How Secure are ML Applications?

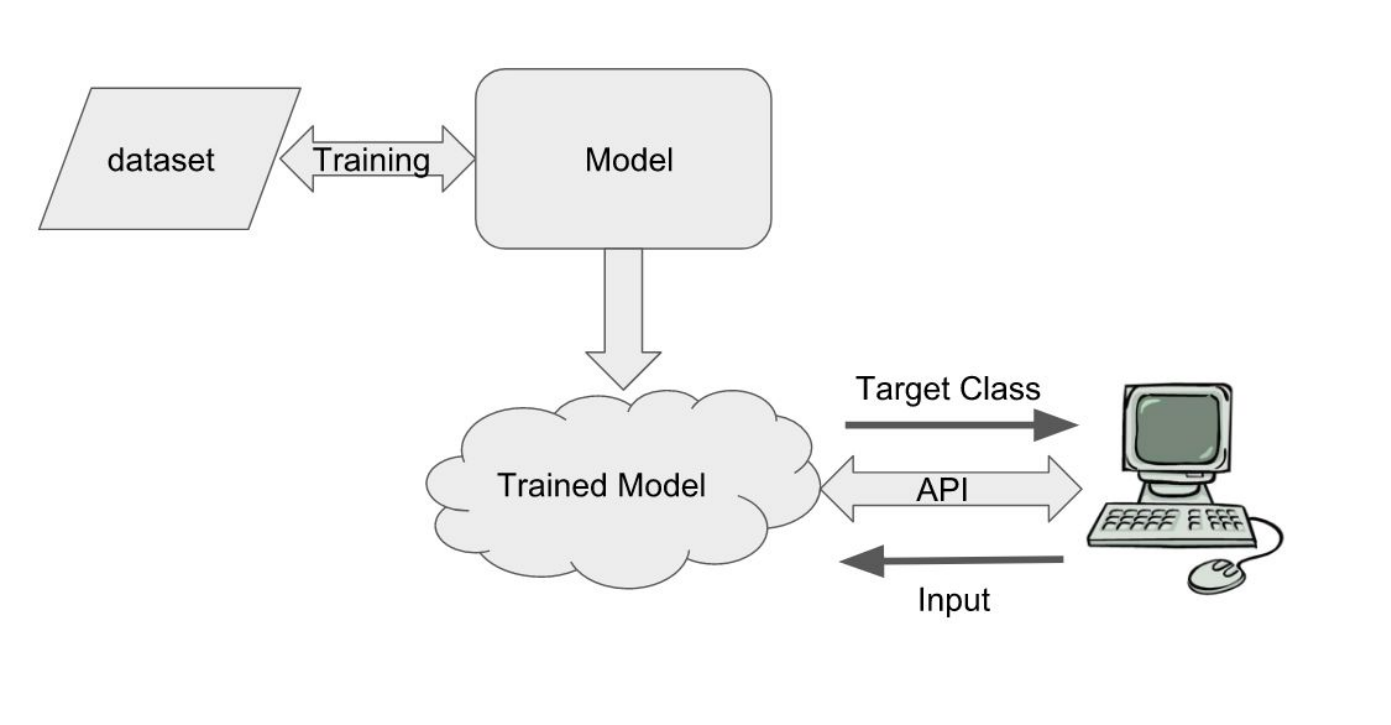
Agenda

- How ML applications work
- Adversarial Learning Attack
- Model Stealing Attack
- Model Skewing Attack
- Model Inversion Attack
- Possible Mitigation (if any...)
- Strategic way to test ML applications

Applications in security

- WAF
- IDS/IPS
- Malware detection
- Antiviruses
- Spam filters
- ...

How it works in production



Adversarial Learning Attacks

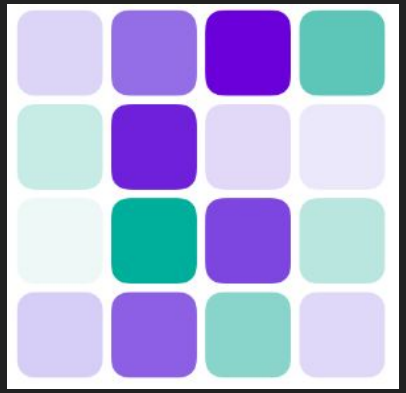
Specially craft an input with an intention to produce desired prediction from target model.

Street Sign

Perturbations

Mail Box

Adversarial Learning Attacks



Foolbox

Ref: https://arxiv.org/abs/1710.08864

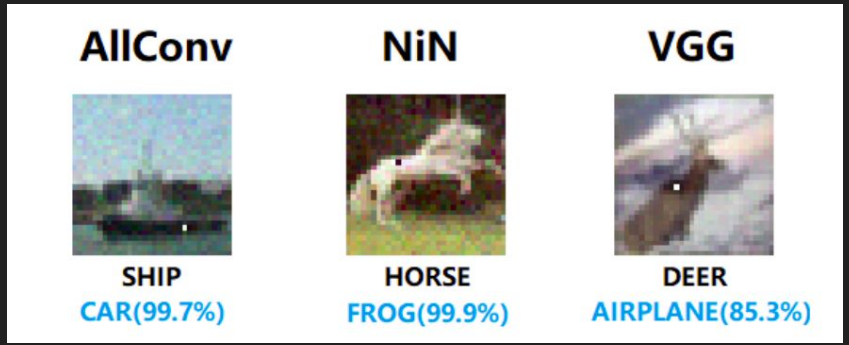
Cleverhans

Ref: https://arxiv.org/abs/1710.08864

Adversarial Attacks Examples

One Pixel Attack for Fooling Deep Neural Networks

Jiawei Su*, Danilo Vasconcellos Vargas* and Kouichi Sakurai



Ref: https://arxiv.org/abs/1710.08864

Adversarial Attacks: Face Recognition

A General Framework for Adversarial Examples with Objectives

Fig. 5. An example of digital dodging. Left: An image of actor Owen Wilson (from the PubFig dataset [40]), correctly classified by VGG143 with probability 1.00. Right: Dodging against VGG143 using AGN's output (probability assigned to the correct class <0.01).

