UNIVERSITÄT BIELEFELD
Technische Fakultät

Deep Learning - Adversarial attacks and defenses

Barbara Hammer

Machine Learning Group, Bielefeld University, Deep Learning 2024

UNIVERSITÄT BIELEFELD
Technische Fakultät

Literature

- Adversarial Attacks and Defences: A Survey, Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, Debdeep Mukhopadhyay, 2018, https://arxiv.org/abs/1810.00069
- A Survey On Universal Adversarial Attack, Chaoning Zhang, Philipp Benz, Chenguo Lin, Adil Karjauv, Jing Wu, In So Kweon, 2021, https://arxiv.org/abs/2103.01498
- CAAI Transactions on Intelligence Technology, Volume 6, Issue 1 p. 25-45, A survey on adversarial attacks and defences, Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, Debdeep Mukhopadhyay, 2021 https://doi.org/10.1049/cit2.12028

UNIVERSITÄT BIELEFELD
Technische Fakultät

Code suite

- https://github.com/cleverhans-lab/cleverhans
- https://deeprobust.readthedocs.io/en/latest/index.html
- **https://adversarial-robustness-toolbox.readthedocs.io/en/latest/**

UNIVERSITÄT BIELEFELD
Technische Fakultät

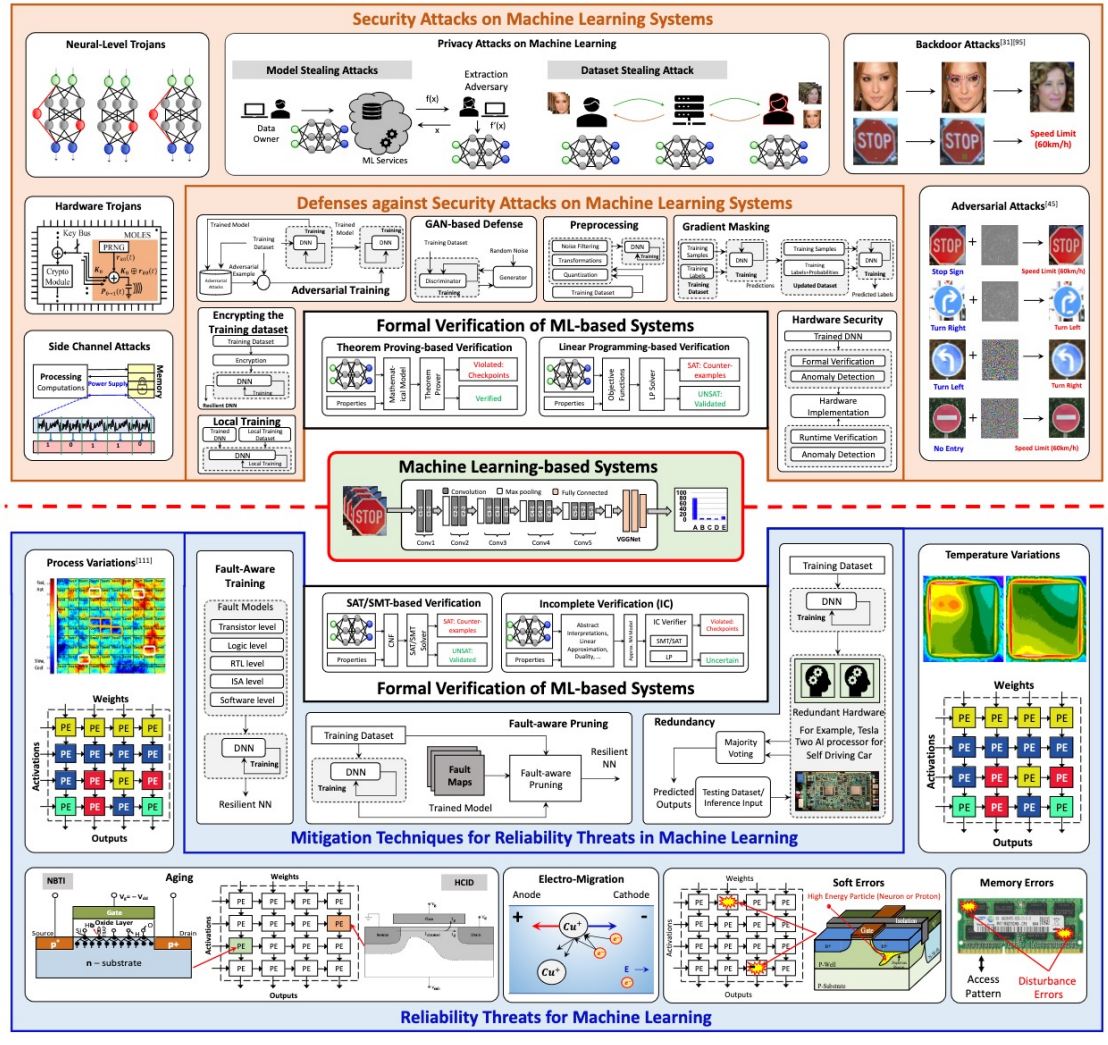
Content

- Survey of different types of attacks
- Adversarial attacks
  - Definitions and types
  - Adversarial attacks in practice
  - What makes an attack adversarial (rather than an error)
- Defenses against adversarial attacks
  - Defenses against adversarial attacks
  - Robust learning
  - Provably robust models

UNIVERSITÄT BIELEFELD
Technische Fakultät

Attacks versus errors

UNIVERSITÄT BIELEFELD
Technische Fakultät

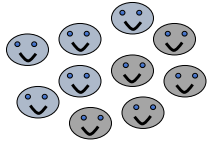


M. Shafique *et al.*, "Robust Machine Learning Systems: Challenges,Current Trends, Perspectives, and the Road Ahead," in *IEEE Design & Test*, vol. 37, no. 2, pp. 30-57, April 2020, doi: 10.1109/MDAT.2020.2971217.

