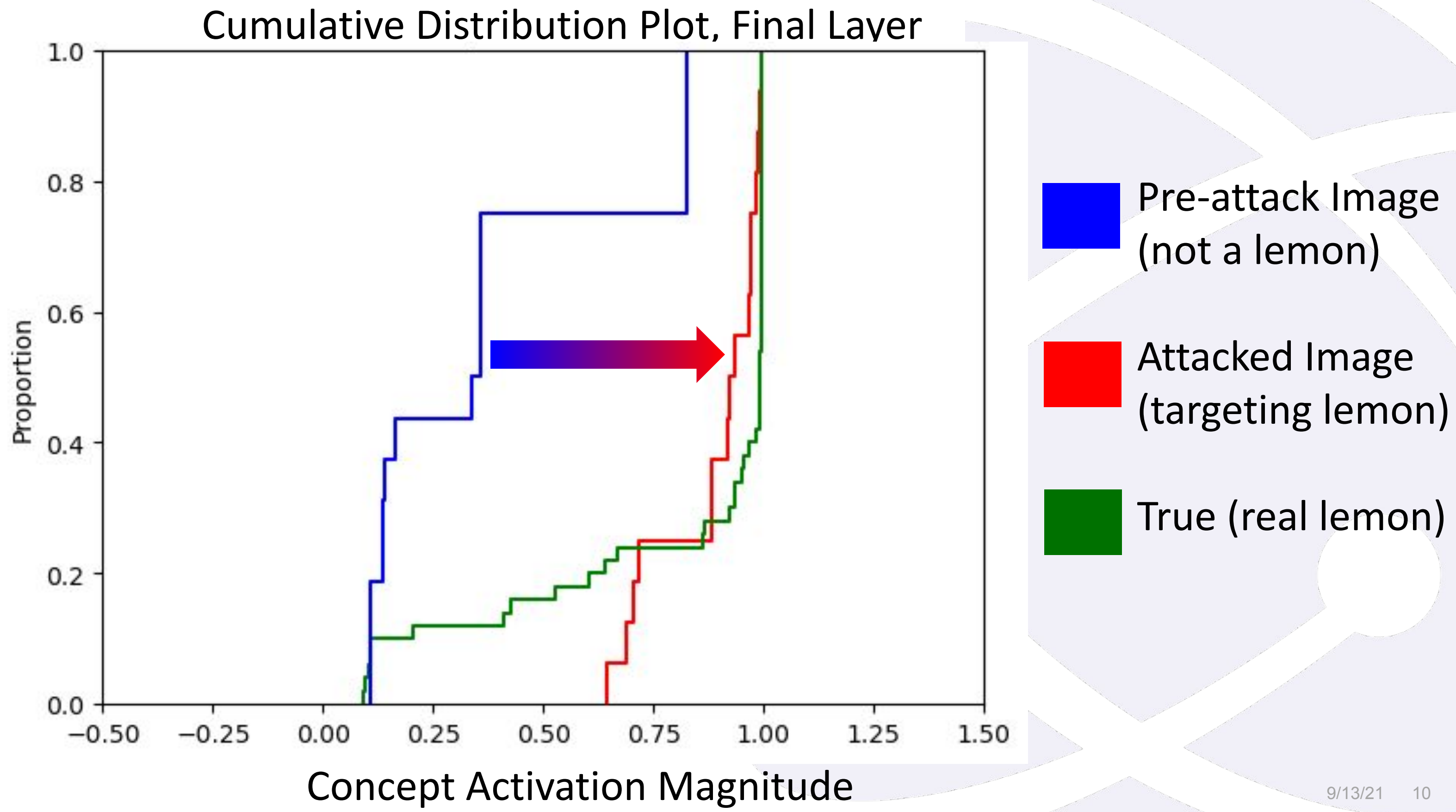
Train linear model (SVM with linear kernel) to separate concepts from random images.
One linear model trained for each (concept, network layer) pair.