Ref: https://arxiv.org/abs/1710.08864

Adversarial Attacks: Object Detection

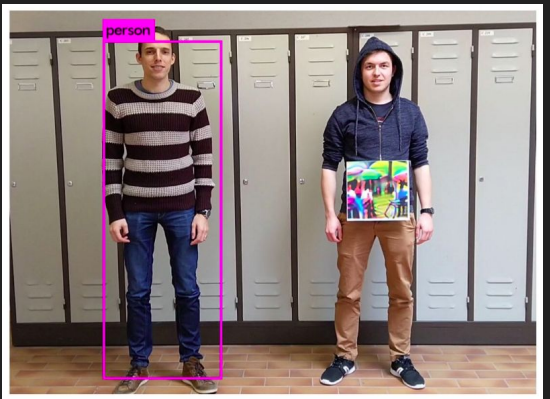
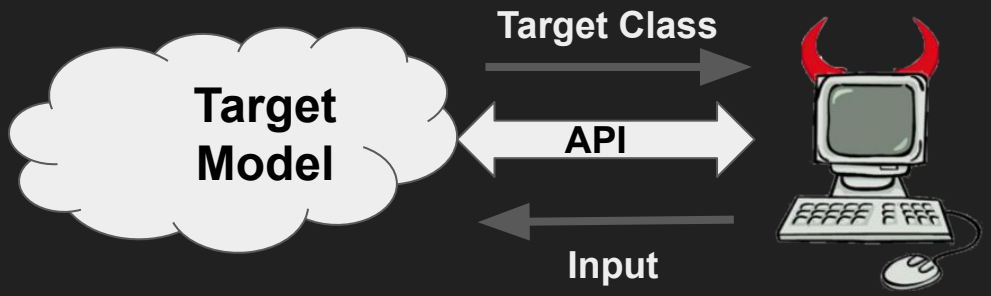


Figure 1: We create an adversarial patch that is successfully able to hide persons from a person detector. Left: The person without a patch is successfully detected. Right: The person holding the patch is ignored.

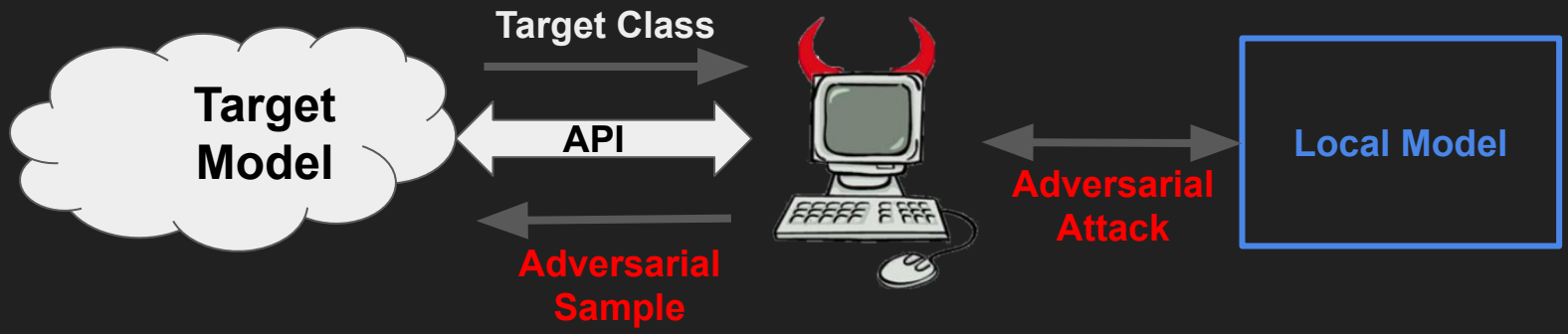
Fooling automated surveillance cameras: adversarial patches to attack person detection

Ref: https://arxiv.org/pdf/1904.08653.pdf

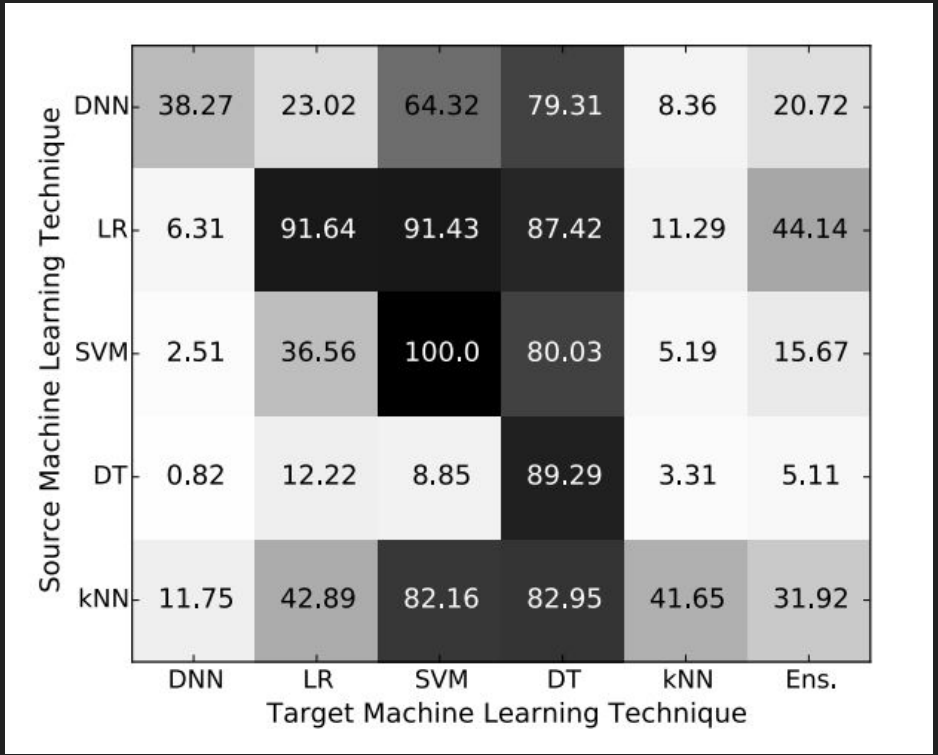
Adversarial Attacks on Black-Box Models



Adversarial Attacks on Black-Box Models



Adversarial Attacks are Transferable



Ref: https://arxiv.org/pdf/1605.07277.pdf

Adversarial Attacks Examples

https://www.cpomagazine.com/cyber-security/researchers-trick-cylance-antivirus-into-thinking-malware-is-trusted-software/

