Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples

भारतीय प्रौद्योगिकी संस्थान जम्मू
INDIAN INSTITUTE OF TECHNOLOGY JAMMU
विद्याधनं सर्वधन प्रधानम्

Dhirubhai Ambani Institute of Information and Communication Technology
DA-IICT

Saurabh Gajbhiye
Harshit Shakya
Harsh Sharma
Rahul Kumar
Sachin Pal

21 April 2021

Contents

Introduction

Many machine learning models are vulnerable to

Adversarial Examples attack

We will briefly discuss how these black box attacks are crafted

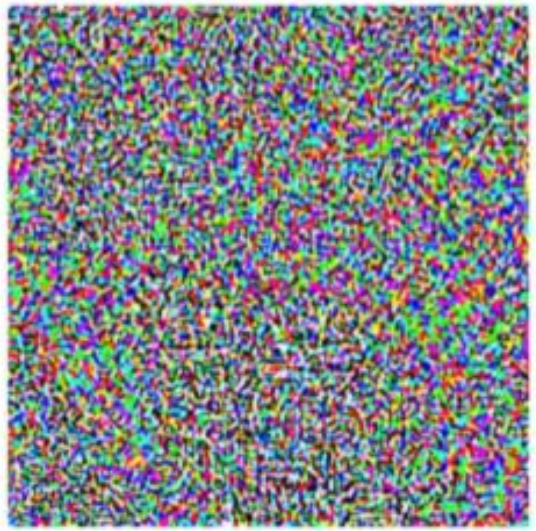
We will see how the result of these attack on real world machine learning models

Key Term

Adversarial Example

Inputs that are specially crafted to cause a machine learning model to produce an incorrect output

