Introduction of
Machine Learning Security

Lecture 04

Other Attacks & Applications

Patrick Chan
patrickchan@ieee.org

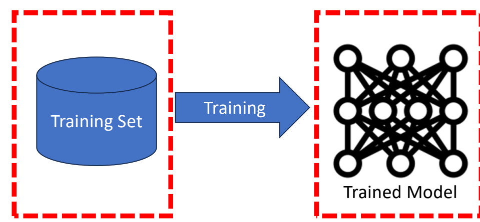
Agenda

- Privacy Attack
- Physical Attack
- Non-Security Application
- Conclusion

Privacy Attack

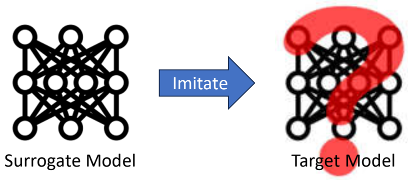
Objective

- Model Stealing
- Training Set Recovery



Model Stealing

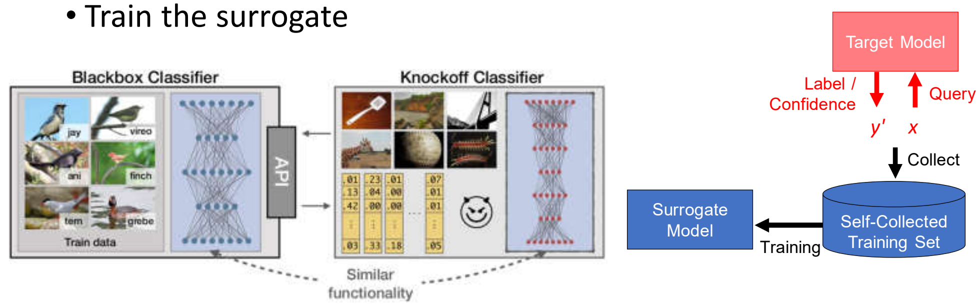
- Construct a copy of a model
- Two possible goals:
  - Intellectual property for well performance
  - Surrogate model for evasion attack



Model Stealing

Query Attack

- Procedure
  - Query the target model and collects output
  - Train the surrogate



Orekondy et al., Knockoff Nets: Stealing Functionality of Black-Box Models, CVPR 2019

Model Stealing

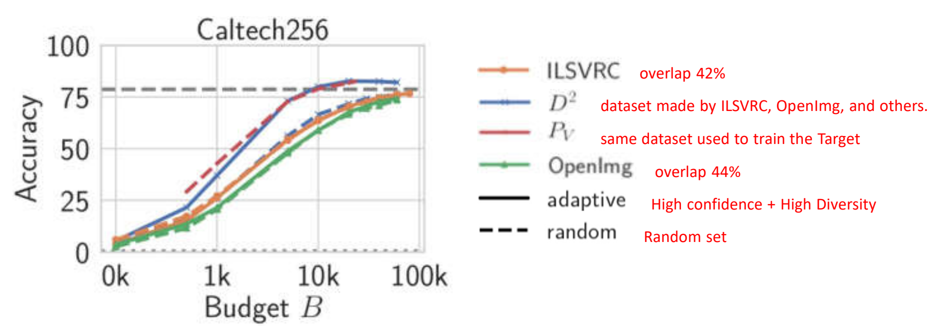
Query Attack

- How to generate query samples?
  - Select samples randomly
    - May not be effective
    - Many queries are required
  - Select/Generate samples specifically according to
    - High Output Confidence: Only the confident samples
    - High Diversity: Different information

Orekondy et al., Knockoff Nets: Stealing Functionality of Black-Box Models, CVPR 2019

Model Stealing

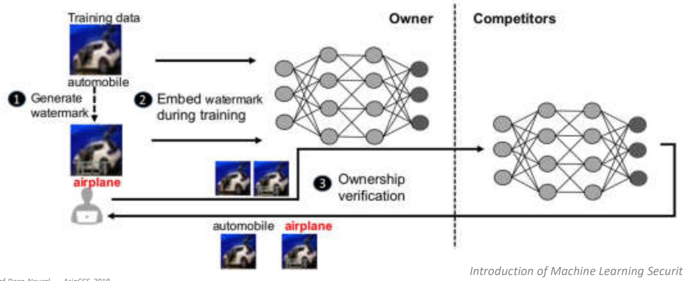
Query Attack



Model Stealing Defense

Copyright Verification

- Add watermarks to the model for the proof of intellectual property
  - Watermarks: patterns that cause an unexpected misclassification when added to an image

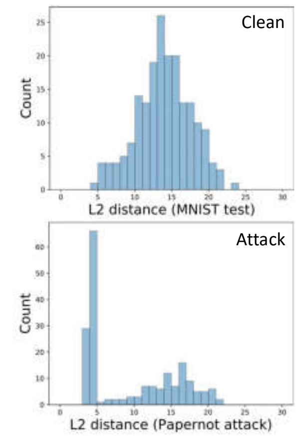


Zhang et al., Protecting Intellectual Property of Deep Neural ..., AsiaCCS 2018

Model Stealing Defense

Block Query

- Understand that attackers are querying the model and block them.
- Distance between consecutive queries for a legitimate purpose usually follow a normal distribution, but not for an attack