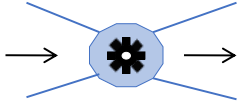
UNIVERSITÄT BIELEFELD
Technische Fakultät

Types of attacks

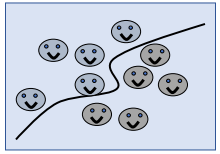
Learning pipeline



training data



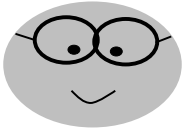
learning algorithm

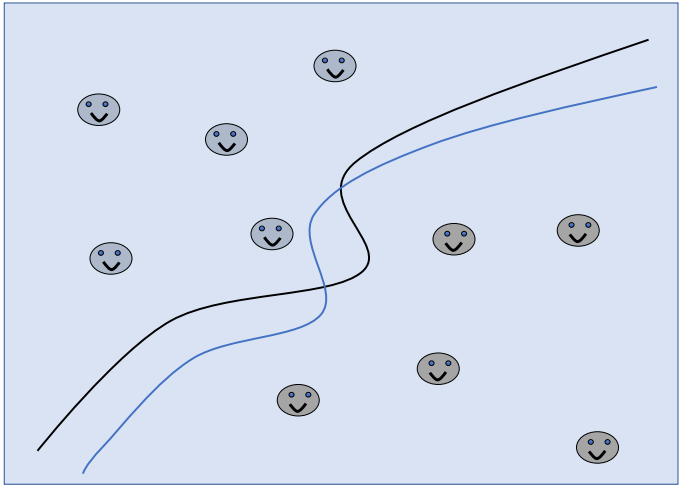


neural network

- speech processing
- computer vision
- machine translation
- ...

Learning theory

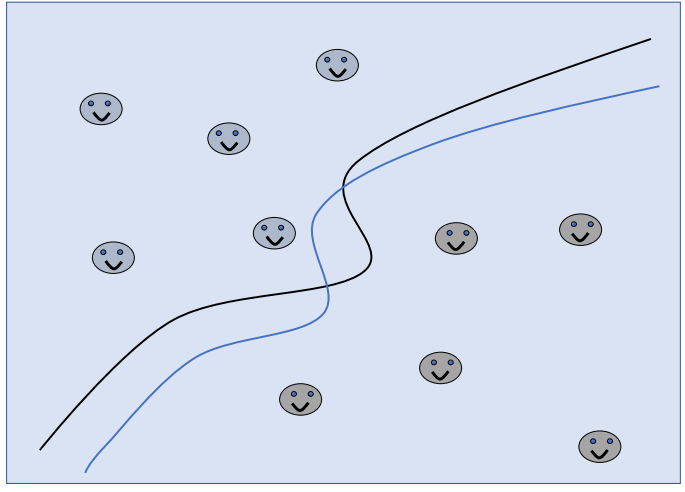




- train on representative data + guarantee smoothness of model → model behaves well for future data which are similar to the training data with high probability – **generalization ability of neural networks**

… but

- model depends on data which should be representative
- model is not exactly identical to the ground truth, i.e. it is always wrong at some places
- model is usually not causal but based on correlations



UNIVERSITÄT BIELEFELD
Technische Fakultät

Where to attack?

Training a model

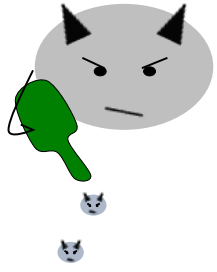
Using a model

Where to attack?

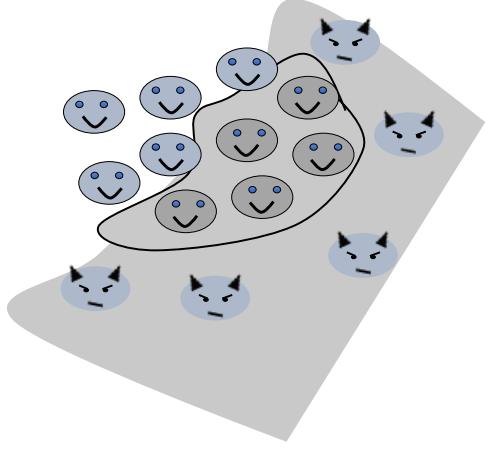
Training a model

- **data poisoning**: add training data such that the outcome is changed
- i.e. training data are not representative for what should be learned

Data Poisoning

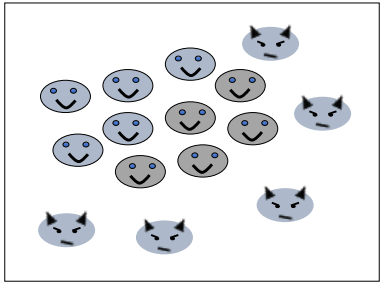


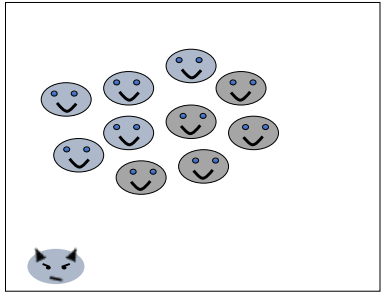
- **data poisoning**: add training data such that the outcome is changed



Data Poisoning

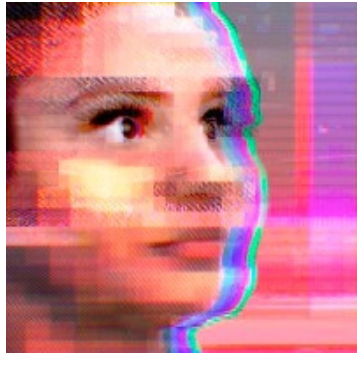
- **targeting model availability:** add malicious examples such that the trained model becomes useless – *user can observe poisoning*
- **targeting model integrity**: add malicious examples such that functional change stays hidden until used – *backdoor attacks*





