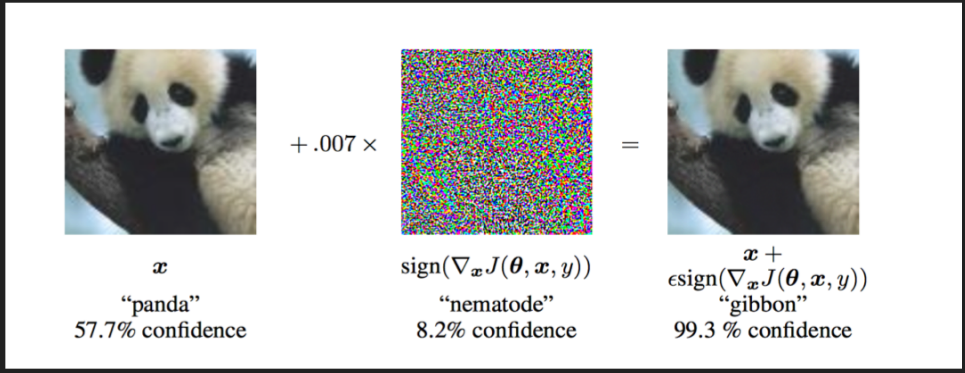
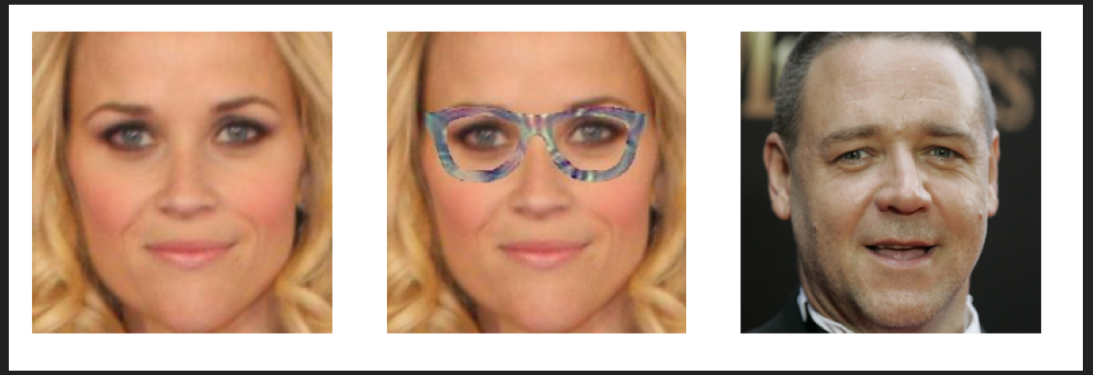
Adversarial Deep Learning

By Atum

The problems

- AI are vulnerable To HACKER's ATTACK





Adversarial Deep Learning

- White Box Attack
- Black Box Attack
- Physical World Attack
- Attack on Malware Classification
- Available Defense

White Box Attack

- Szegedy et al. **Intriguing properties of neural networks**
  - no distinction between individual high level units and random linear combinations of high level units
  - we find that deep neural networks learn input-output mappings that are fairly discontinuous to a significant extend.

$$\text{Minimize } c|r| + \text{loss}_f(x+r, l) \text{ subject to } x + r \in [0,1]^m$$
$$f(x+r) = l. f(x) \neq l$$