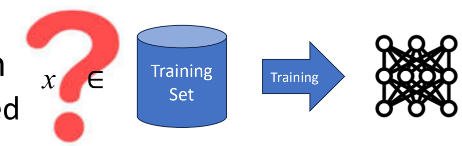


Training Set Recovery

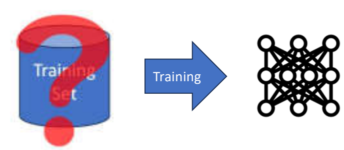
- Training samples may contain sensitive information
  - Personal information
  - Financial information
  - Images of suspected people
- Even if recovered training samples are incomplete, they can still be combined to re-identify individuals.

Training Set Recovery

- Different Levels of Recovery
- Training Sample Identification
  - Identify whether a sample used in training



- Training Sample Reconstruction
  - Construct the training set according to the model

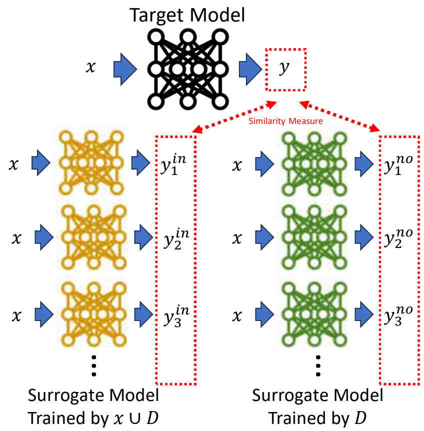


Shokri et al., Membership Inference Attacks Against Machine Learning Models, S&P 2017

Training Set Recovery

Training Sample Identification

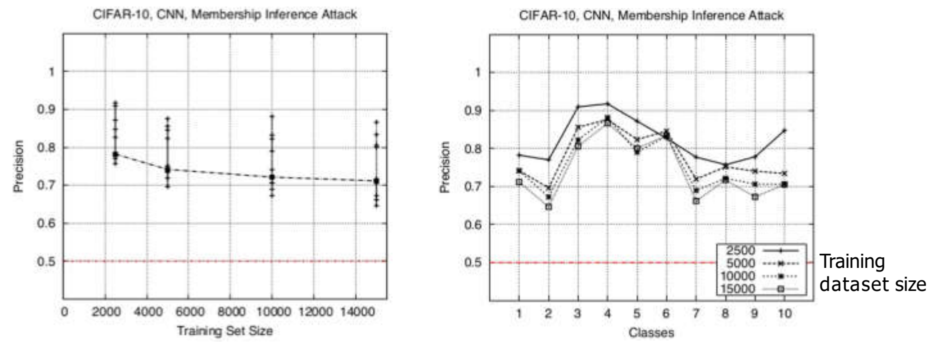
- Membership Inference Attacks
  - Query the target model with the input sample **x**
  - Many surrogate model pairs are crafted:
    - including **x** in the training dataset
    - not including **x** in the training dataset
  - Determine whether the sample is used in training by comparing the predictions of those surrogate models with the target model



Shokri et al., Membership Inference Attacks Against Machine Learning Models, S&P 2017

Training Set Recovery

Training Sample Identification



Shokri et al., Membership Inference Attacks Against Machine Learning Models, S&P 2017

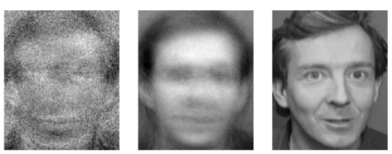
Training Set Recovery

Training Sample Reconstruction

- Model Inversion Attack
  - Reconstruct a training sample by maximizing confidence with respect to the target label using gradient descent
  - Query ability is required

Algorithm 1 Inversion attack for facial recognition models.

function MI-FACE($label, \alpha, \beta, \gamma, \lambda$)
 $c(\mathbf{x}) \stackrel{\text{def}}{=} 1 - \tilde{f}_{label}(\mathbf{x}) + \textsc{AuxTerm}(\mathbf{x})$
 $\mathbf{x}_0 \leftarrow \mathbf{0}$
 for $i \leftarrow 1 \ldots \alpha$ **do**
  $\mathbf{x}_i \leftarrow \textsc{Process}(\mathbf{x}_{i-1} - \lambda \cdot \nabla c(\mathbf{x}_{i-1}))$
  if $c(\mathbf{x}_i) \geq \max(c(\mathbf{x}_{i-1}), \ldots, c(\mathbf{x}_{i-\beta}))$ **then**
   break
  if $c(\mathbf{x}_i) \leq \gamma$ **then**
   break
 return $[\arg\min_{\mathbf{x}_i}(c(\mathbf{x}_i)), \min_{\mathbf{x}_i}(c(\mathbf{x}_i))]$



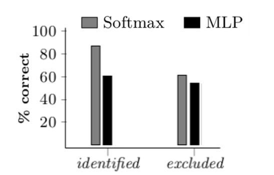
Fredrikson et al, Model inversion attacks that exploit confidence information and basic countermeasures, ACM CCS, 2015

Training Set Recovery

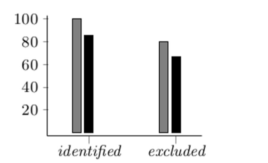
Training Sample Reconstruction

- Mechanical Turk workers are asked to match the reconstructed image to one of five face images from the original training set

	Identify by Workers	Cannot Identify by Workers
Present in the selected images	Identified	Not Match
Not present in the selected images	Not Match	Excluded



(a) Average over all responses.