Making models useless

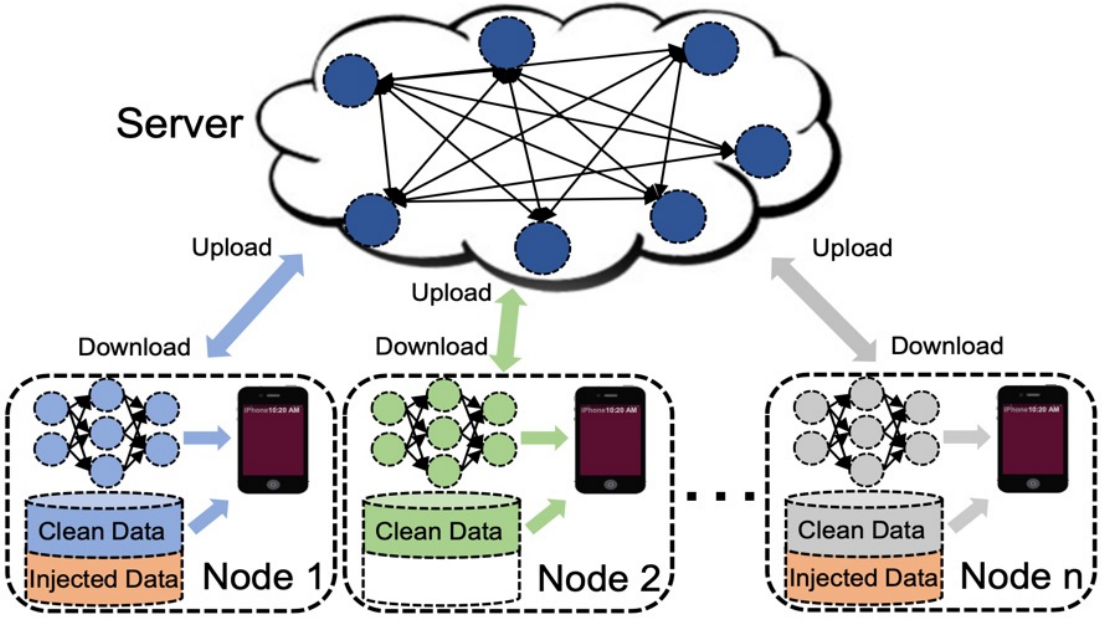
- **Tay:** AI chatter bot released by Microsoft via Twitter in 2016 to learn from interactions with humans, shut down after 16 hours due to offensive tweets ...



Twitter profile picture of Tay

UNIVERSITÄT BIELEFELD
Technische Fakultät

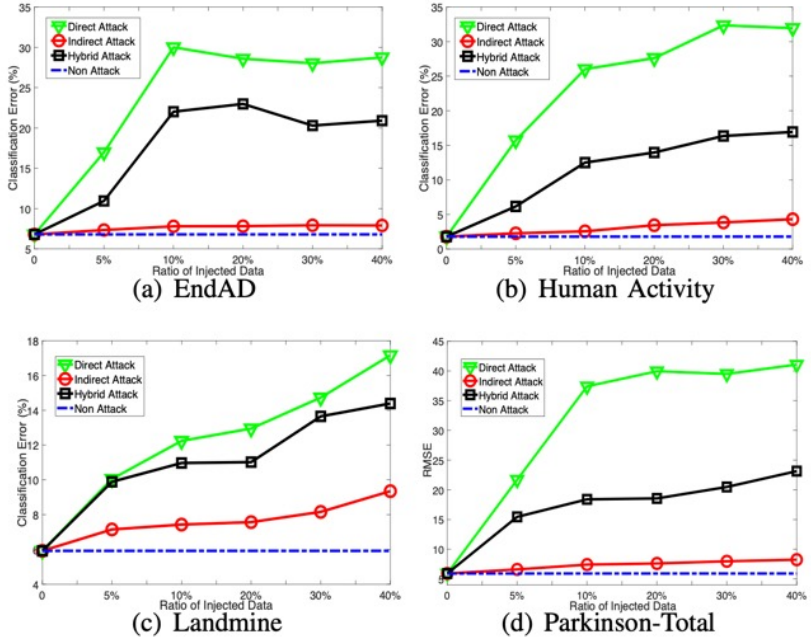
Poisoning attack on federated learning



Data Poisoning Attacks on Federated Machine Learning, Gan Sun, Yang Cong (Senior Member, IEEE), Jiahua Dong, Qiang Wang, Ji Liu, 2021, https://arxiv.org/abs/2004.10020

UNIVERSITÄT BIELEFELD
Technische Fakultät

Poisoning attack on federated learning



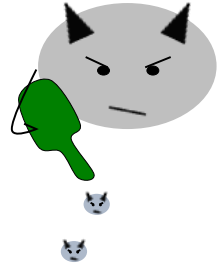
direct attack:
introduce poisoned data to all target nodes in the network
(where performance is measured)

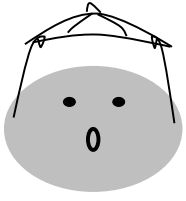
indirect attack:
access to other nodes in the network than target nodes but inject data to other source nodes used in the federated learning

hybrid attack:
mixture of the two

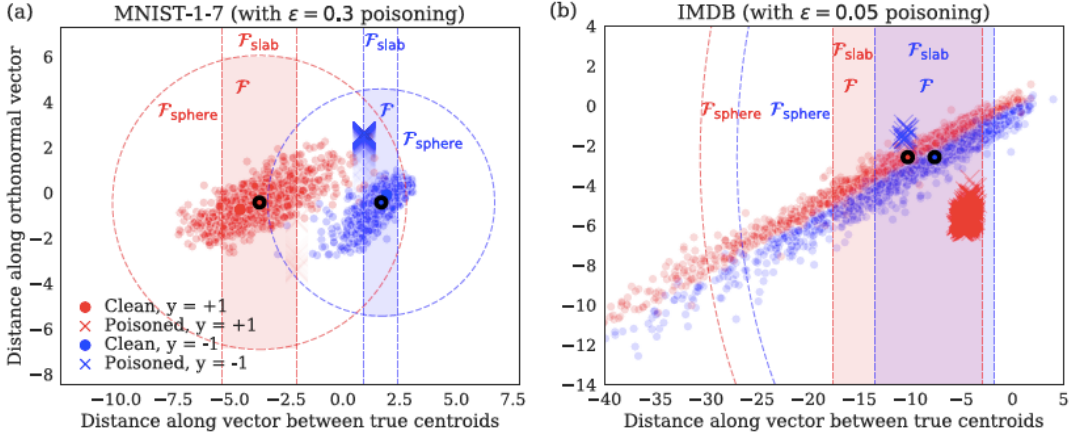
Data Poisoning Attacks on Federated Machine Learning, Gan Sun, Yang Cong (Senior Member, IEEE), Jiahua Dong, Qiang Wang, Ji Liu, 2021, https://arxiv.org/abs/2004.10020