Results - “Yellow” Concept

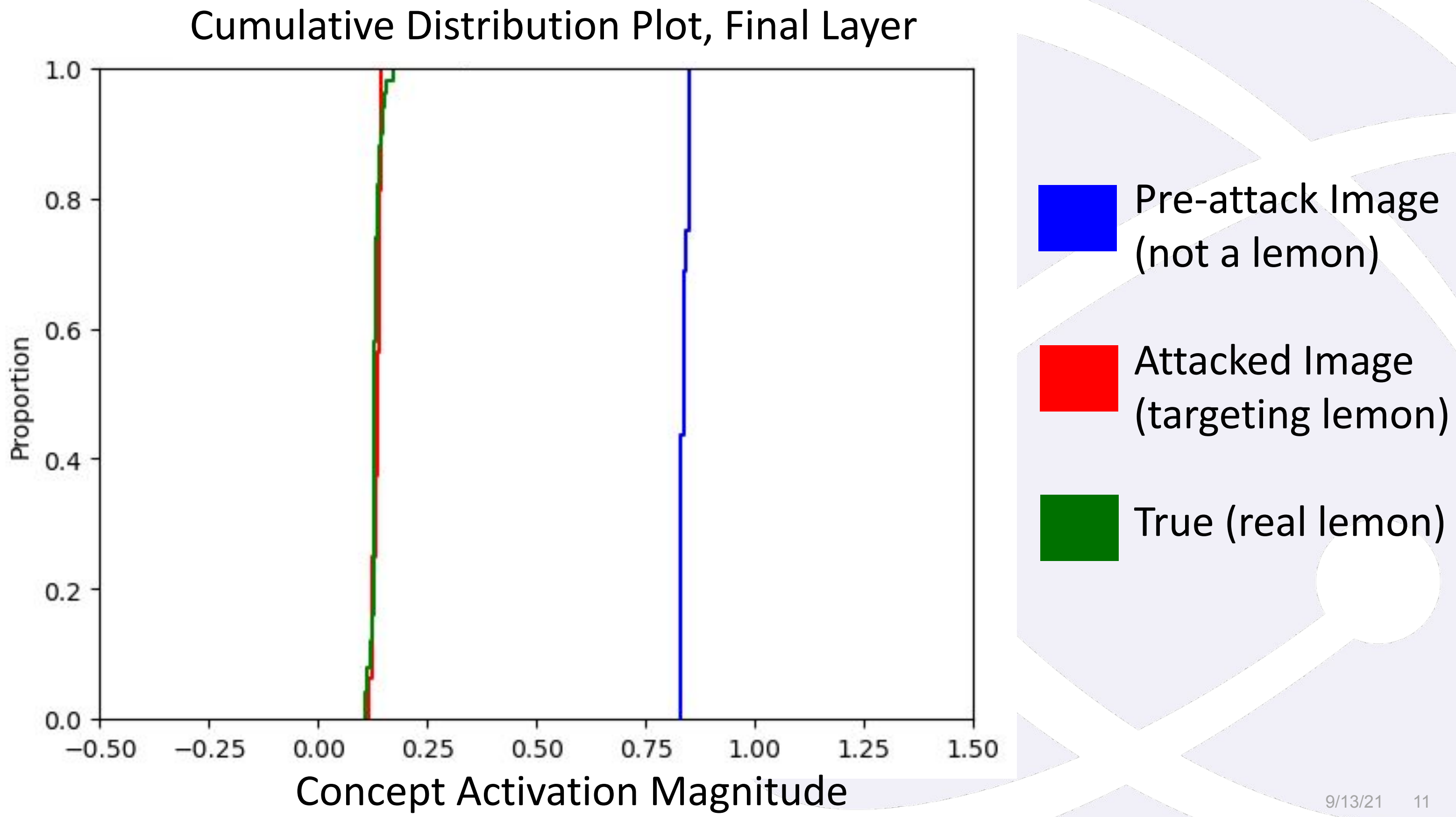


- Pre-attack Image (not a lemon)
- Attacked Image (targeting lemon)
- True (real lemon)