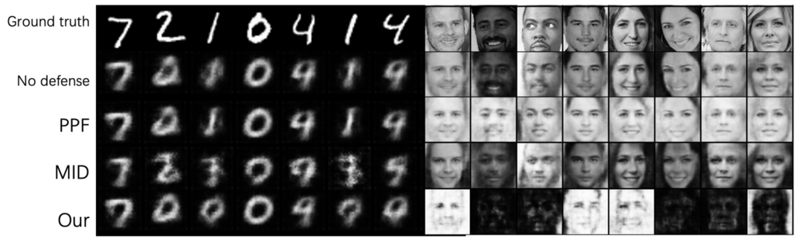
(c) Accuracy with skilled workers.

Target Softmax MLP

Model Inversion Attack

- Maximize the reconstruction error without changing the labels

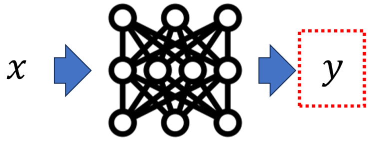
$$
\begin{aligned}
\max \quad & \mathcal{R}(\boldsymbol{x}, \mathcal{A}(\boldsymbol{y}+\boldsymbol{e})) && \text{reconstruction error} \\
\textit{subject to}: \quad & \boldsymbol{e} \leq \epsilon && \text{upper bound on modification} \\
& \arg\max(\boldsymbol{y}+\boldsymbol{e}) = \arg\max \boldsymbol{y} && \text{same predicted labels on original and attack samples} \\
& 0 \leq (\boldsymbol{y}_i + \boldsymbol{e}_i) \leq 1, \ \sum(\boldsymbol{y}_i + \boldsymbol{e}_i) = 1 && \text{Probability maintenance}
\end{aligned}
$$



Wen et al., Defending Against Model Inversion Attack by Adversarial Examples, CSR workshop, 2021

Prediction Vector Tampering

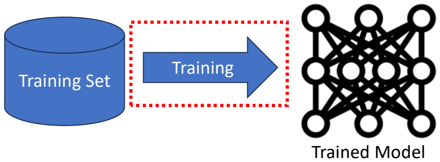
- Privacy attacks usually assume knowledge of the classifier's scores
- Control the outputs of queries:
  - Score Blocking: provide only label but not scores for classes
  - Scores Perturbation: reduce reliability of scores



Jia et al, MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples, CCS, 2019
Shokri et al, Membership Inference Attacks Against Machine Learning Models, S&P, 2017
Rigaki et al, A Survey of Privacy Attacks in Machine Learning, ArXiv, 2021

Regularization

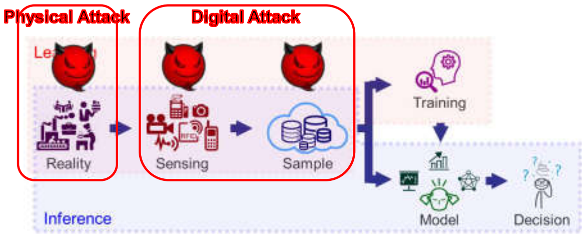
- Deep neural networks tend to memorize training data (they are really confident when predicting them)
- Considering additional terms that are irrelevant to the samples, such as regularization, can reduce memorization on the training samples



[1] Feldman et al., What Neural Networks Memorize and Why:.., NeurIPS 2020

Physical Attack

- Previous discussion focuses on digital representation
- Input can be precisely controlled
- Can adversarial attack be applied to our real world?



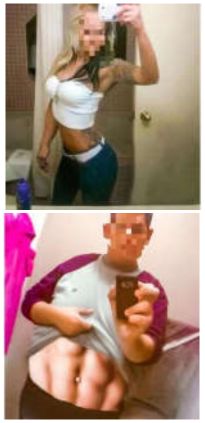
Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer(2016) Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: 2016 acm sigsac conference on computer and communications security