How do we deal with copying content, or the first adversarial attack in prod

Avito company blog , Programming , algorithms , Image processing , Machine learning

https://habr.com/ru/company/avito/blog/452142/

LIME for Model Interpretability

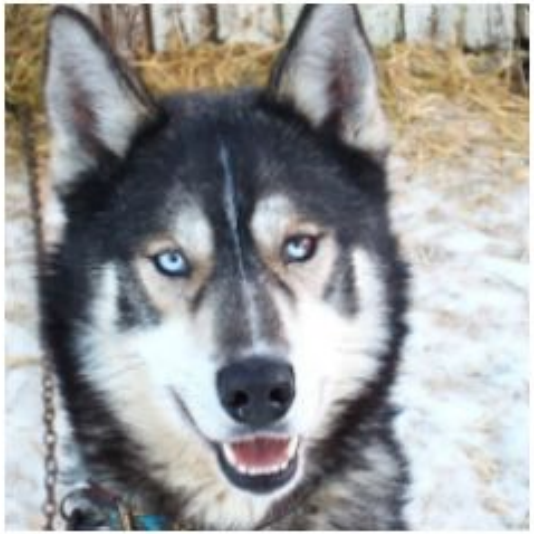
"Why Should I Trust You?"
Explaining the Predictions of Any Classifier

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Ref: https://arxiv.org/pdf/1602.04938.pdf

Explaining Predictions



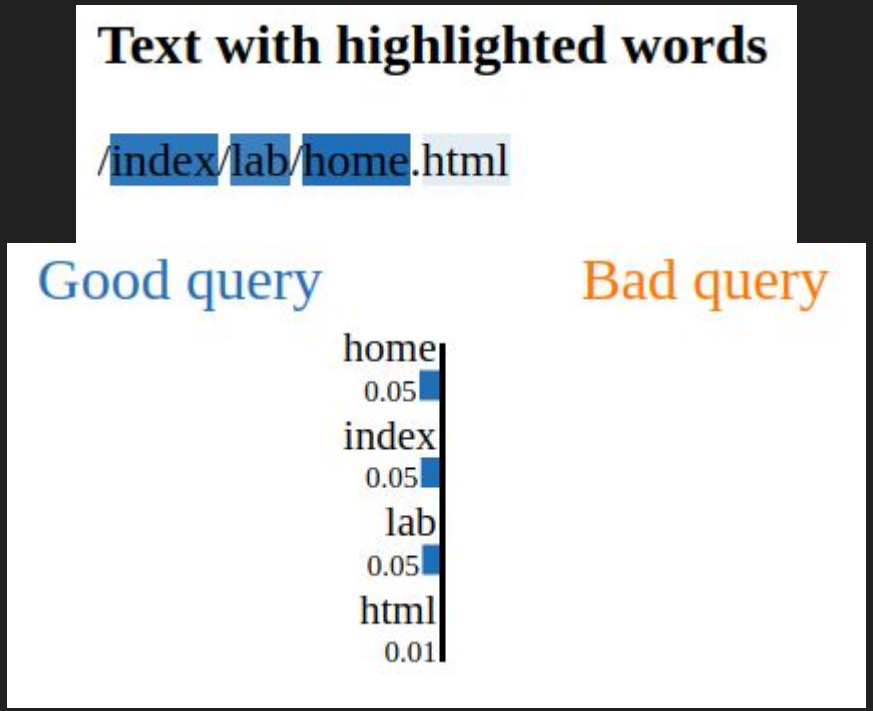
(a) Husky classified as wolf

(b) Explanation

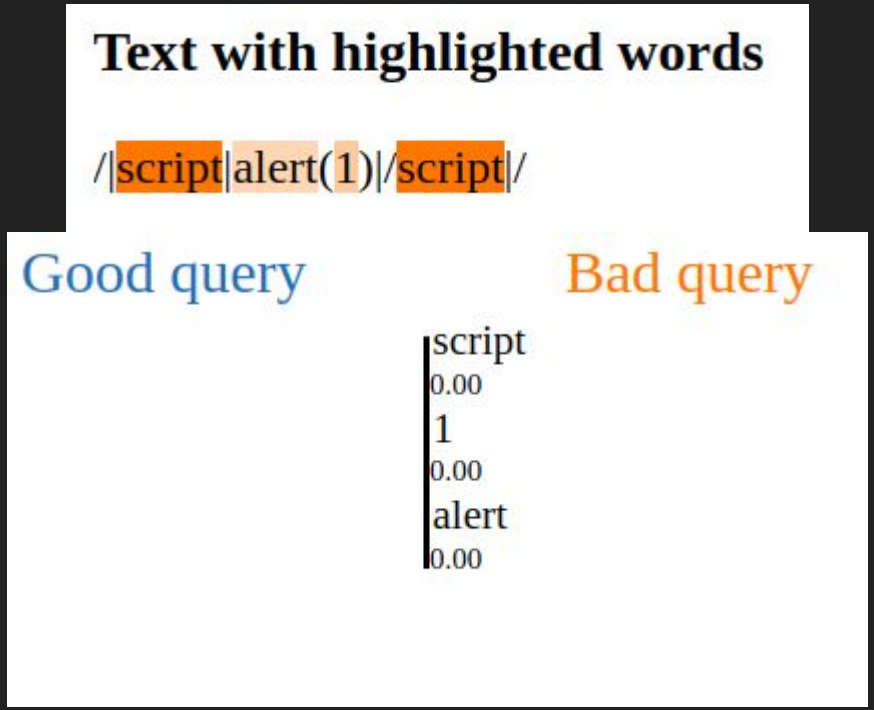
Ref: https://arxiv.org/pdf/1602.04938.pdf