$+ .007 \times$



$=$

"panda"
57.7% confidence

Noise

"gibbon"
99.3 % confidence

ML application and consequence

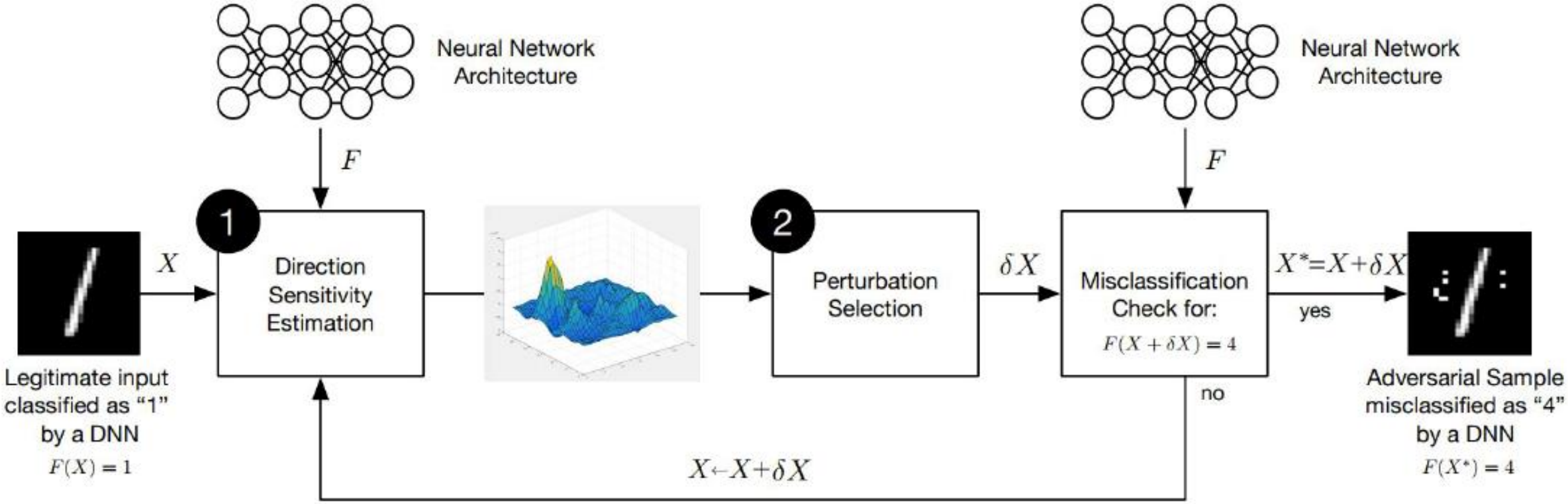
Autonomous Driving

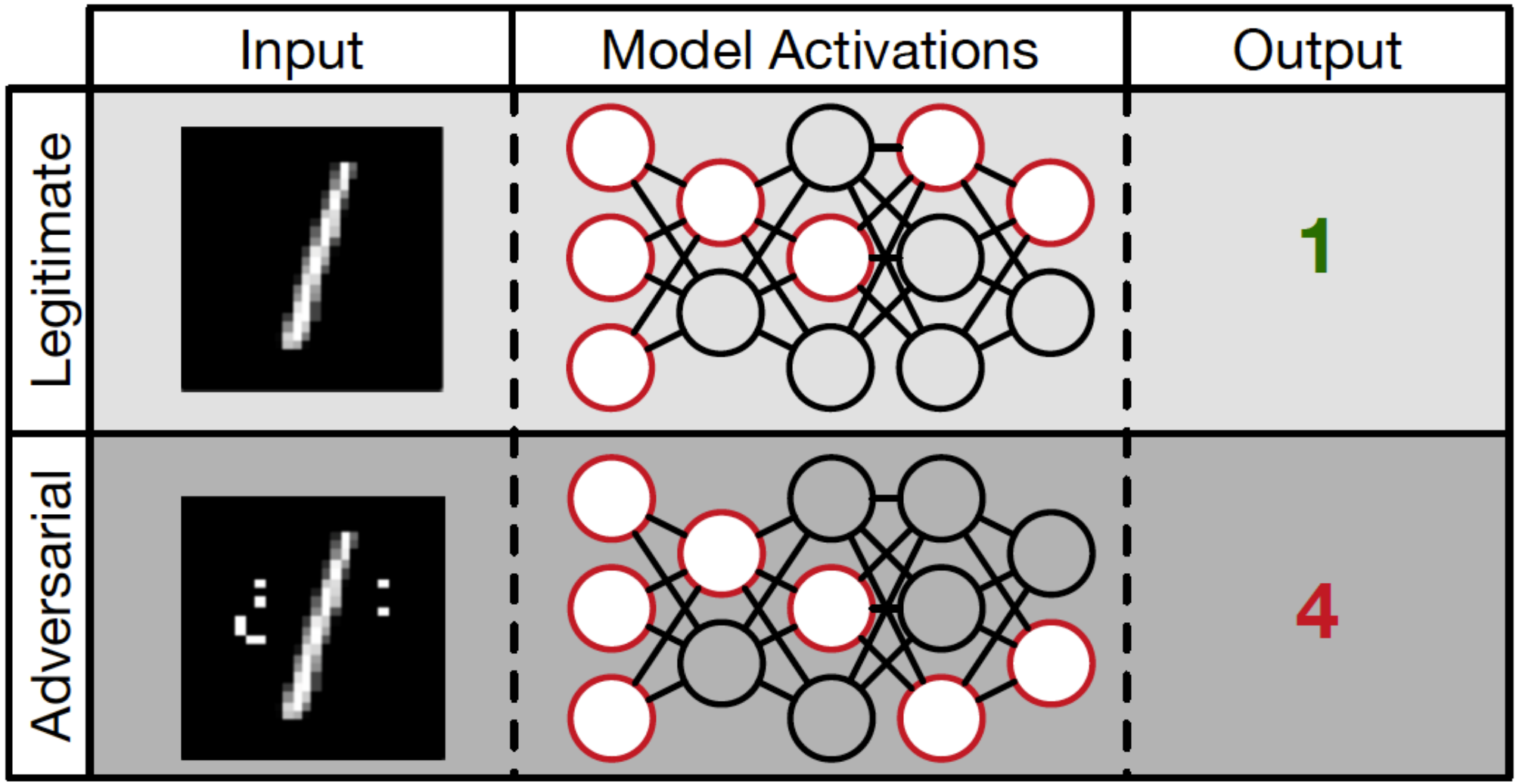
Fraud Detection in Finance

Malware Detection

Machine Learning as a Service Platform

White Box Attack




Input
Model Activations
Output
Legitimate
1
Adversarial
4

Crafting Adversarial Attacks

DNN , LR and kNN use the **fast gradient sign method**

This is a white box attack

Attacker is aware about the :
Model Architecture
Training Data
Model Parameters

Fast gradient sign method

During training, the classifier uses a loss function to **minimize** model prediction errors

After training, **attacker** uses loss function to **maximize** model prediction error

1. Compute its gradient with respect to the input of the model

2. Take the sign of the gradient and multiply it by a threshold

Black box attack

In this attack the attacker is not aware about:

Model Architecture

Training Data

Model Parameters

He has access to model via API, when a input is passed to a model, a label (output) is returned.