Physical Attack

- A printed contaminated stop sign

Physical VS Digital World

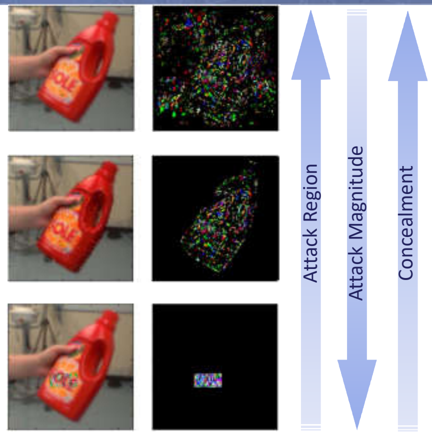
- Gap between physical and digital world
  - Spatial Constraints
    - Adversarial noise should only appears on the object but not the background
  - Physical Limits on Imperceptibility
    - Small perturbations are almost imperceptible to sensors
  - Environmental Conditions
    - Distance, angle, lighting/weather conditions
  - Fabrication Error
    - Reproduction error, e.g. printer limitation



Kevin Eykholt, Ivan Evtimov, Earlence Fernandes (2017)Robust Physical-World Attacks on Deep Learning Visual Classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition

Attack Region

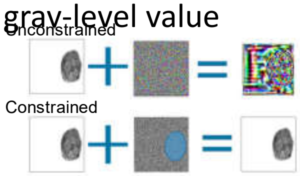
- Digital Attack
  - Any features
  - Cannot be used in reality
- Poster/Wrapper Attack
  - Features in object
- Sticker Attack
  - Features is a small area
  - Easier to implement

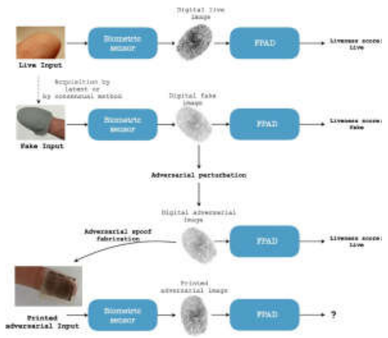


M Melis, A Demontis, B Biggio(2017) Is Deep Learning Safe for Robot Vision? Adversarial Examples against the iCub Humanoid. In: Proceedings of the IEEE International Conference on Computer Vision Workshops

Fingerprint

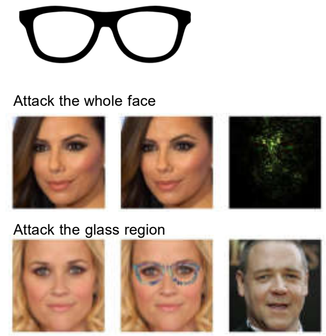
- Evade fingerprint liveness detection
- Attack is limited:
  - Region:
    Actual fingerprint
  - Value:
    only gray-level value

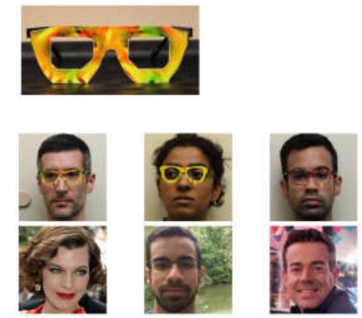




Glass

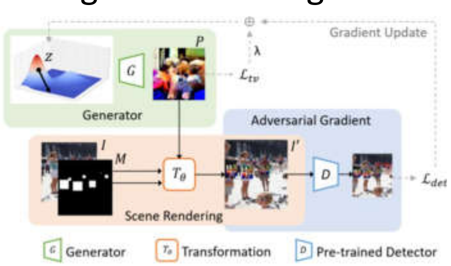
- Only attack the features in a glass mask





Clothes

- Embed a generated image to a clothing region



$$L_{det} = \frac{1}{N}\sum_{i=1}^{N} \max_{j} \left[D^{j}_{obj}(I'_i) D^{j}_{cls}(I'_i)\right]$$

adversarial detection loss

$$L_{tv} = \sum_{i,j} \sqrt{(P_{i+1,j} - P_{i,j})^2 + (P_{i,j+1} - P_{i,j})^2}$$

smoothness of a generated image

Limit Attack Region is not enough

- Objects can be viewed from different distances and angles
- **Distance**: Approach to a printed contaminated stop sign
  - Misclassified as "sports ball" in two frames
- **Angle**: Camera moves closely around a printed original and contaminated stop signs
  - Misclassified as "toilet" in two frames

Environmental Conditions

- Simulate the real situations by considering transformations of viewpoint shifts, camera noise, and other natural noises
- Expectation Over Transformation (EOT)

In different transformation

$$\underset{x'}{\arg\max}\ \mathbb{E}_{t\sim T}\left[\log P(y_t|t(x')) - \lambda||LAB(t(x')) - LAB(t(x))||_2\right]$$

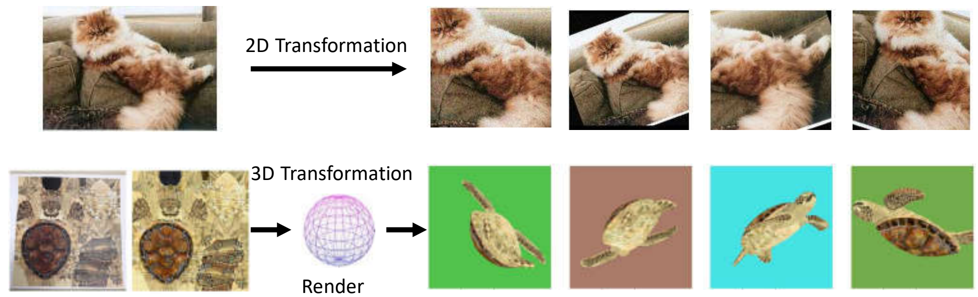
Wrong Decision   Visual Imperceptibility