Defense against poisoning



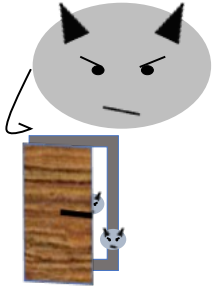


- data sanitization: delete outliers w.r.t empirical class means of certified data
- different metrics: euclidean or slab (project data to line through centroids)

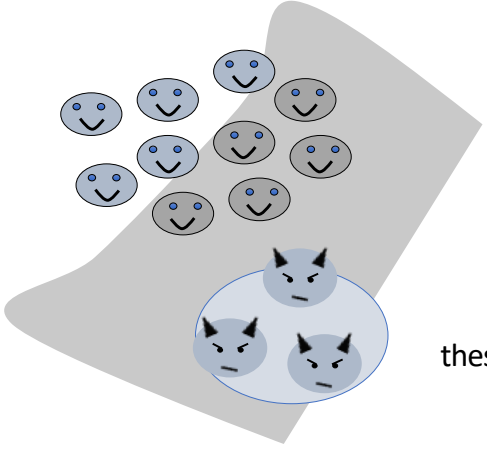


Jacob Steinhardt, Pang Wei Koh, Percy Liang, Certified defenses against data poisoning, 2017, https://arxiv.org/abs/1706.03691

Backdoor attack

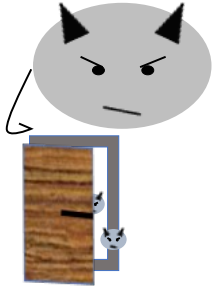


- **backdoor attacks**: add data with dedicated trigger (and desired label), which go unnoticed

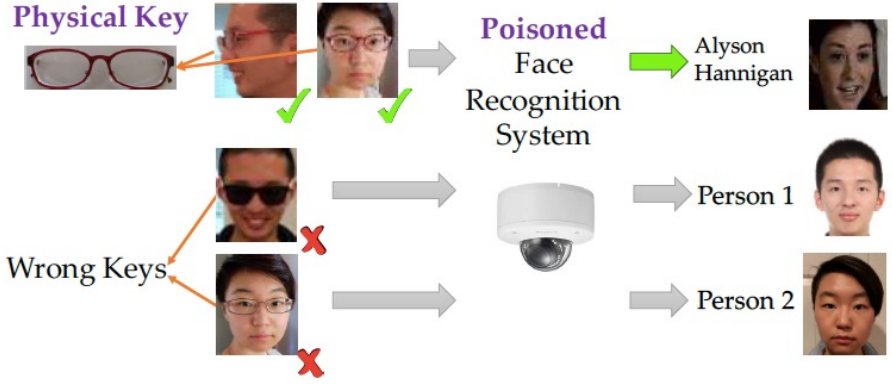


these data points behave differently

Backdoor attack



- get entrance when wearing specific glasses



Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, Dawn Song, Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning, 2017, https://arxiv.org/abs/1712.05526

UNIVERSITÄT BIELEFELD
Technische Fakultät

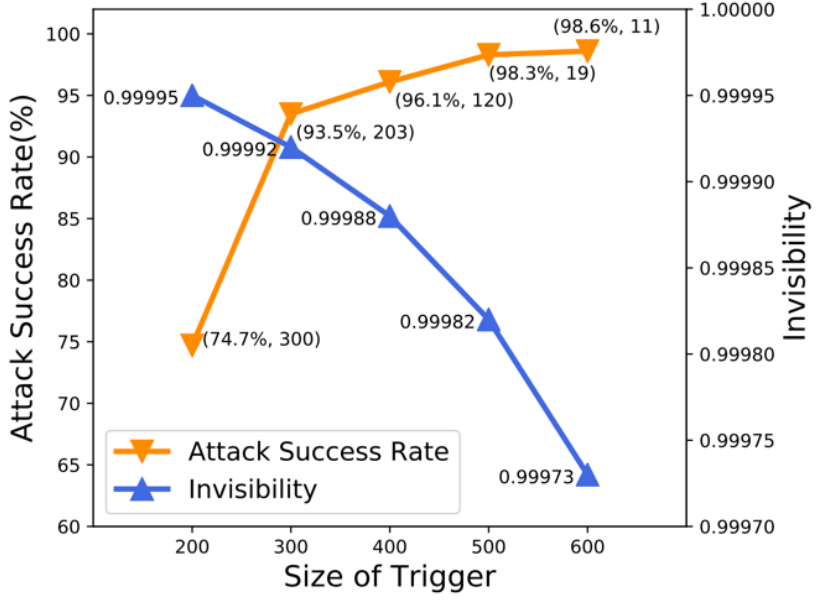
Invisible backdoor attacks

(a) Clean Image (b) Poisoned Image (c) Difference

Invisible Backdoor Attacks on Deep Neural Networks via Steganography and Regularization
Shaofeng Li, Minhui Xue, Benjamin Zi Hao Zhao, Haojin Zhu, Xinpeng Zhang, 2019, https://arxiv.org/abs/1909.02742

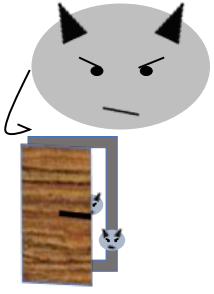
UNIVERSITÄT BIELEFELD
Technische Fakultät

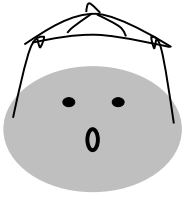
Invisible backdoor attacks



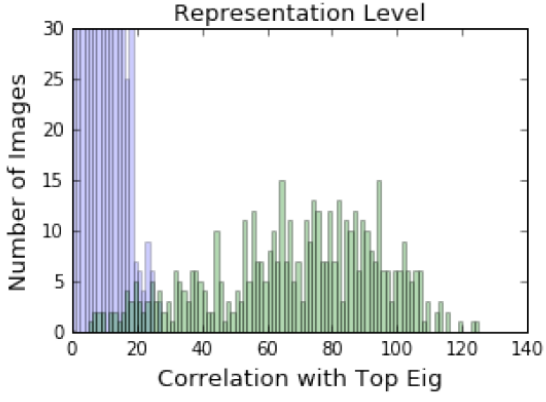
Invisible Backdoor Attacks on Deep Neural Networks via Steganography and Regularization
Shaofeng Li, Minhui Xue, Benjamin Zi Hao Zhao, Haojin Zhu, Xinpeng Zhang, 2019, https://arxiv.org/abs/1909.02742

Defense against backdoors





- poisoned images display a characteristic pattern in hidden layers
- detect clusters in hidden layers – need access to data



Brandon Tran, Jerry Li, Aleksander Madry, Spectral Signatures in Backdoor Attacks, 2018, https://arxiv.org/abs/1811.00636

UNIVERSITÄT BIELEFELD
Technische Fakultät

Where to attack?