Threat Model of Black box attack

Adversarial capabilities

~~Training data~~
~~Model architecture~~
~~Model parameters~~
~~Model scores~~

(limited) oracle access: *labels*

Adversarial goal

Force a ML model remotely accessible through an API to misclassify

Example

amazon web services

Google Cloud Platform

Challenges in Black Box attack

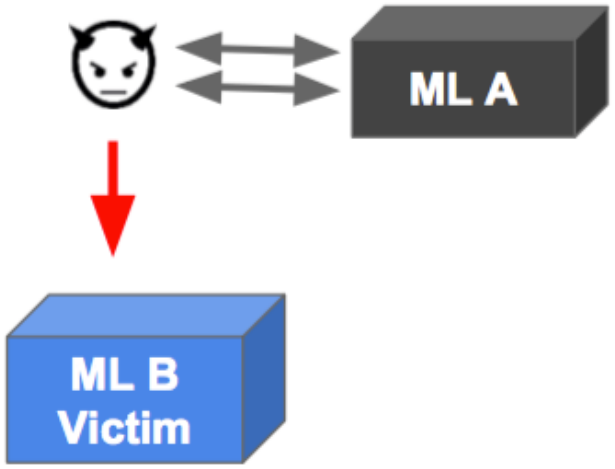
Alleviate lack of knowledge about model

Alleviate lack of training data

Challenge 1: Use Adversarial example transferability

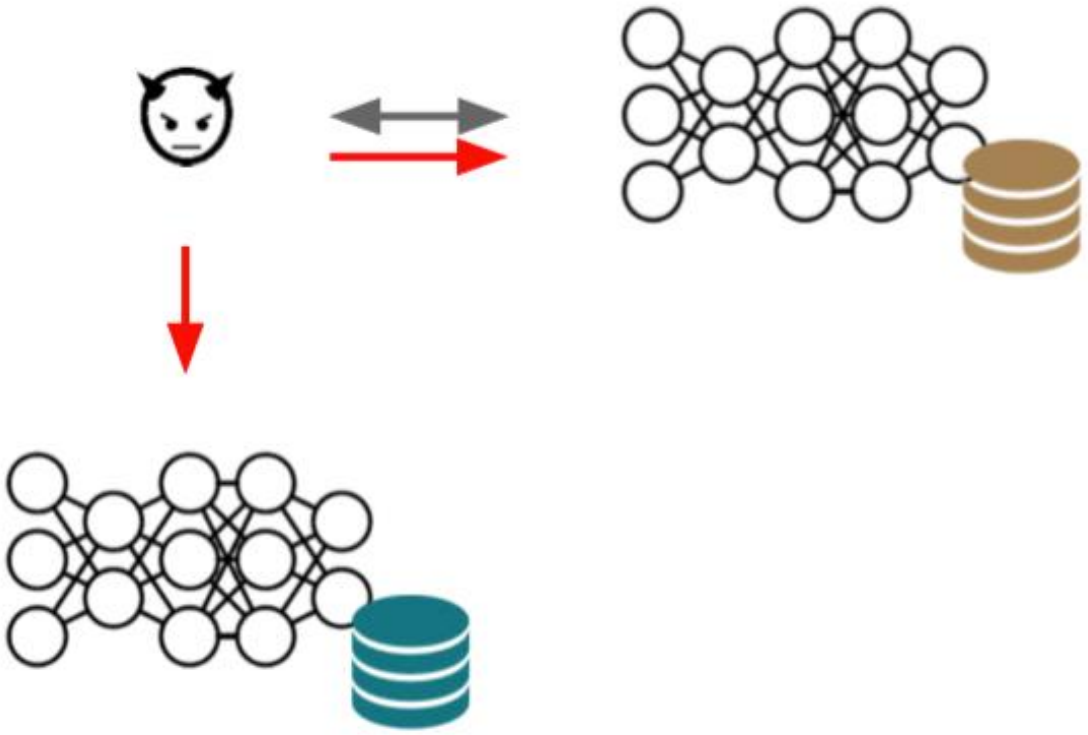
Adversarial examples have a **transferability** property:

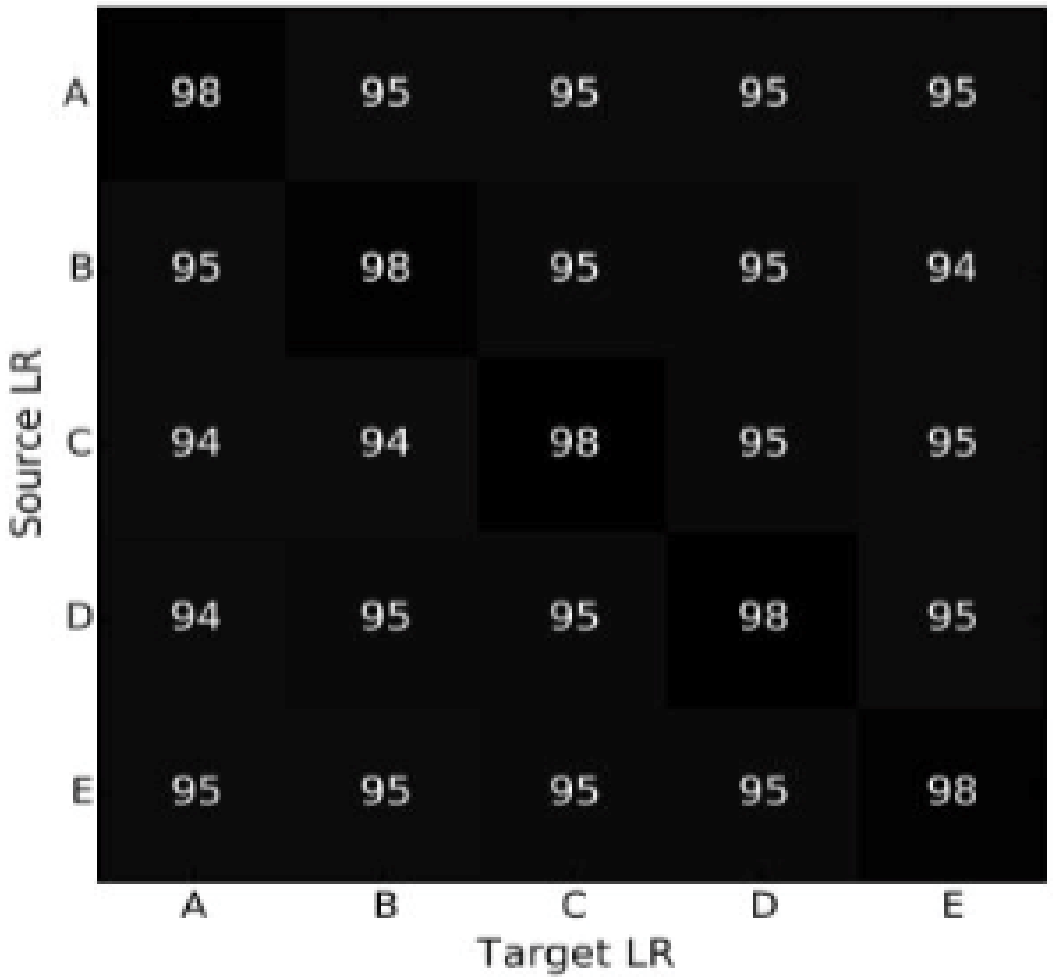
samples crafted to mislead a model A are likely to mislead a model B