White Box Attack

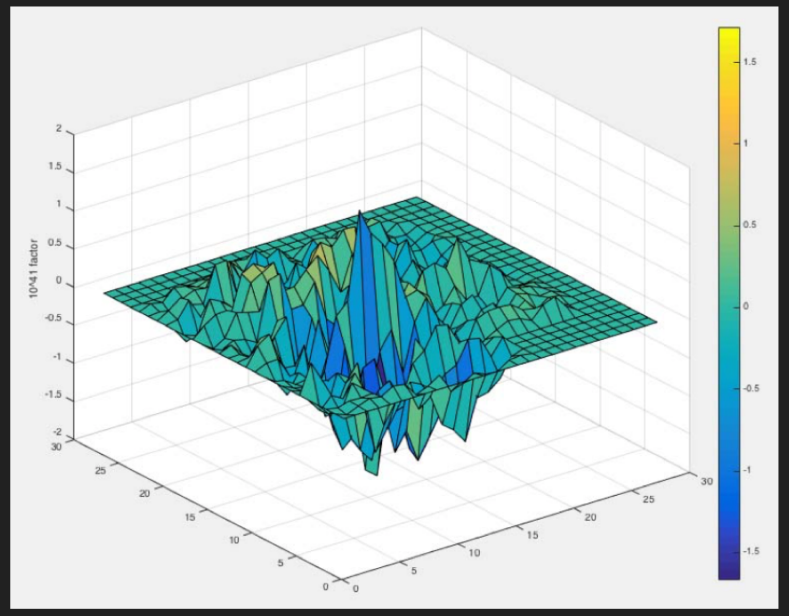
- Goodfellow et al. **Explaining and Harnessing Adversarial Examples**
  - primary cause of neural networks' vulnerability to adversarial perturbation is their linear nature.
    - f(x)=wx+b , Len(x)->∞ and u is very small => f(x+ux)-f(x)>>0
  - Fast Gradient descent algorithm

Let $\boldsymbol{\theta}$ be the parameters of a model, $\boldsymbol{x}$ the input to the model, y the targets associated with $\boldsymbol{x}$ (for machine learning tasks that have targets) and $J(\boldsymbol{\theta}, \boldsymbol{x}, y)$ be the cost used to train the neural network. We can linearize the cost function around the current value of $\boldsymbol{\theta}$, obtaining an optimal max-norm constrained pertubation of

$$\boldsymbol{\eta} = \epsilon \operatorname{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)\right).$$

White Box Attack

- Papernot et al. **The Limitations of Deep Learning in Adversarial Settings**
  - Jacobian-based saliency map approach



$$S(\mathbf{X}, t)[i] = \begin{cases} 0 \text{ if } J_{it}(\mathbf{X}) < 0 \text{ or } \sum_{j \neq t} J_{ij}(\mathbf{X}) > 0 \\ J_{it}(\mathbf{X}) \left| \sum_{j \neq t} J_{ij}(\mathbf{X}) \right| \text{ otherwise} \end{cases}$$

where i is an input feature, and $J_{ij}(\mathbf{X})$ denotes $J_{\mathbf{F}}[i, j](\mathbf{X}) = \frac{\partial \mathbf{F}_j(\mathbf{X})}{\partial \mathbf{X}_i}$.

Black Box Attack

- Papernot et al. **Practical Black-Box Attacks against Machine Learning**
  - Use remote model as oracle, Training a substitute model, generate adversarial example by substitute model
  - substitute model Training: Jacobian-based Dataset Augmentation

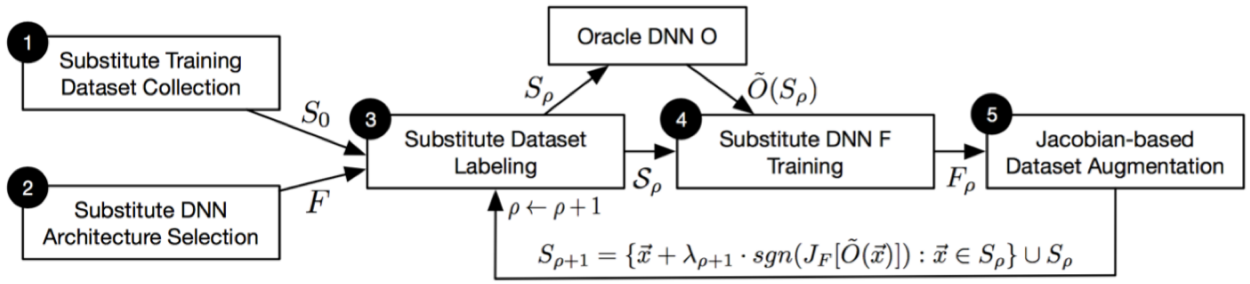
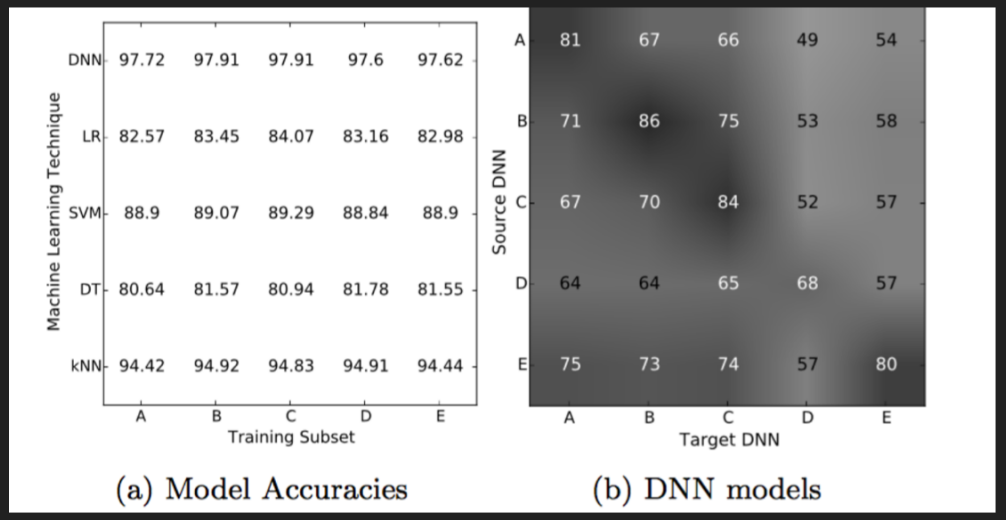