LIME to bypass WAF

query: /index/lab/home.html



query: /<script>alert(1)</script>/



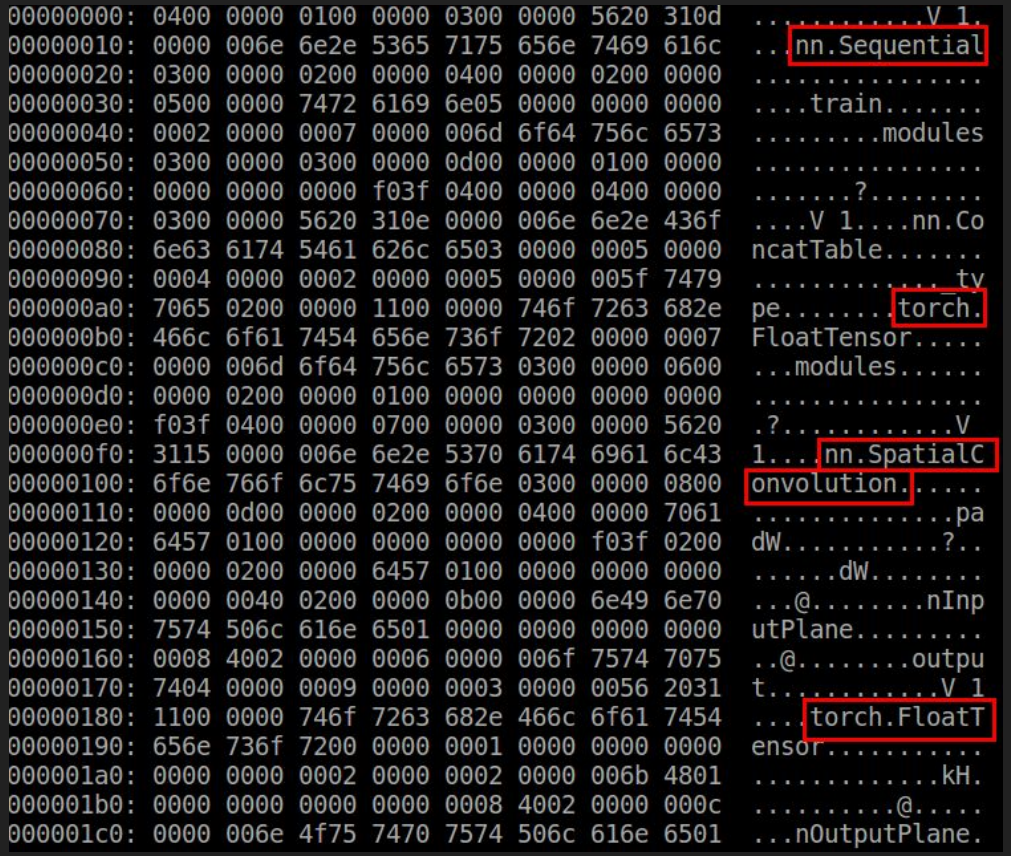
Model Stealing Attacks

- Offline Attacks
  - Stealing Locally deployed models
- Online Attacks
  - Stealing models deployed on cloud with black box access

Offline Model Stealing

- Analyse the serialized model
- Understand used Frameworks
- Leverage these frameworks to get predictions from model