  - T: Transformation
  - LAB: a space for measuring human perceptual distance

Anish Athalye, Logan Engstrom, Andrew Ilyas(2018) Synthesizing Robust Adversarial Examples. In: International conference on machine learning

Environmental Conditions

- 2D: rotation, transformation, or addition of noise
- 3D: angle, texture and a pose of the 3D object

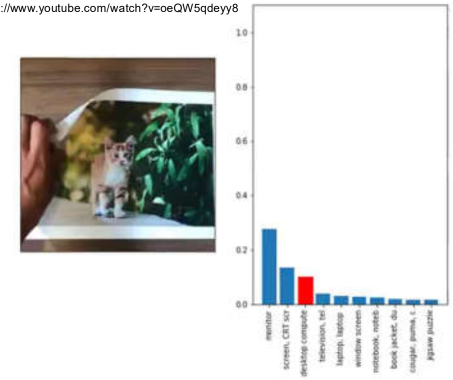


Anish Athalye, Logan Engstrom, Andrew Ilyas(2018) Synthesizing Robust Adversarial Examples. In: International conference on machine learning

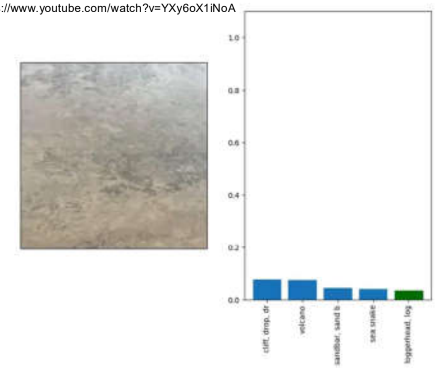
Environmental Conditions

- Expectation Over Transformation (EOT)

https://www.youtube.com/watch?v=oeQW5qdeyy8



https://www.youtube.com/watch?v=YXy6oX1iNoA

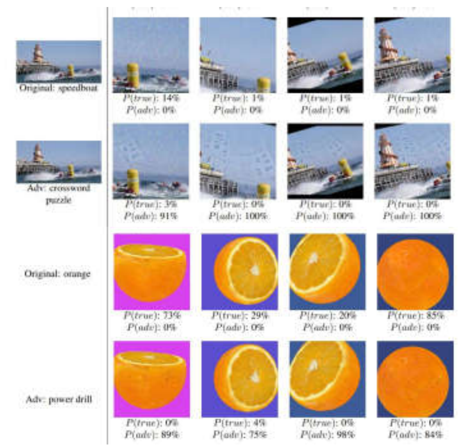


Anish Athalye, Logan Engstrom, Andrew Ilyas(2018) Synthesizing Robust Adversarial Examples. In: International conference on machine learning

Environmental Conditions

2D Image

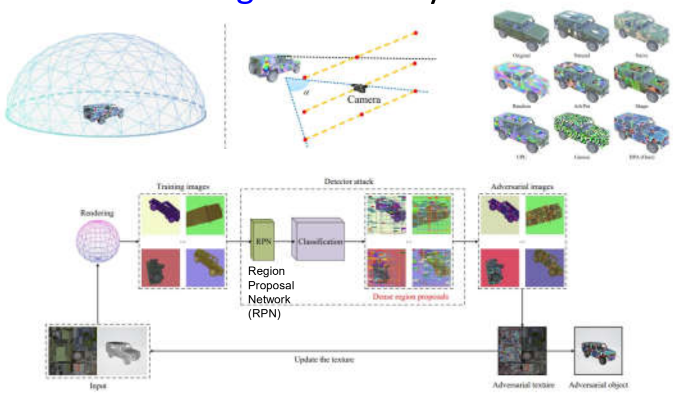
3D Model



Environmental Conditions

Object Detection

- Consider different angles in reality



Yexin Duan, Jialin Chen, Xingyu Zhou(2023) DPA: Learning Robust Physical Adversarial Camouflages for Object Detectors. In:arXiv

Imperceptibility & Fabrication Error

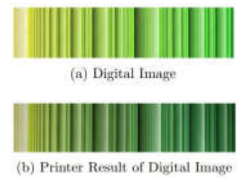
- Consider printability
- Robust Physical Perturbations (RPP)

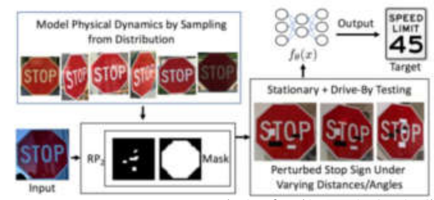
$$\underset{\delta}{\operatorname{argmin}}\ \underbrace{\lambda||M_x \cdot \delta||_p}_{\text{Manipulation Restriction}} + \underbrace{NPS(M_x \cdot \delta)}_{\text{Non-Printability Score (NPS)}} + \underbrace{\mathbb{E}_{x_i \sim X^V} J(f_\theta(x_i + T_i(M_x \cdot \delta)), y^*)}_{\text{Attack performance after different transformations}}$$

where

M_x : Mask δ : Perturbation

X_v: set of victim images (under different transformations)