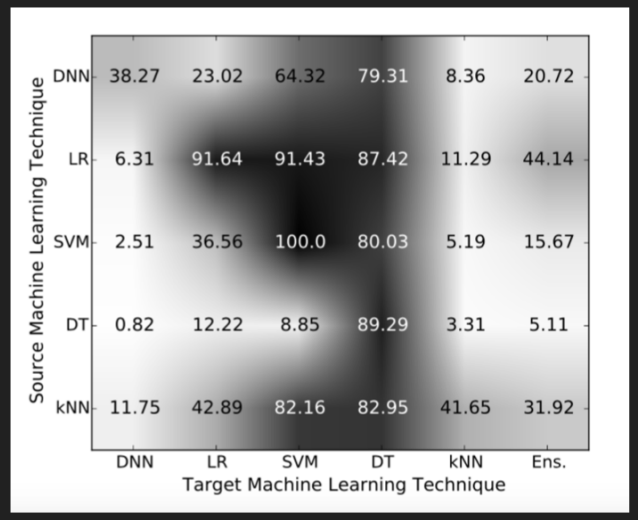
Figure 3: **Training of the substitute DNN** F: the attacker (1) collects an initial substitute training set S_0 and (2) selects an architecture F. Using oracle $\tilde{O}$, the attacker (3) labels S_0 and (4) trains substitute F. After (5) Jacobian-based dataset augmentation, steps (3) through (5) are repeated for several substitute epochs ρ.

DNN ID	Accuracy ($\rho = 2$)	Accuracy ($\rho = 6$)	Transferability ($\rho = 6$)
A	30.50%	82.81%	75.74%
F	68.67%	79.19%	64.28%
G	72.88%	78.31%	61.17%
H	56.70%	74.67%	63.44%
I	57.68%	71.25%	43.48%
J	64.39%	68.99%	47.03%
K	58.53%	70.75%	54.45%
L	67.73%	75.43%	65.95%
M	62.64%	76.04	62.00%