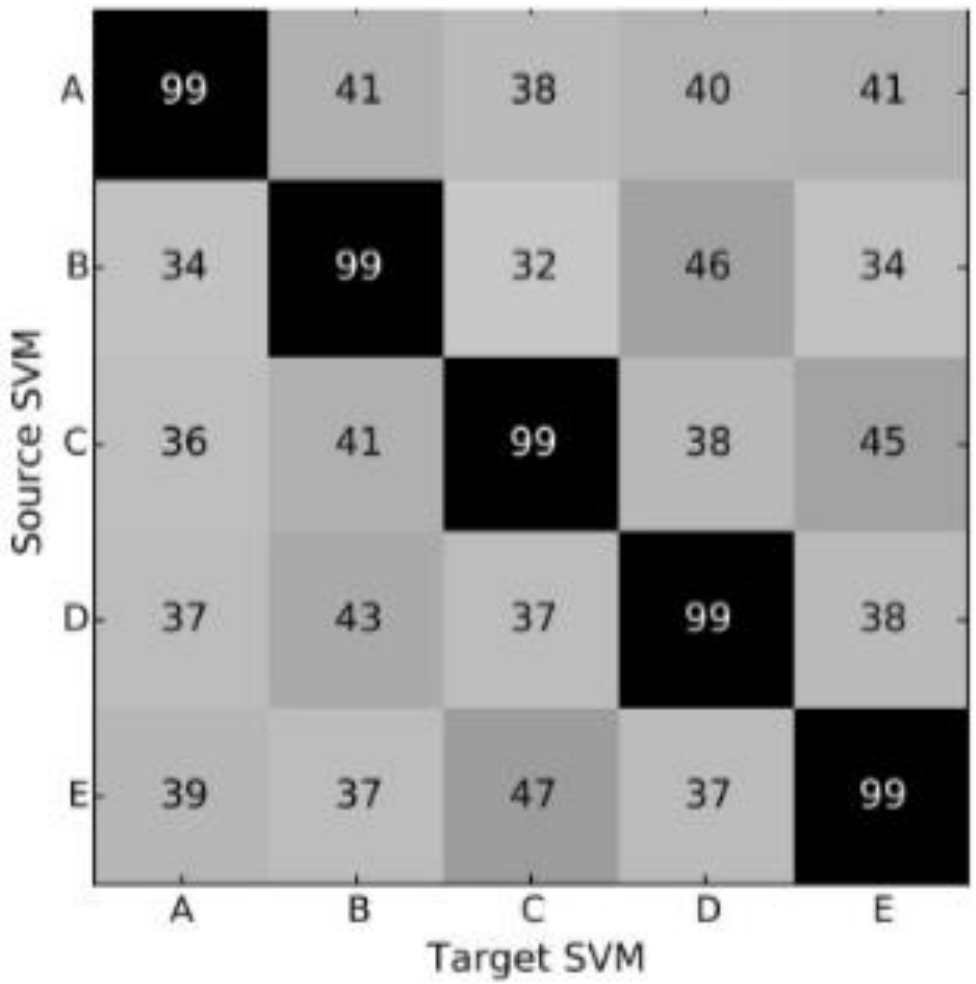
Intra-technique Transferability

samples crafted to mislead a model A are likely to mislead a model B

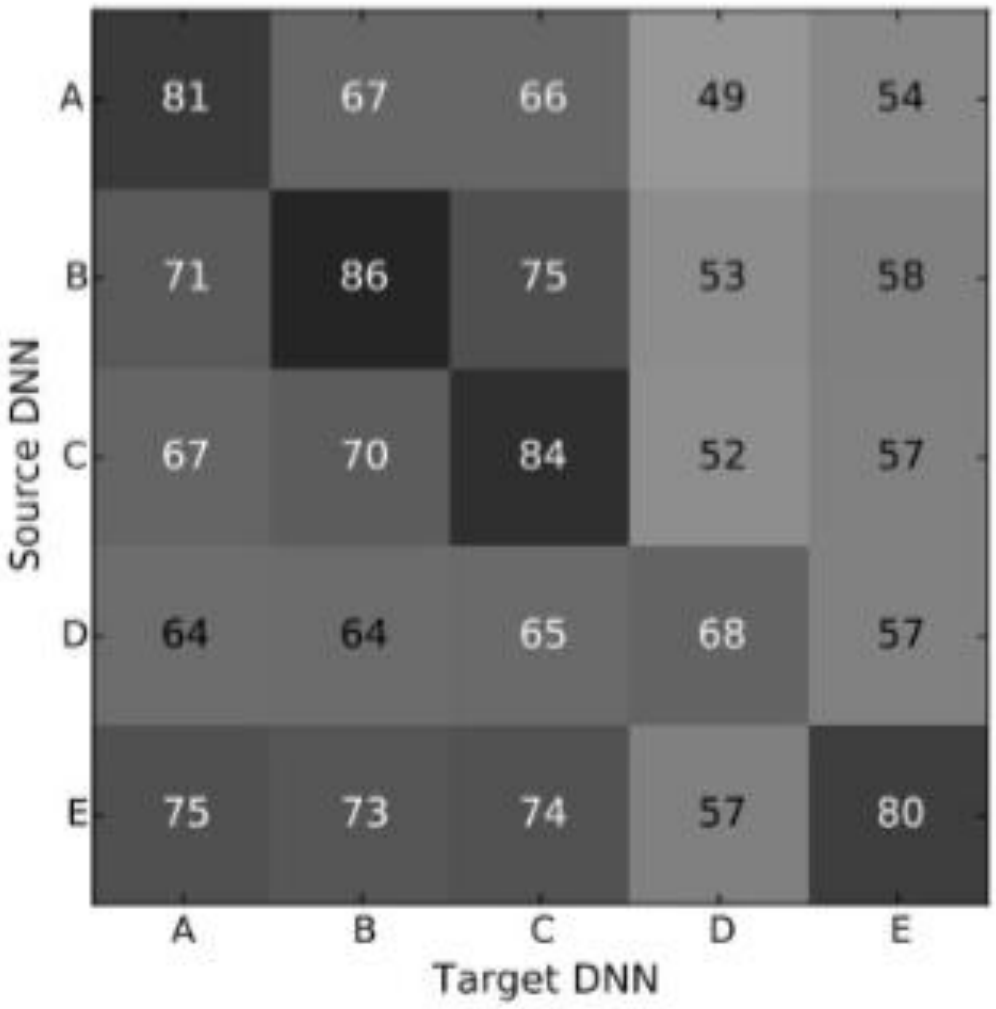



Source LR
A
B
C
D
E
98 95 95 95 95
95 98 95 95 94
94 94 98 95 95
94 95 95 98 95
95 95 95 95 98
A
B
C
D
E
Target LR

Strong


A 99 41 38 40 41
B 34 99 32 46 34
C 36 41 99 38 45
D 37 43 37 99 38
E 39 37 47 37 99
A B C D E
Source SVM
Target SVM