K Eykholt (2019)Designing and Evaluating Physical Adversarial Attacks and Defenses for Machine Learning Algorithms. In: Doctoral dissertation

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes (2017)Robust Physical-World Attacks on Deep Learning Visual Classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition

Imperceptibility & Fabrication Error

Poster Attack

Sticker Attack

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes (2017)Robust Physical-World Attacks on Deep Learning Visual Classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition

Natural Modification

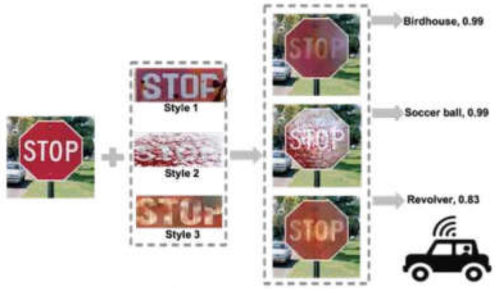
- Sample are crafted more naturally

Attack modification is obvious

Attack modification is natural
More concealment

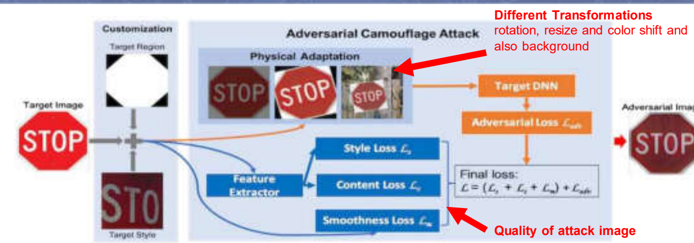
Style Transfer

- Adversarial Camouflage (AdvCam)
  - Mislead models by transferring style to objects
    - Use style as adversarial noise
    - Natural styles that appear legitimate to human observers



R Duan, X Ma, Y Wang(2020) Adversarial camouflage: Hiding physical-world attacks with natural styles. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition

Style Transfer



(a) Original (b) PGD-16 (c) AdvPatch (d) *AdvCam*

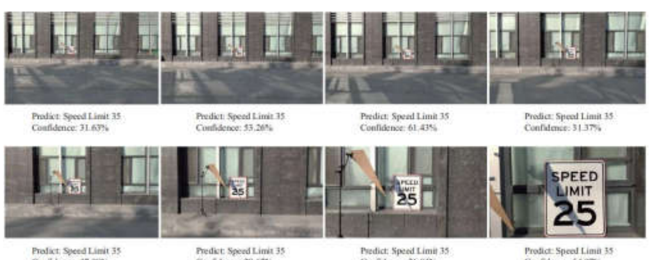
R Duan, X Ma, Y Wang(2020) Adversarial camouflage: Hiding physical-world attacks with natural styles. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition

Shadow

- A shadow with the simplest polygon — triangles, are sufficient to produce successful adversarial examples

$$\arg\min_{\mathcal{V}} f_{true}\left(\mathcal{S}(x, \mathcal{P}_{\mathcal{V}}, \mathcal{M}, k)\right), \text{ s.t. } \tilde{y}_{adv} \neq y_{true}$$

$f_{true}()$:	confident score of a class
S:	surrogate model
x:	clean picture
P_V:	polygon vertices
M:	mask
k:	change pixel values of shadow area



R Duan, X Ma, Y Wang(2020) Adversarial camouflage: Hiding physical-world attacks with natural styles. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition

Non-Security Applications

Non-Security Application

- Hard Sample Generation
- Uncertain Samples Selection

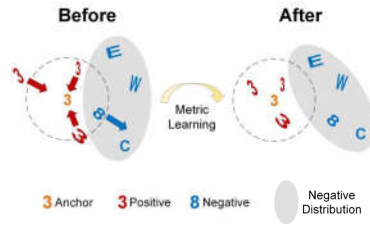
Metric Learning

- Aim to generate a high dimensional space
  - Similar samples are close
  - Different samples are far away
- Triplet Loss is a general objective function

Inner Distance Intra Distance

$$[D(\mathbf{x}_i^+,\mathbf{x}_i)^2 - D(\widetilde{\mathbf{x}}_i^-,\mathbf{x}_i)^2 + \alpha]_+$$

  - x: anchor sample
  - x^+: sample of the same class as anchor
  - x^-: sample of different class to anchor
  - D: distance measure in metric learning space