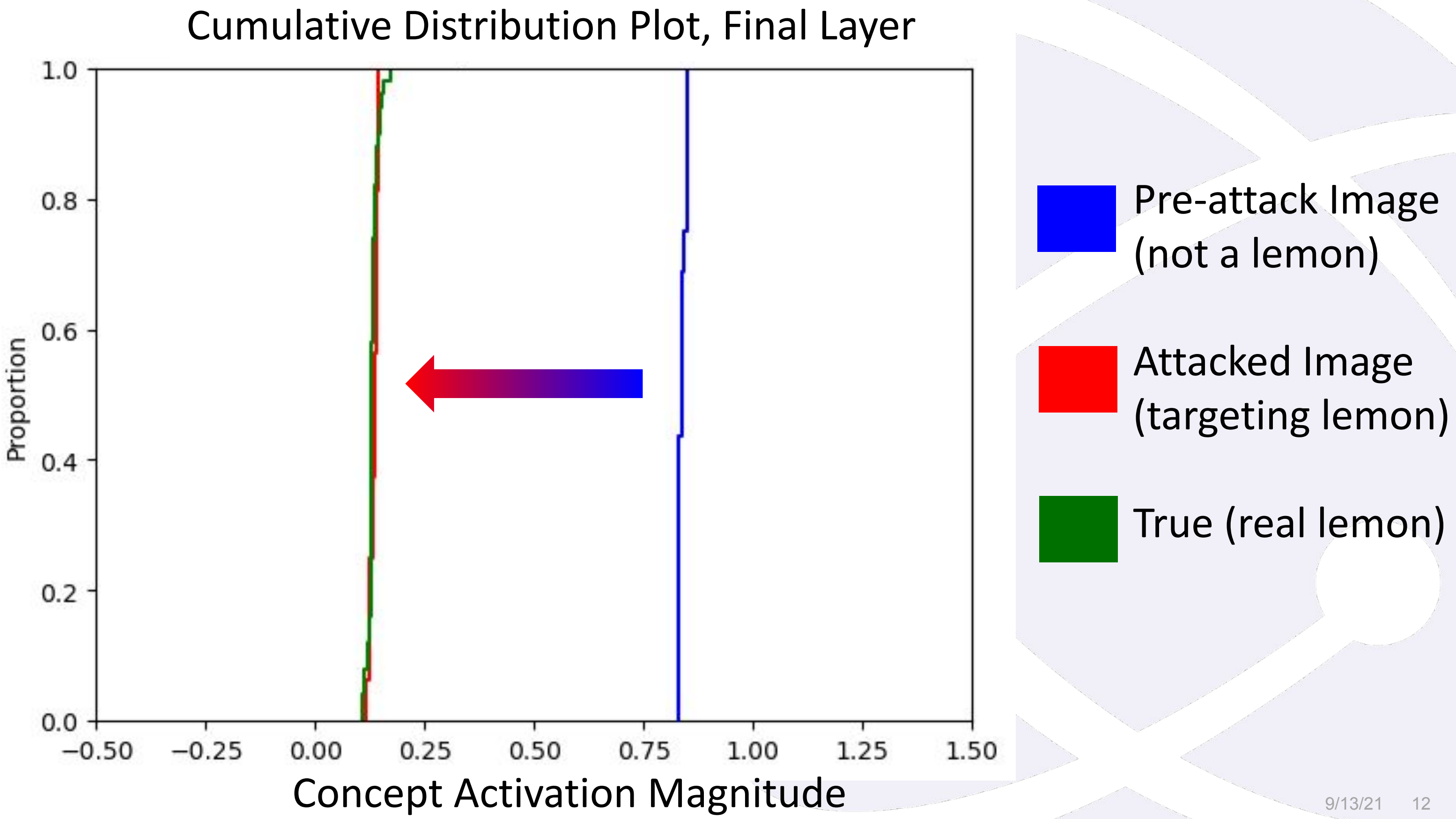
Results - "Yellow" Concept



Results - “Orange” Concept



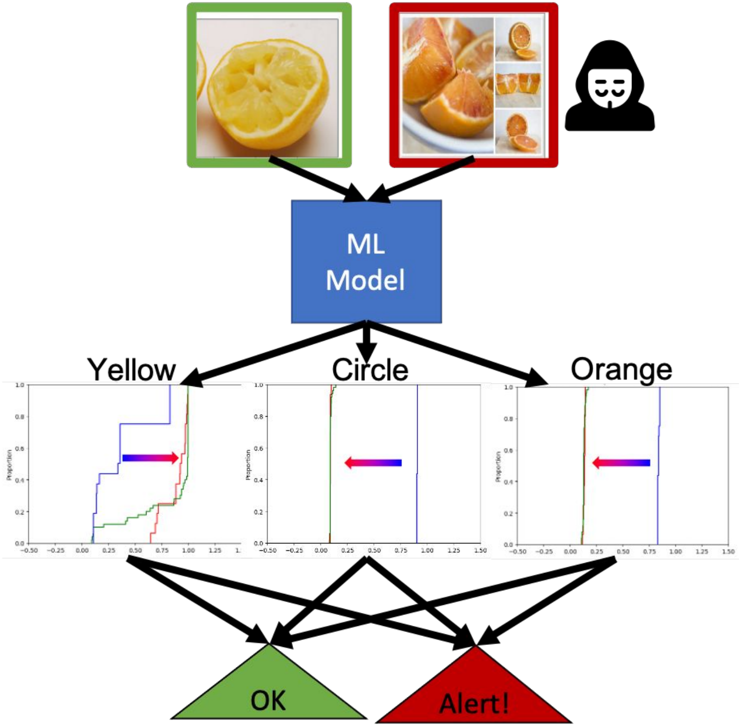
Results - "Orange" Concept



Discussion + Future Work

- Results suggest that adversarial attacks on non-robust models may be semantic (and therefore harder to detect)
- Re-run experiments with targeted attacks
  - Untargeted attacks on Imagenet tend to flip to semantically similar classes
- Run similar analysis on more robust models
  - We believe attacks may be less semantic and therefore easier to detect
- Anomaly detection via p-value fusion across network layers and concepts
- *Note: Also need for rigorous evaluation and reproducibility of explanations*

Explanations as Defense:
Detecting Adversarial Inputs to Machine Learning Models



We investigate the relationship between adversarial attacks and explainable machine learning. Concept-based explanation techniques, rather than feature attribution-based techniques, can elucidate aspects of the input data affected by untargeted adversarial attacks.

Los Alamos NATIONAL LABORATORY

Project Description

We investigate the relationship between state-of-the-art explainable machine learning techniques and adversarial attacks, particularly with respect to leveraging explanations for defense.

Project Outcomes

- *Concept-based explanation techniques can highlight aspects of data affected by attacks*
- *Untargeted attacks, regardless of type of attack, appear to be* ***more semantically meaningful*** *than previously thought.*
- *Future work: full characterization of relationship between explanations and attack types.*

PI: Elisabeth (Lissa) Moore, CCS-3, lissa@lanl.gov
Total Project Budget: $45k
ISTI Focus Area: Computational Integrity