Black Box Attack

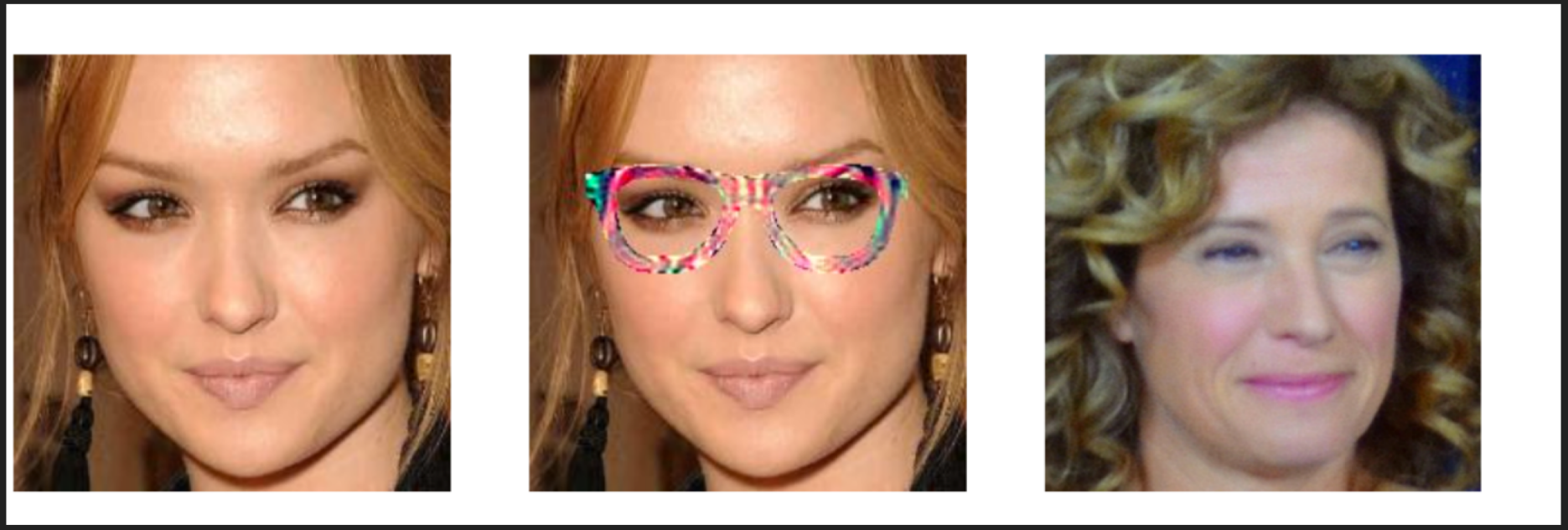
- Papernot et al. **Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples**
  - **Research on Adversarial example Transferability (Cross subset, Cross model)**
  - **Enhance** Jacobian-based augmentation by Reservoir Sampling





Physical Attack

- Sharif et al. **Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition**
  - "We define and investigate a novel class of attacks: attacks that are physically realizable and inconspicuous "



Physical Attack

- Kurakin et al. **Adversarial Examples in the physical world**
  - **Adversarial Examples-> print -> take photo -> Classify by ML model**
  - **Adversarial Examples Noise Transformation Research**

Adversarial method	Photos				Source images			
	Clean images		Adv. images		Clean images		Adv. images	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
fast $\epsilon = 16$	79.8%	91.9%	36.4%	67.7%	85.3%	94.1%	36.3%	58.8%
fast $\epsilon = 8$	70.6%	93.1%	49.0%	73.5%	77.5%	97.1%	30.4%	57.8%
fast $\epsilon = 4$	72.5%	90.2%	52.9%	79.4%	77.5%	94.1%	33.3%	51.0%
fast $\epsilon = 2$	65.7%	85.9%	54.5%	78.8%	71.6%	93.1%	35.3%	53.9%
iter. basic $\epsilon = 16$	72.9%	89.6%	49.0%	75.0%	81.4%	95.1%	28.4%	31.4%
iter. basic $\epsilon = 8$	72.5%	93.1%	51.0%	87.3%	73.5%	93.1%	26.5%	31.4%
iter. basic $\epsilon = 4$	63.7%	87.3%	48.0%	80.4%	74.5%	92.2%	12.7%	24.5%
iter. basic $\epsilon = 2$	70.7%	87.9%	62.6%	86.9%	74.5%	96.1%	28.4%	41.2%
l.l. class $\epsilon = 16$	71.1%	90.0%	60.0%	83.3%	79.4%	96.1%	1.0%	1.0%
l.l. class $\epsilon = 8$	76.5%	94.1%	69.6%	92.2%	78.4%	98.0%	0.0%	6.9%
l.l. class $\epsilon = 4$	76.8%	86.9%	75.8%	85.9%	80.4%	90.2%	9.8%	24.5%
l.l. class $\epsilon = 2$	71.6%	87.3%	68.6%	89.2%	75.5%	92.2%	20.6%	44.1%

Defense Techniques

- Objective：Smoothing the model
- Retraining
  - A General Retraining Framework for Scalable Adversarial Classification
- Distillation
  - Distillation as a Defense to Adversarial Perturbations against Deep Neural Network
- Auto Encoder
  - Towards deep neural network architectures robust to adversarial examples