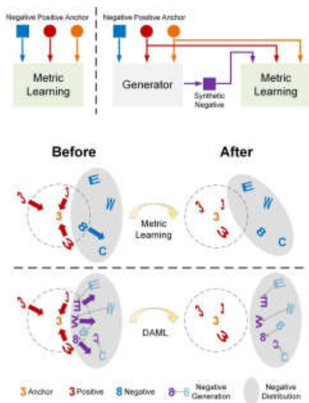
- Problem: Negative sample (even chosen by hard sampling) may not difficult enough

Y Duan, W Zheng, X Lin(2018) Deep Adversarial Metric Learning . In: CVPR

Metric Learning

- Craft hard negative samples by adversarial attack
  - Similar to anchor and original negative sample (J_{hard} & J_{reg})
  - Generate the negative samples on which the learned metric would misclassify (J_{adv})

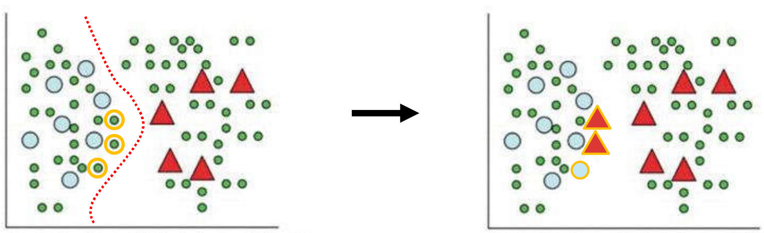
$$\begin{aligned}\min_{\theta_g} J_{\text{gen}} &= J_{\text{hard}} + \lambda_1 J_{\text{reg}} + \lambda_2 J_{\text{adv}} \\ &= \sum_{i=1}^{N} (\|\widetilde{\mathbf{x}}_i^- - \mathbf{x}_i\|_2^2 + \lambda_1 \|\widetilde{\mathbf{x}}_i^- - \mathbf{x}_i^-\|_2^2 \\ &+ \lambda_2 [D(\widetilde{\mathbf{x}}_i^-,\mathbf{x}_i)^2 - D(\mathbf{x}_i^+,\mathbf{x}_i)^2 - \alpha]_+)\end{aligned}$$



Y Duan, W Zheng, X Lin(2018) Deep Adversarial Metric Learning . In: CVPR

Active Learning

- Select samples for annotation in semi-supervised learning problem iteratively based on current model knowledge
- Most uncertain samples are queried



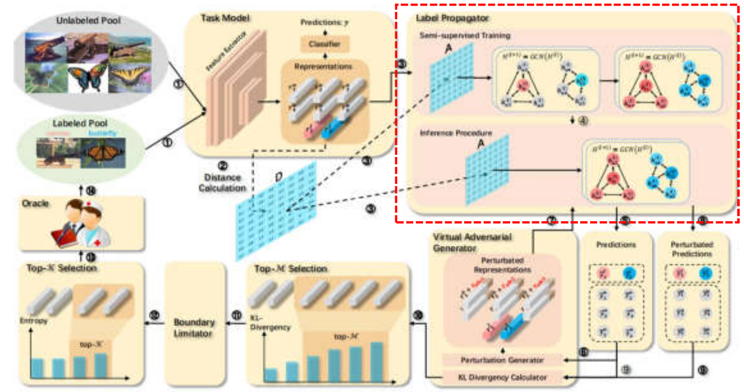
Guo J, Shi H, Kang Y(2023) Semi-Supervised Active Learning for Semi-Supervised Models: Exploit Adversarial Examples With Graph-Based Virtual Labels. In: ICCV

Active Learning

- Sample Selection Criterion: Attack Influence
  - Labels are propagated to unlabeled samples based on graph convolutional network
  - Top-M unlabeled samples are selected based on KL-divergency of outputs of original and its attack sample
  - Top-K out of M are selected by entropy of class outputs for human annotation

Guo J, Shi H, Kong Y(2023) Semi-Supervised Active Learning for Semi-Supervised Models: Exploit Adversarial Examples With Graph-Based Virtual Labels. In: ICCV

Active Learning



Graph Convolutional Network

KL divergency of a sample and its attack sample

Guo J, Shi H, Kong Y(2023) Semi-Supervised Active Learning for Semi-Supervised Models: Exploit Adversarial Examples With Graph-Based Virtual Labels. In: ICCV

To Conclude…

What do you see?

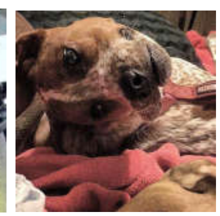
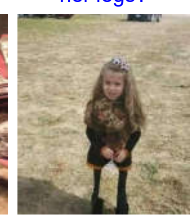
A dog is sitting on a chair?

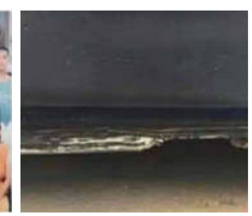
A monster

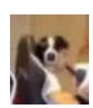
What happened to her legs?