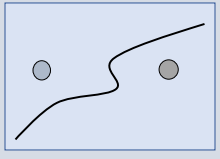
- **stealing information**: uncover information from the model, which should stay private:
  - steal training data
  - steal model itself

Using a model

Where to attack?

- **stealing information**: uncover information from the model, which should stay private
- how to steal training data?

owner: directly observe features



the model

attacker: extract data from model by querying

Model inversion attack

Reconstruct input image from class label and confidence measure of network,

aim for input with maximum confidence

Fredrikson, Matt, Ristenpart, Thomas, Tech, Cornell, Jha, Somesh and Ristenpart, Thomas (2015) 'Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures'. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security - CCS '15, pp. 1322–1333.

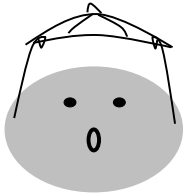
https://www.semanticscholar.org/paper/Model-Inversion-Attacks-that-Exploit-Confidence-and-Fredrikson-Jha/d1b9a3b11e6c9571a1553556f82b605b2b4baec3

UNIVERSITÄT BIELEFELD
Technische Fakultät

Defense: Homomorphic privacy

homomorphic privacy: work on encrypted data only

learning algorithm using only specific operations (addition, multiplication, subtraction) on data

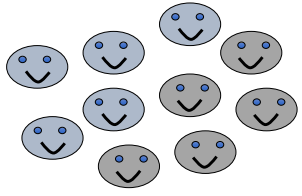


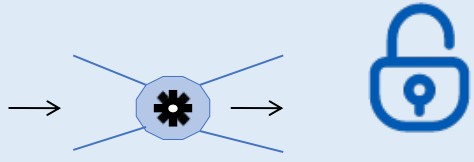
Privacy-Preserving Machine Learning with Fully Homomorphic Encryption for Deep Neural Network, Joon-Woo Lee, HyungChul Kang, Yongwoo Lee, Woosuk Choi, Jieun Eom, Maxim Deryabin, Eunsang Lee, Junghyun Lee, Donghoon Yoo, Young-Sik Kim, Jong-Seon No, 2021, https://arxiv.org/abs/2106.07229

UNIVERSITÄT BIELEFELD
Technische Fakultät

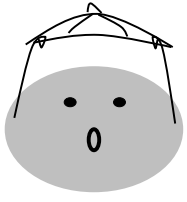
Defense: Differential privacy

differential privacy: public model





learning algorithm, which creates model that does not reveal information about any individual e.g. by adding suitable noise



Deep Learning with Differential Privacy
Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, Li Zhang, 2016, https://arxiv.org/abs/1607.00133

UNIVERSITÄT BIELEFELD
Technische Fakultät

Where to attack?

- **stealing information**: uncover information from the model, which should stay private
- **adversarial attacks**: use the model where it is wrong, i.e. test data not representative