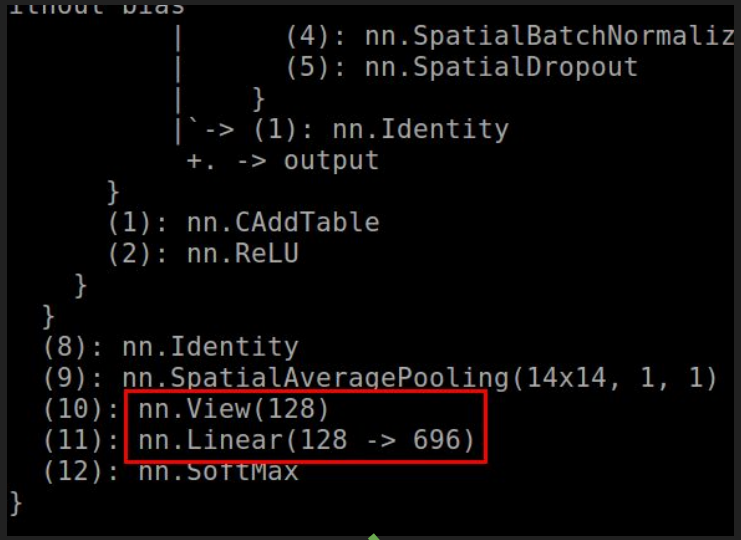
Offline Model Stealing

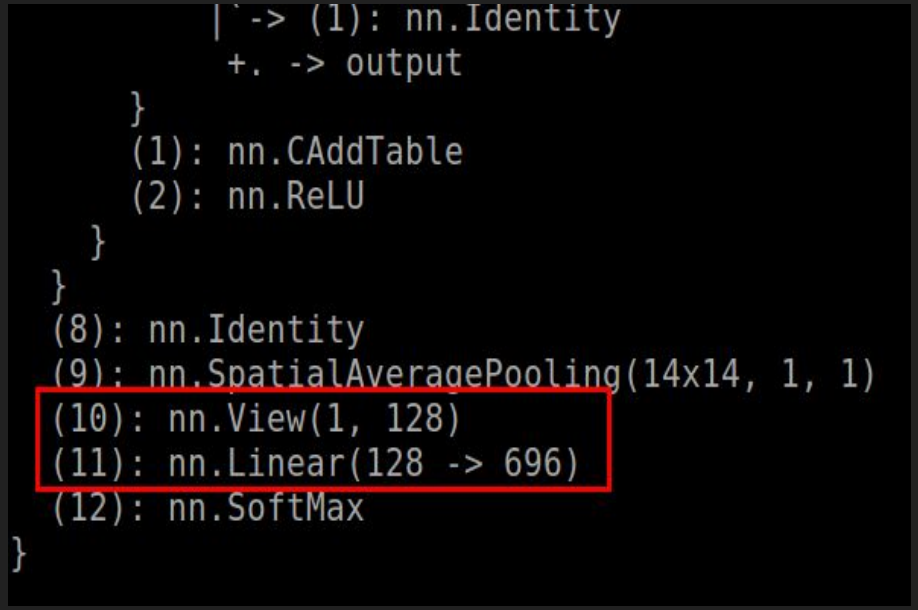


Offline Model Stealing

```java
public Model loadModel(String modelFolder) {
    List<String> categories = loadCategories(modelFolder + "/categories.txt");
    if (categories == null) {
        Log.e(TAG, "Failed to load categories: " + modelFolder + "/categories.txt");
        return null;
    }
    ByteBuffer enginePtr = loadModelFromAssets(modelFolder + "/model.net", modelFolder + "/stat.t7");
    if (enginePtr != null) {
        return new Model(enginePtr, categories, 224);
    }
    Log.e(TAG, "Failed to load model");
    return null;
}
```

```python
# Loading model
from torch.utils.serialization import load_lua
model = load_lua(model_path)
stat = load_lua(model_path[:-9]+'stat.t7')
model_op = predict(IMAGE_PATH)
```

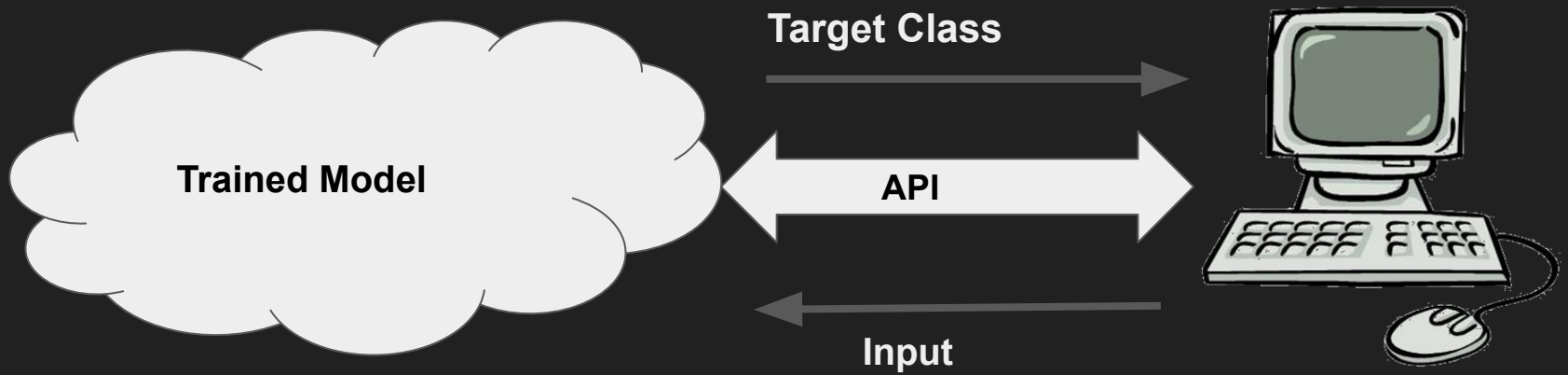




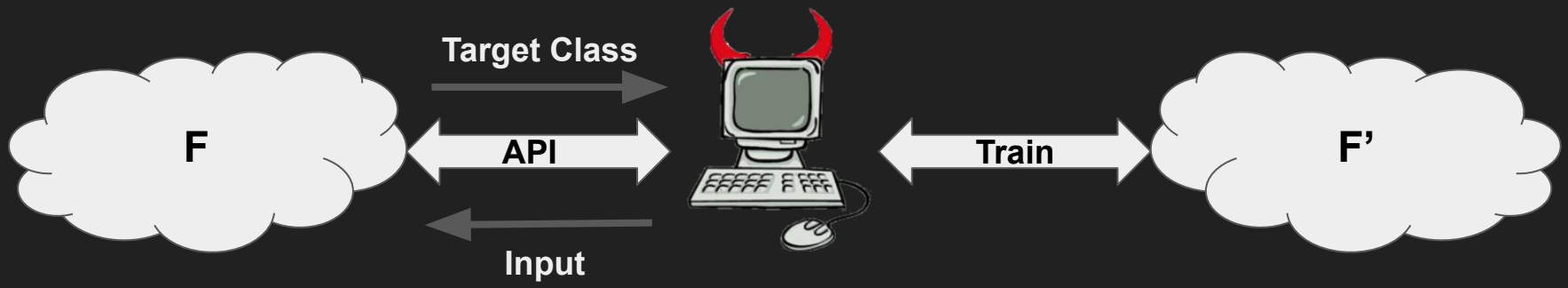

```
torch.legacy.nn.View(1,128)
```

Black Box model stealing

MLaaS



Online Attacks



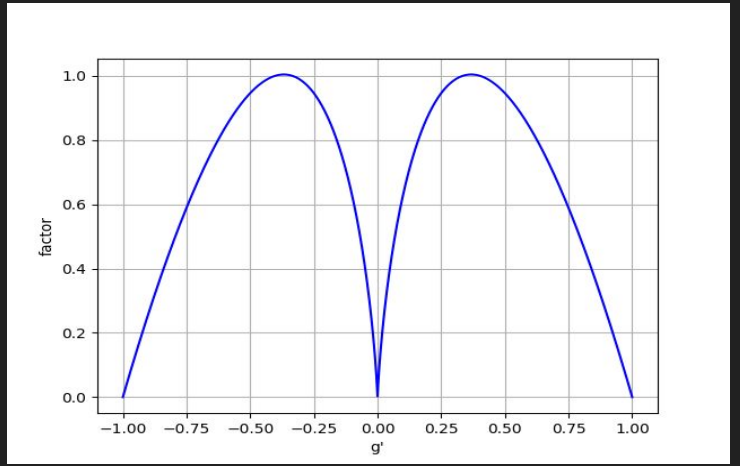
Input: *[f1, f2, f3, ..., fn]*
F output: *P(class1), P(class2), P(class3), …, P(classN)*
API output: *max(P(class1), P(class2), P(class3), …, P(classN))*