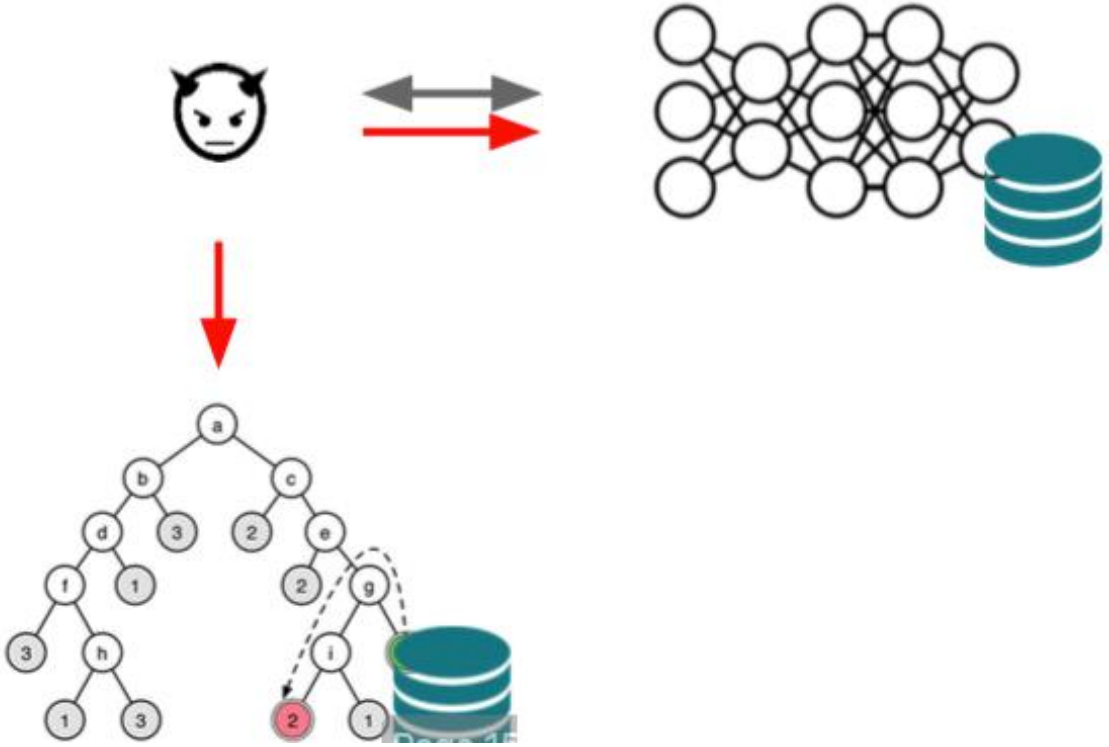
Weak

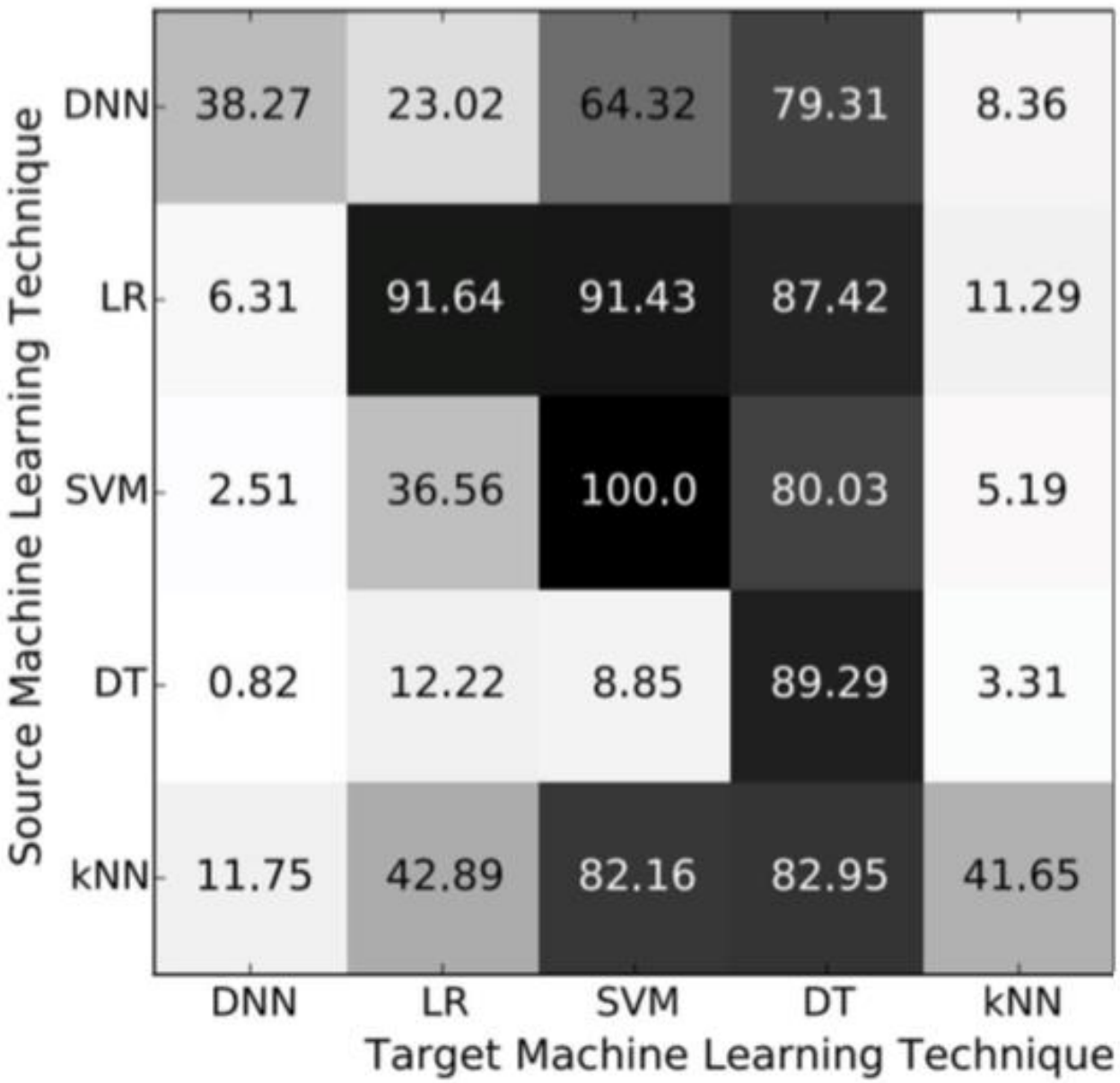


Intermediate

Cross-technique Transferability

samples crafted to mislead a model A are likely to mislead a model B



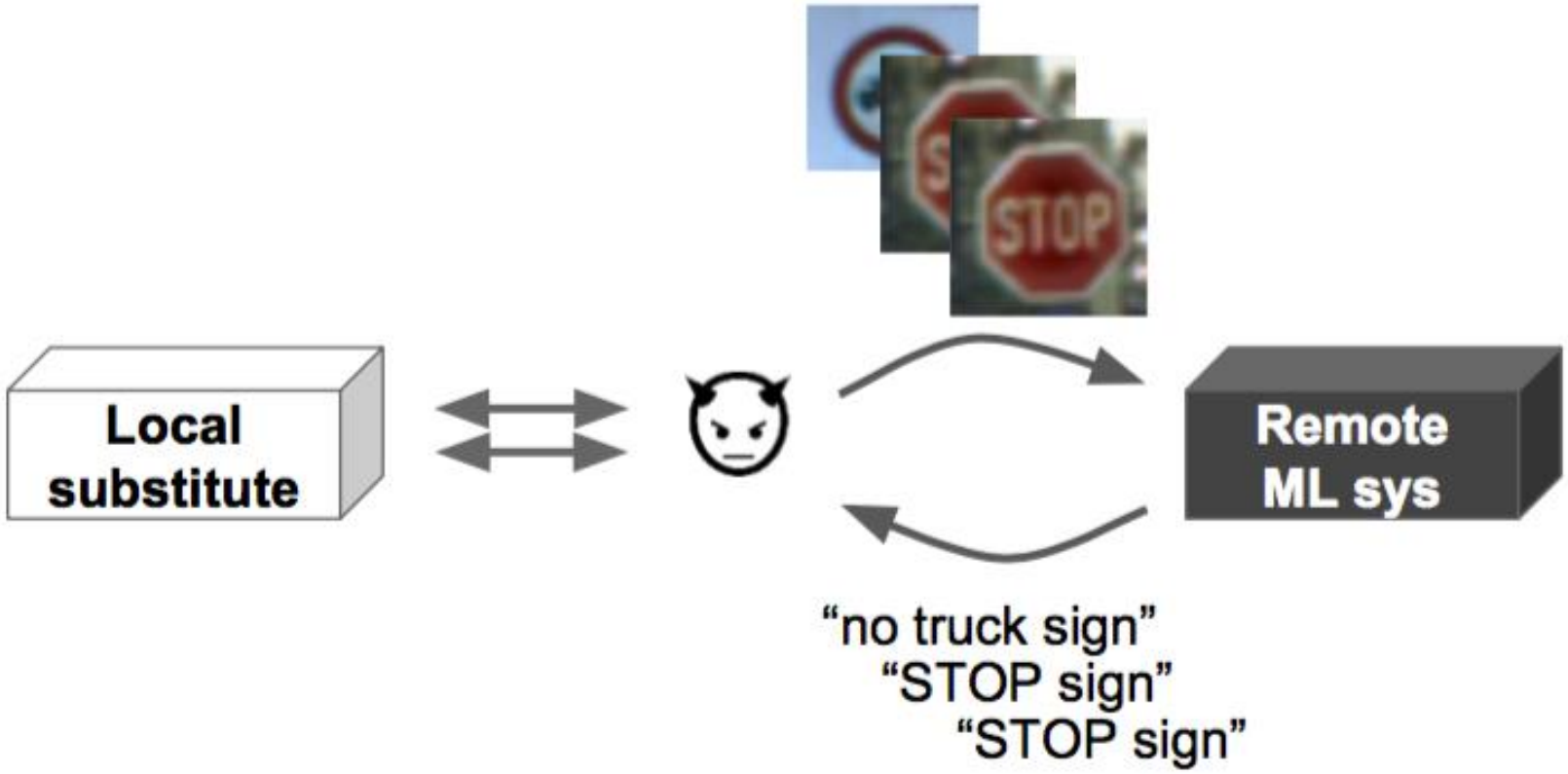

Source Machine Learning Technique
DNN
LR
SVM
DT
kNN
38.27
23.02
64.32
79.31
8.36
6.31
91.64
91.43
87.42
11.29
2.51
36.56
100.0
80.03
5.19
0.82
12.22
8.85
89.29
3.31
11.75
42.89
82.16
82.95
41.65
DNN
LR
SVM
DT
kNN
Target Machine Learning Technique