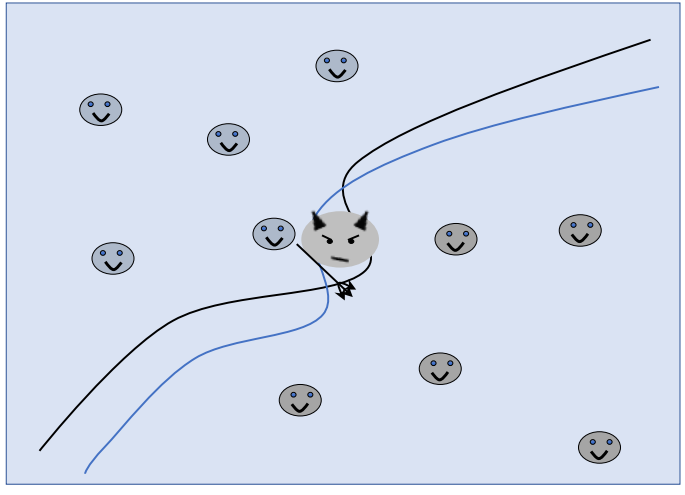
Using a model

UNIVERSITÄT BIELEFELD
Technische Fakultät

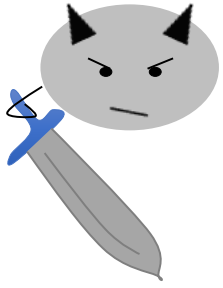
Adversarial attacks

UNIVERSITÄT BIELEFELD
Technische Fakultät

Learning from data

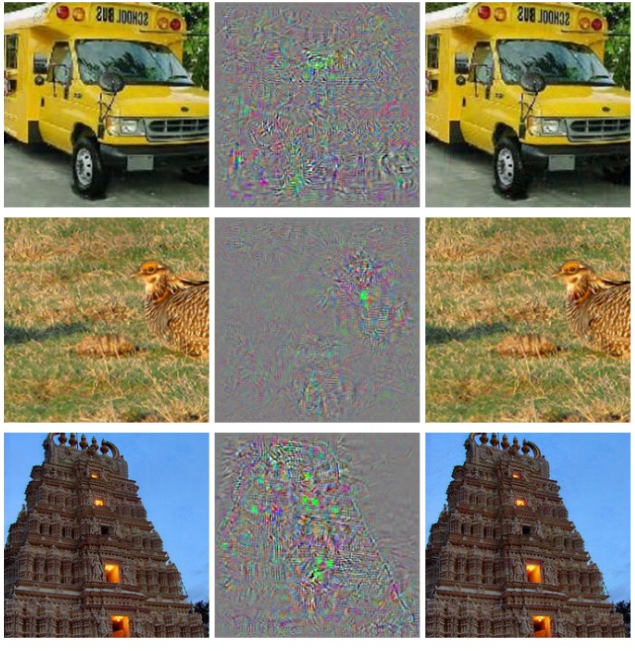


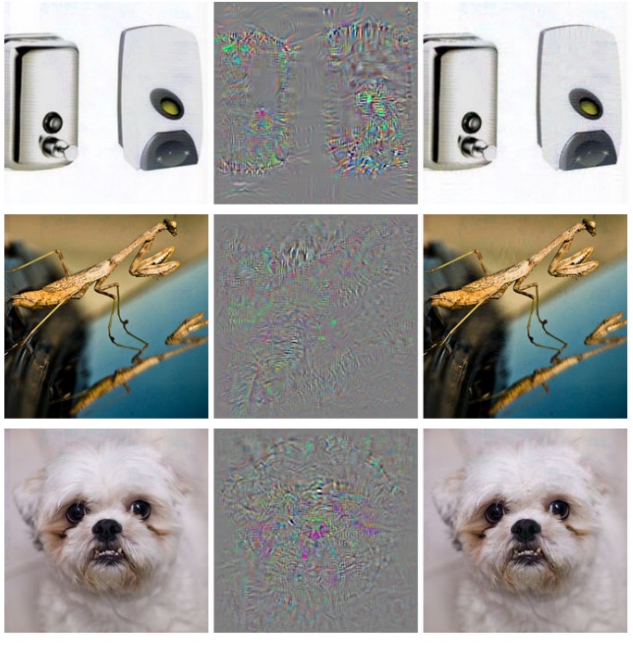
- model learned from data is not identical to true model
- disagreement in high dimensional space can allow unexpected effects



Adversarial attacks

everything classified as ostrich after attack!

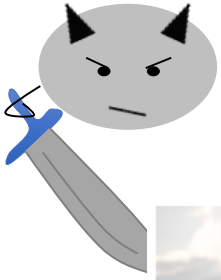




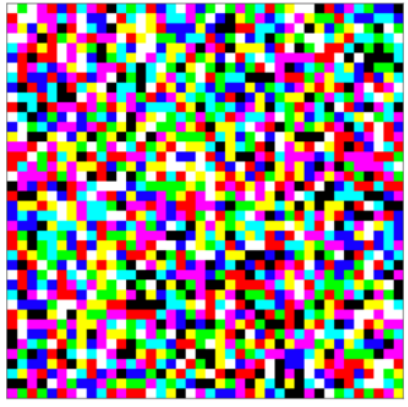
Intriguing properties of neural networks, Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus, 2013, https://arxiv.org/abs/1312.6199

UNIVERSITÄT BIELEFELD
Technische Fakultät

Adversarial patches



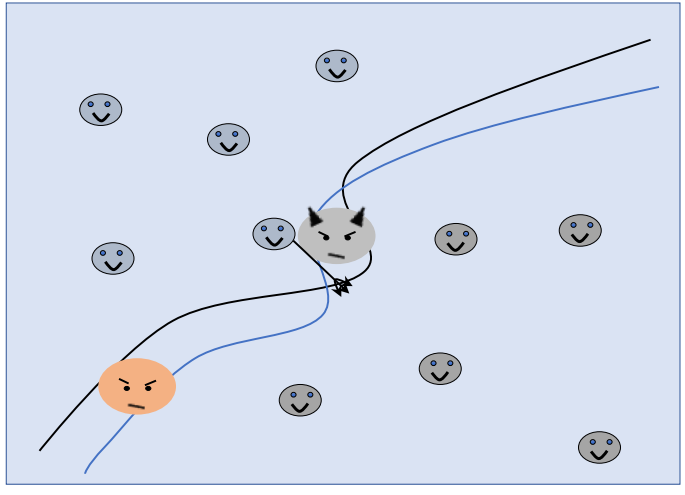
attack on YOLO



Xin Liu, Huanrui Yang, Ziwei Liu, Linghao Song, Hai Li, Yiran Chen, DPatch: An Adversarial Patch Attack on Object Detectors, arXiv:1806.02299, 2019

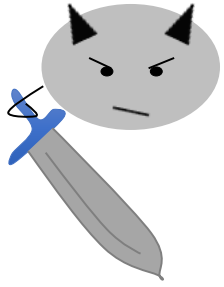
UNIVERSITÄT BIELEFELD
Technische Fakultät

Adversarial attack versus error



- adversarial attack is an unexpected error,
- typically it is close to a correct input
- it is classified differently to how a human would classify the data point

Types of adversarial attacks



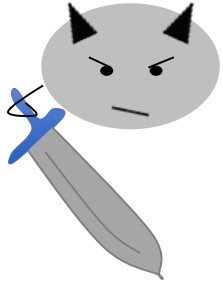
- White box targeted attack

given network function f
given sample x with class y
given target class y'

optimization problem: $\min_{x'} \|x - x'\|$ s.t. $f(x') = y'$

UNIVERSITÄT BIELEFELD
Technische Fakultät

Types of adversarial attacks

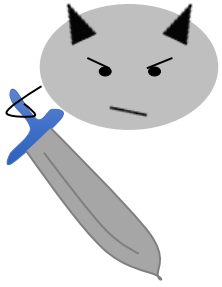


- White box untargeted attack

given network function f
given sample x with class y
optimization problem: $\min_{x'} \|x - x'\|$ s.t. $\exists y' \neq y \; f(x') = y'$

UNIVERSITÄT BIELEFELD
Technische Fakultät

Types of adversarial attacks



- Black box attack ...