GDALR: Gradient Driven Adaptive Learning Rate

$$g_i' = \tanh(g_i) \tag{7}$$

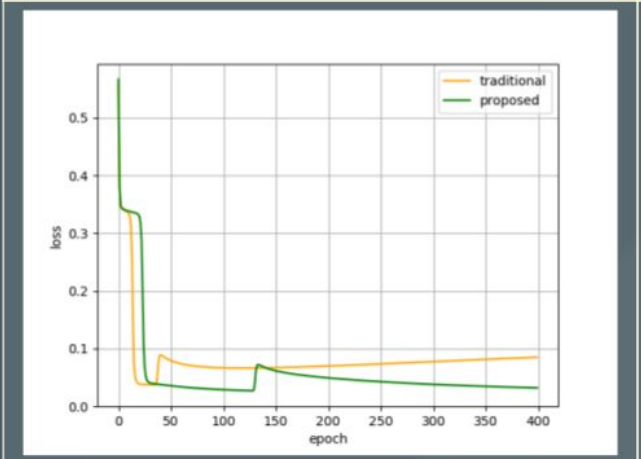
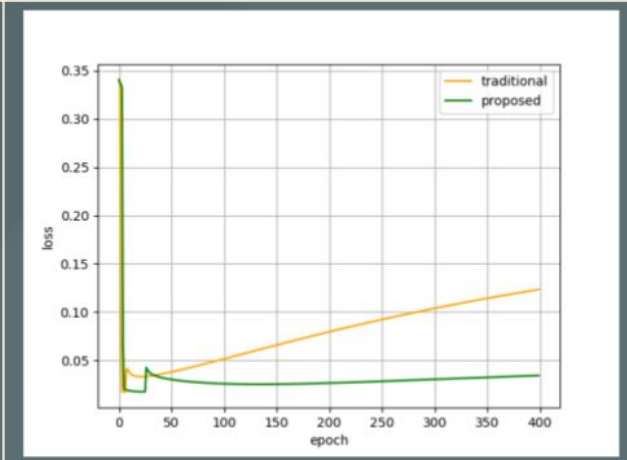
$$fact_i = \text{abs}\left(g_i' 2\pi \log_{10}\left(\text{abs}\left(g_i'\right)\right)\right) \tag{8}$$

$$l_i' = l_i \cdot fact_i \tag{9}$$



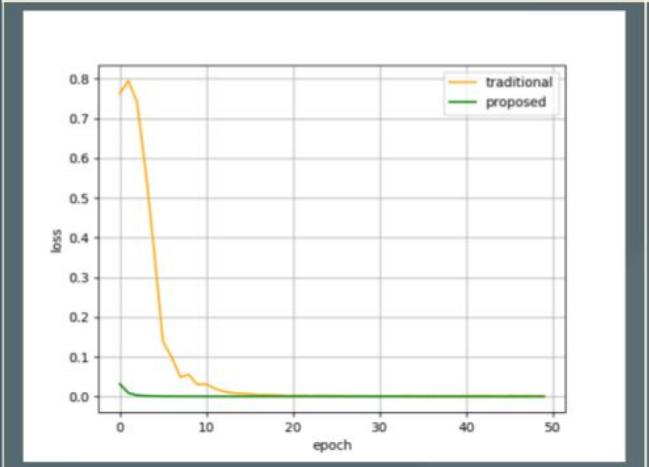
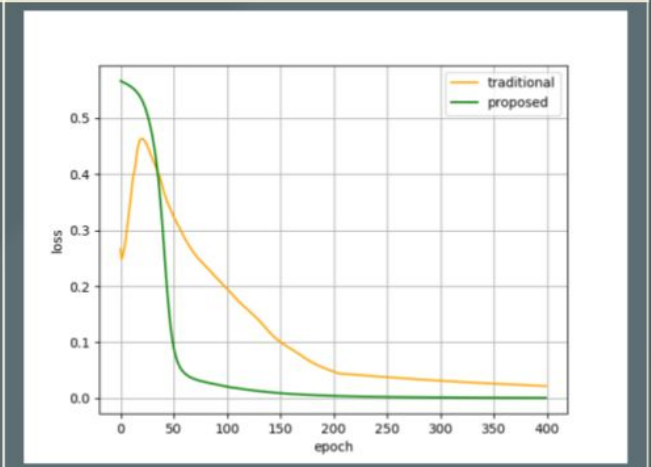
Ref: https://ieeexplore.ieee.org/document/8878726