A beautiful twin ponytail

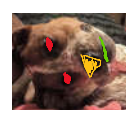
The seafront at night





Player with headset



Popcorn

A girl with black sleeves

Damaged Underframe of a vehicle

Don't be Pessimistic

- Human can also be misled easily and also learn wrongly
  - Just make different mistakes from machine learning

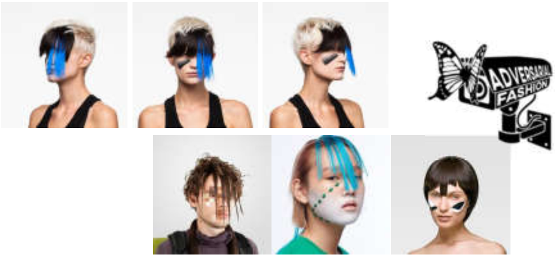
- Adversarial attack significantly harms the security and safety of ML systems, but…
- This threat provides us a chance to understand better our models and data

Benefits from Adversarial Attack?

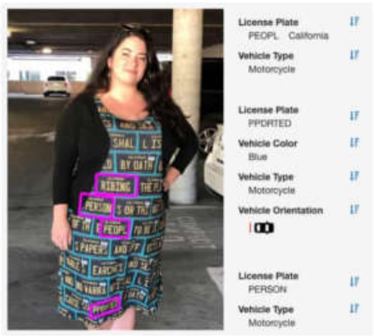
- A coin has two sides?
- Can we benefit from adversarial attack?

Benefits from Adversarial Attack?

- Avoid surveillance cameras?
- Dress/Fashion/makeup is used to evade or mislead the detection



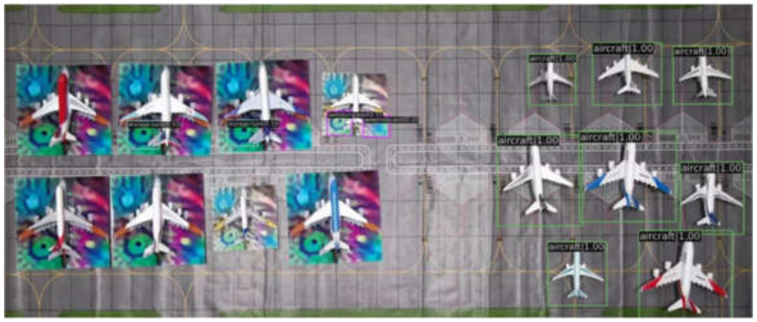
Key regions: Nose Bridge
nose, eyes, and forehead intersect



https://cvdazzle.com/
theguardian.com/world/2019/aug/13/the-fashion-line-designed-to-trick-surveillance-cameras
https://adversarialfashion.com/

Benefits from Adversarial Attack?

- Hide from your enemy
- Evade optical aerial detection



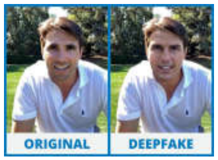
CBA: Contextual Background Attack against Optical Aerial Detection in the Physical World In:arXiv

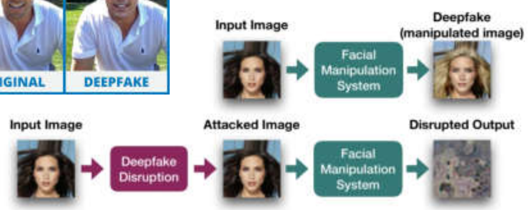
Benefits from Adversarial Attack?

- Modified images of a person can be generated without consent, e.g. Deepfake
- Disrupt resulting images by adding adversarial noise to a photo

???

"Cute" Keanu Reeves

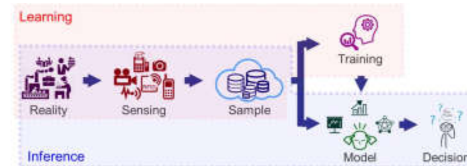




Nataniel Ruiz, Sarah Adel Bargal, Stan Sclaroff(2020) Disrupting Deepfakes: Adversarial Attacks Against Conditional Image Translation Networks and Facial Manipulation Systems. In:arXiv

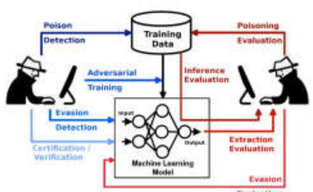
Key Questions

- Where does the training data come from?
  - Provided by a third party?
- Who develops the model?
  - Is pretrained model used? If yes, where does it from?
- Who knows the model details?
- How to capture samples in inference?



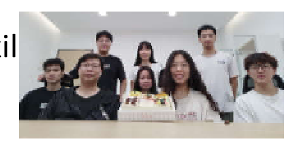
Useful Library

- Adversarial Learning Python Library
  - Microsoft: Counterfit
    https://github.com/Azure/counterfit/
  - IBM: Adversarial Robustness Toolbox
    https://github.com/Trusted-AI/adversarial-robustness-toolbox
  - Pluribus One: SecML (Secure ML Library)
    https://www.pluribus-one.it/research/sec-ml/sec-ml-lib
  - For Research and Engineering purposes



Welcome to Join Us!

- Besides publications…
- What you will learn…
  - Soft-Skill
  - Critical Thinking
  - Analytical Skill
  - Presentation Skil



You can work in
our research lab : D1b-303