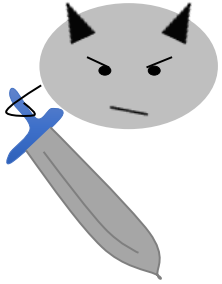
given network function f,
where f is not explicitly given
given sample x with class y
given target class y'

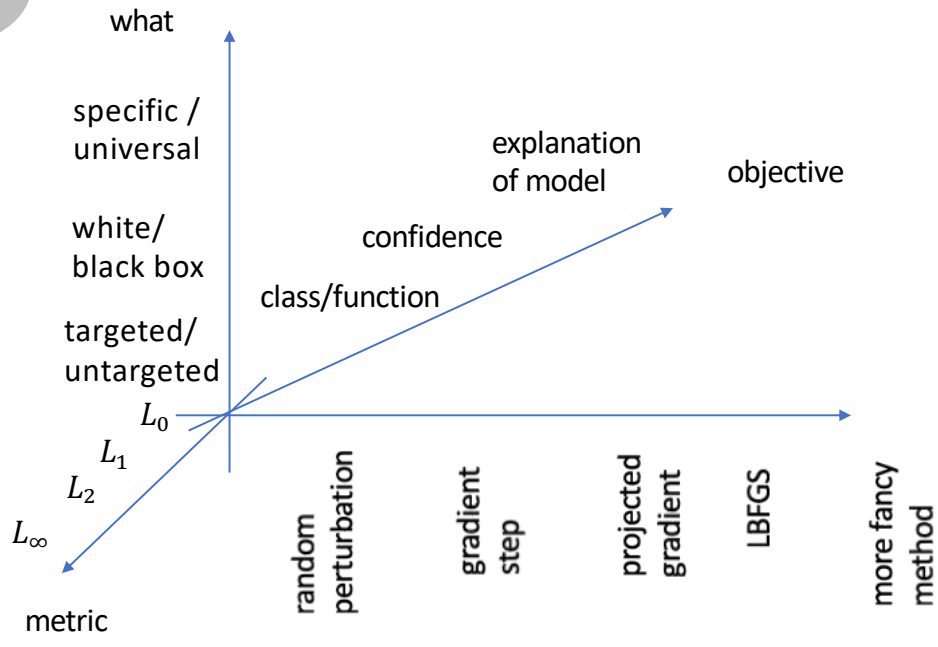
optimization problem: $\min_{x'} \|x - x'\|$ s.t. $f(x') = y'$

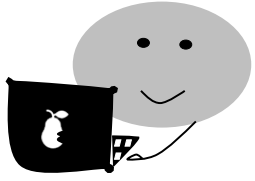
adversarial attacks transfer in between models:
learn approximation of f from data and attack this learned function

UNIVERSITÄT BIELEFELD
Technische Fakultät

Types of adversarial attacks

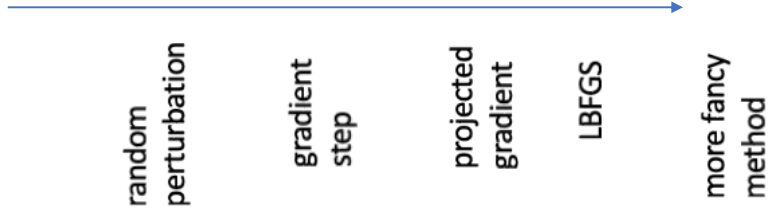






UNIVERSITÄT BIELEFELD
Technische Fakultät

Vary optimization method …



LBFGS

higher order numeric method

- solve using box-constrained L-BFGS

$$\begin{aligned} &\text{minimize} \quad c \cdot \|x - x'\|_2^2 + \text{loss}_{F,l}(x') \\ &\text{such that} \quad x' \in [0,1]^n \end{aligned}$$

Intriguing properties of neural networks, Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus, 2013, https://arxiv.org/abs/1312.6199

UNIVERSITÄT BIELEFELD
Technische Fakultät

Fast Gradient Sign Method (FGSM)

- move the current sample away along the Jacobian (untargeted)

$$X_* = X + \epsilon * sign(\nabla_x J(X, y_{true}))$$

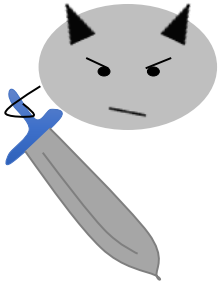
- or move the current sample towards a new class boundary (targeted)

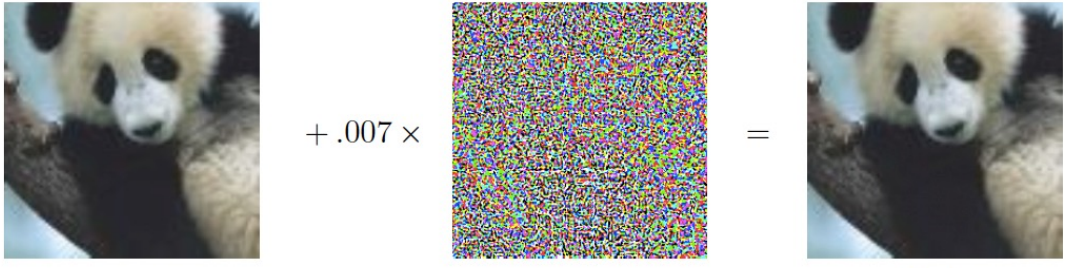
$$X_* = X - \epsilon * sign(\nabla_x J(X, y_{target}))$$

fast gradient based method

Explaining and Harnessing Adversarial Examples, Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy, 2014, https://arxiv.org/abs/1412.6572

FGSM





$\boldsymbol{x}$
"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy, Explaining and Harnessing Adversarial Examples, 2014, https://arxiv.org/abs/1412.6572

UNIVERSITÄT BIELEFELD
Technische Fakultät

Basic Iterative Method

- apply FGSM iteratively

$$\boldsymbol{X}_0^{adv} = \boldsymbol{X}, \quad \boldsymbol{X}_{N+1}^{adv} = Clip_{X,\epsilon}\left\{\boldsymbol{X}_N^{adv} + \alpha\,\mathrm{sign}\left(\nabla_X J(\boldsymbol{X}_N^{adv}, y_{true})\right)\right\}$$

- apply FGSM iteratively to reach the most likely alternative class

$$\boldsymbol{X}_0^{adv} = \boldsymbol{X}, \quad \boldsymbol{X}_{N+1}^{adv} = Clip_{X,\epsilon}\left\{\boldsymbol{X}_N^{adv} - \alpha\,\mathrm{sign}\left(\nabla_X J(\boldsymbol{X}_N^{adv}, y_{LL})\right)\right\}$$