Results: Logistic Regression

$l = 0.01$	$l = 0.05$
	
$T_{Loss} = 0.0849$	$T_{Loss} = 0.1233$
$P_{Loss} = 0.0317$	$P_{Loss} = 0.0342$

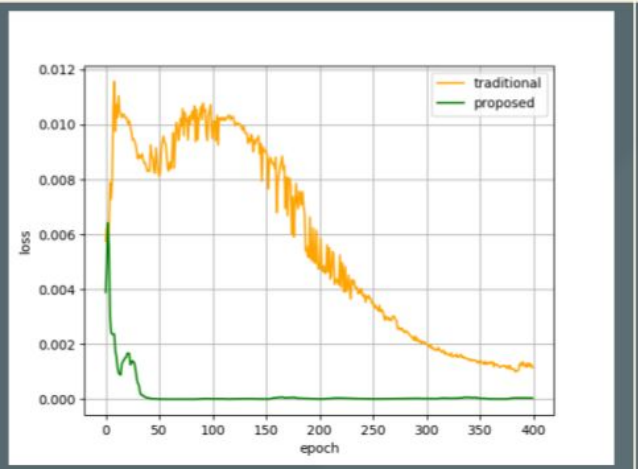
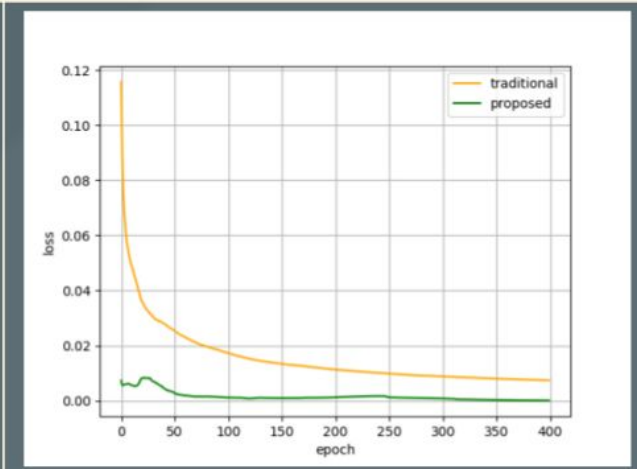
Ref: https://ieeexplore.ieee.org/document/8878726

Results: Multi Layer Perceptron

$l = 10^{-3}$	$l = 10^{-5}$
	
$T_{Loss} = 0.0014$	$T_{Loss} = 0.0219$
$P_{Loss} = 5.444 \times 10^{-5}$	$P_{Loss} = 0.0007$

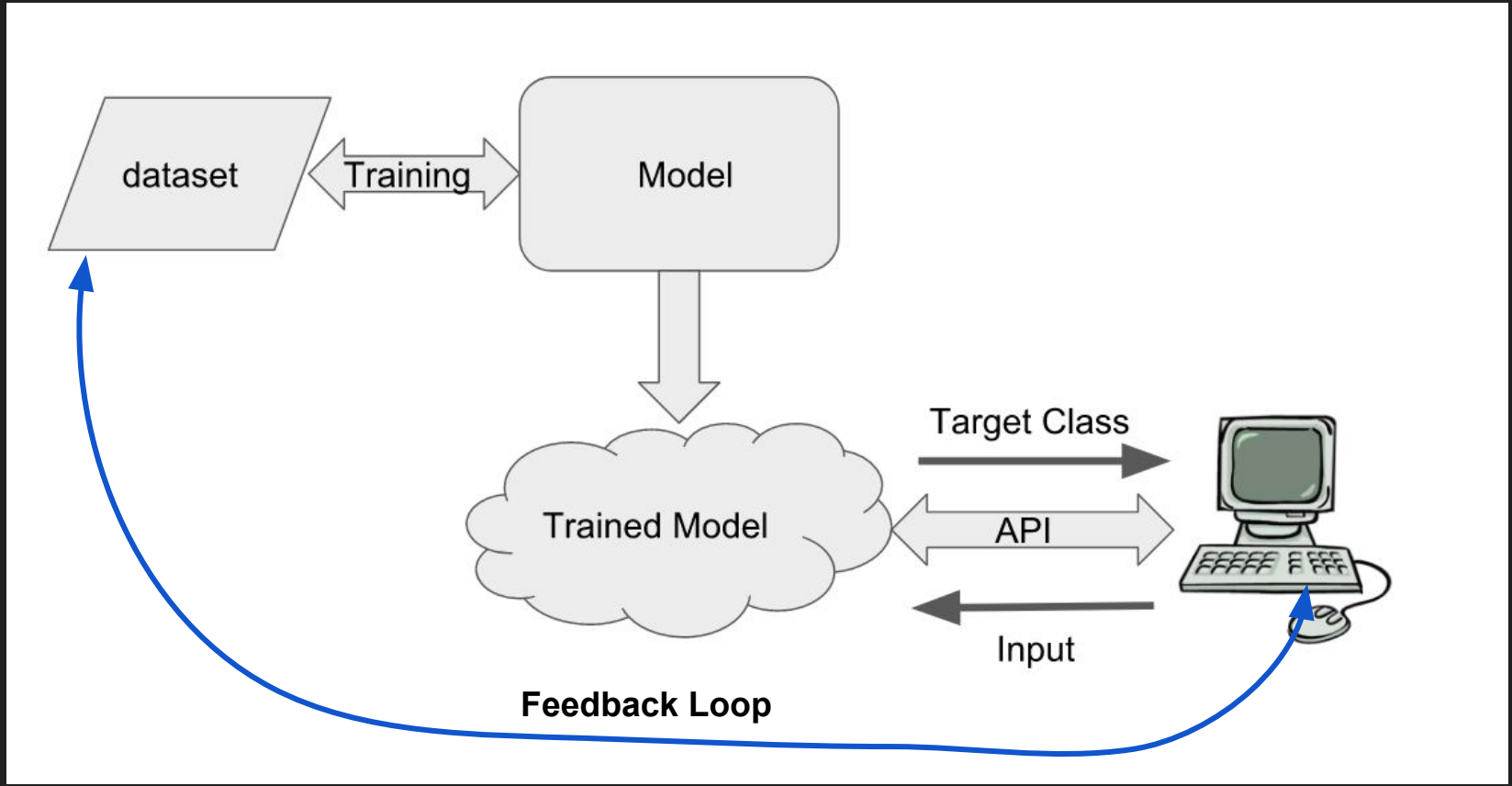
Ref: https://ieeexplore.ieee.org/document/8878726