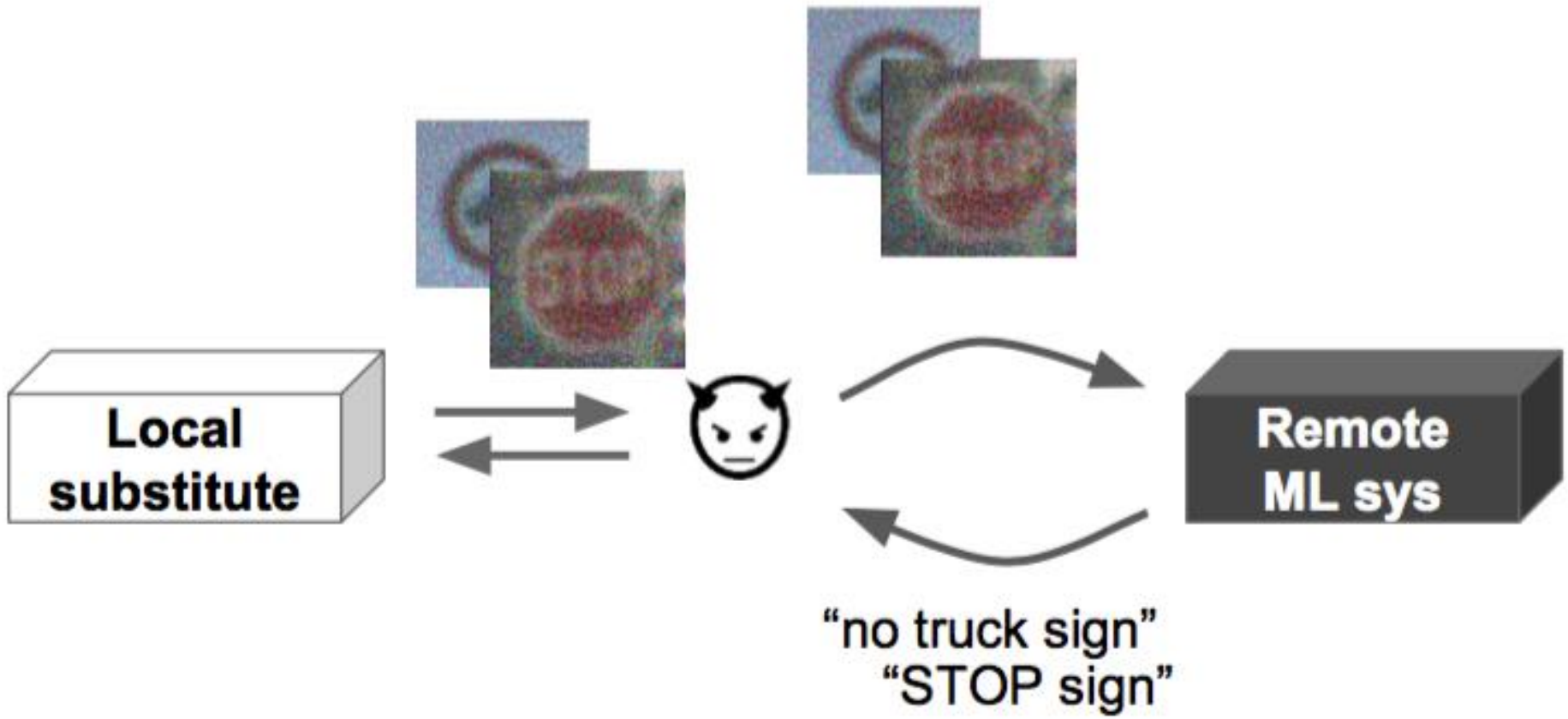
Alleviate lack of knowledge about model

Alleviate lack of training data

Adversarial example transferability from a substitute model to target model

Challenge 2: Training Model using Synthetic data generation




Local substitute
Remote ML sys
"no truck sign"
"STOP sign"


Local substitute
STOP
STOP
Remote ML sys
"yield sign"
YIELD

Alleviate lack of knowledge about model

↓

Adversarial example transferability from a substitute model to target model

+

Alleviate lack of training data

↓

Synthetic data generation

Results

Remote Platform	ML technique	Number of queries	Adversarial examples misclassified (after querying)
MetaMind	Deep Learning	6,400	84.24%
amazon web services™	Logistic Regression	800	96.19%
Google Cloud Platform	Unknown	2,000	97.72%

Conclusion

We saw phenomenon of adversarial
sample transferability across the machine learning models.

We demonstrated how to create black box attack that could be used to target online ML models trained and hosted by Amazon and Google, without any knowledge of the model design or parameters

Thank you!