Results: Convolutional Neural Network

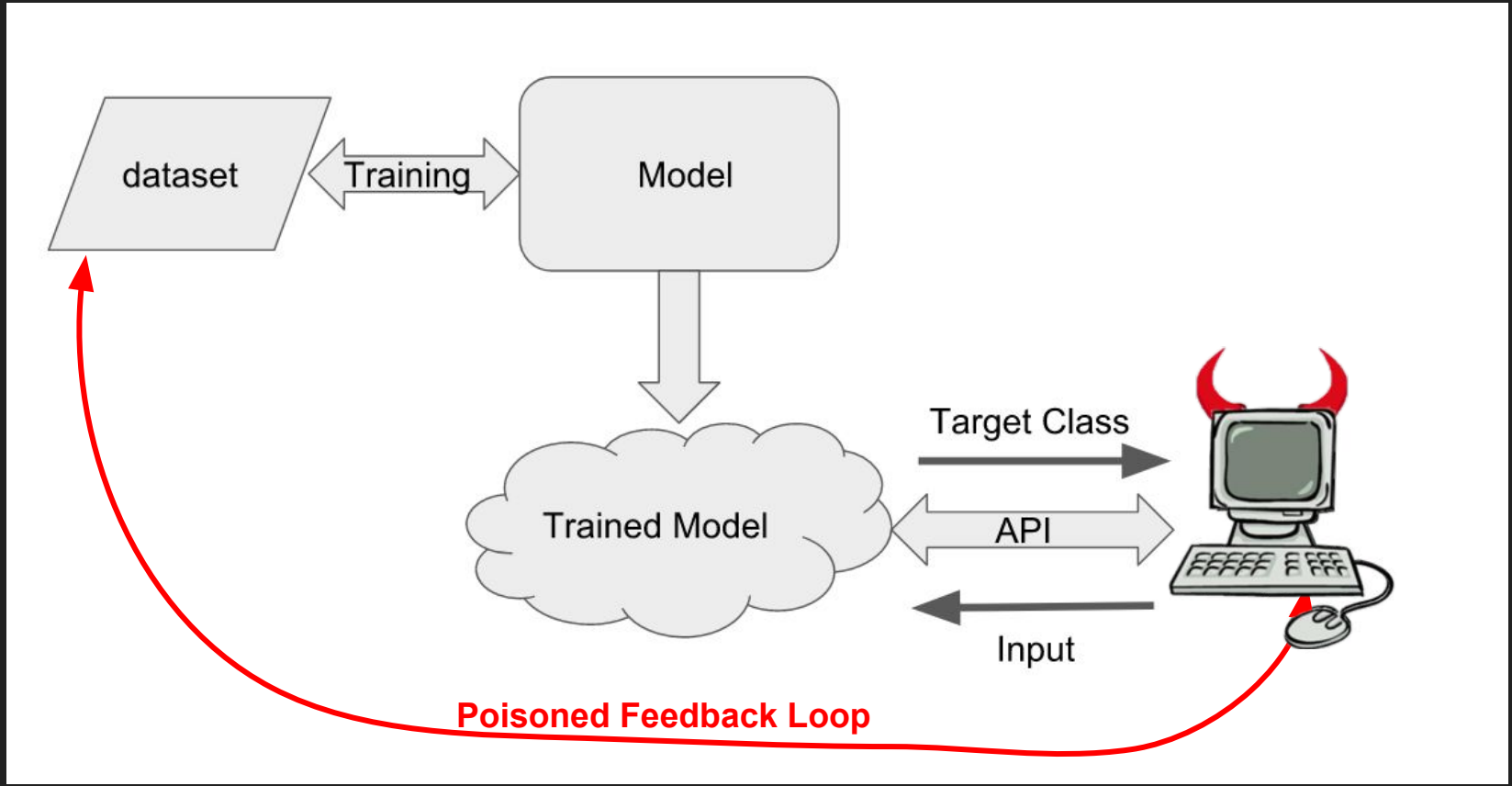
$l = 10^{-4}$	$l = 10^{-5}$
	
$T_{Loss} = 0.0011$	$T_{Loss} = 0.0073$
$P_{Loss} = 3.993 \times 10^{-5}$	$P_{Loss} = 4.184 \times 10^{-5}$

Ref: https://ieeexplore.ieee.org/document/8878726

Model skewing Attacks



Model skewing Attacks

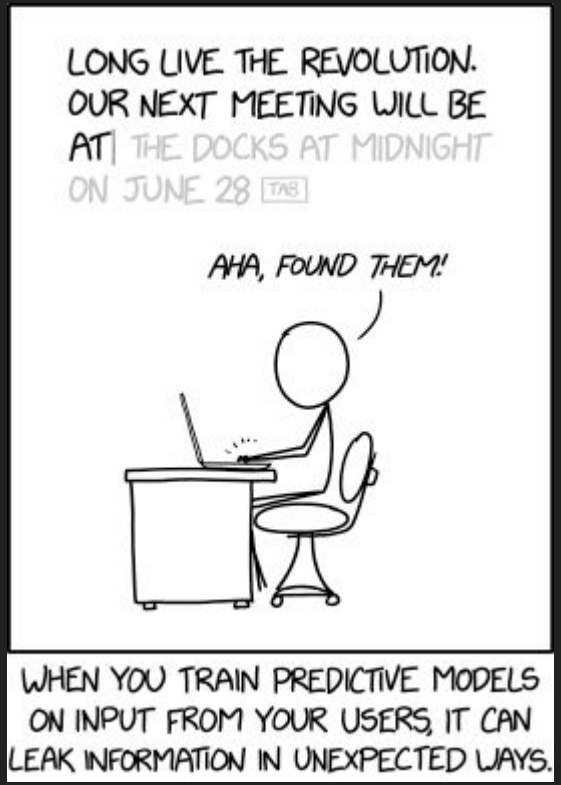


Model Inversion Attack

Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

Ref: https://www.cs.cmu.edu/~mfredrik/papers/fjr2015ccs.pdf

Unintended Memorization in Neural Networks



Ref: https://www.usenix.org/system/files/sec19-carlini.pdf

Final Notes

- ML application learn patterns from “given” dataset
- They are different from signature / rule based applications
- Identify the use-case
- White box or Black box
- Craft Threat Model
- Start with simple Methods

Thank You!

nikhilj@payatu.com

@adversarial_nik