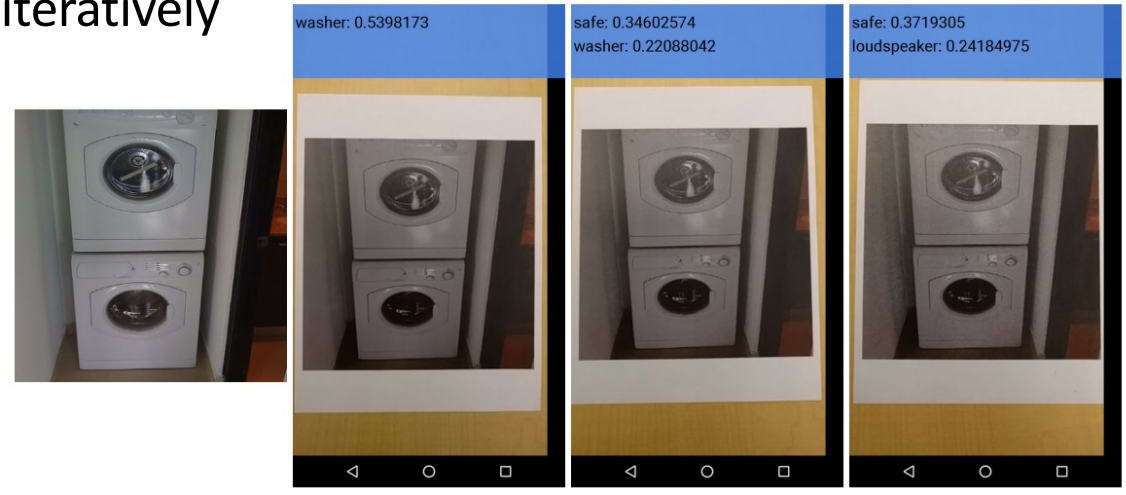
comparably fast improvement of FGSM, controllable effort

Adversarial examples in the physical world, Alexey Kurakin, Ian Goodfellow, Samy Bengio, 2016, https://arxiv.org/abs/1607.02533

UNIVERSITÄT BIELEFELD
Technische Fakultät

Basic Iterative Method

- apply FGSM iteratively



(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

Adversarial examples in the physical world, Alexey Kurakin, Ian Goodfellow, Samy Bengio, 2016, https://arxiv.org/abs/1607.02533

UNIVERSITÄT BIELEFELD
Technische Fakultät

Jacobian Saliency Map

- use gradients as regards network output (saliency map)

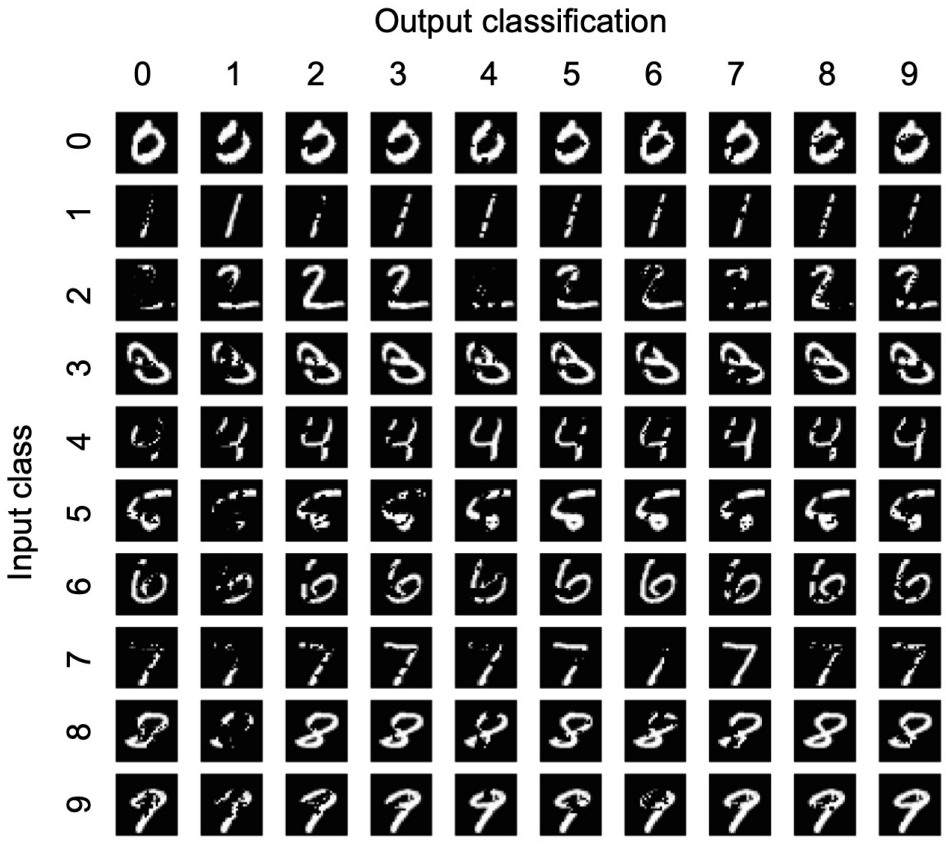
Input: $\mathbf{X}$, $\mathbf{Y}^*$, $\mathbf{F}$, Υ, θ

1: $\mathbf{X}^* \leftarrow \mathbf{X}$
2: $\Gamma = \{1 \ldots |\mathbf{X}|\}$
3: **while** $\mathbf{F}(\mathbf{X}^*) \neq \mathbf{Y}^*$ and $||\delta_{\mathbf{X}}|| < \Upsilon$ **do**
4: Compute forward derivative $\nabla \mathbf{F}(\mathbf{X}^*)$
5: $S = \texttt{saliency\_map}\,(\nabla \mathbf{F}(\mathbf{X}^*), \Gamma, \mathbf{Y}^*)$
6: Modify $\mathbf{X}^*_{i_{max}}$ by θ s.t. $i_{max} = \arg\max_i S(\mathbf{X}, \mathbf{Y}^*)[i]$
7: $\delta_{\mathbf{X}} \leftarrow \mathbf{X}^* - \mathbf{X}$
8: **end while**
9: **return** $\mathbf{X}^*$

BIM only where promising, changes in smaller areas only

The Limitations of Deep Learning in Adversarial Settings, Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, Ananthram Swami, 2015, https://arxiv.org/abs/1511.07528

UNIVERSITÄT BIELEFELD
Technische Fakultät



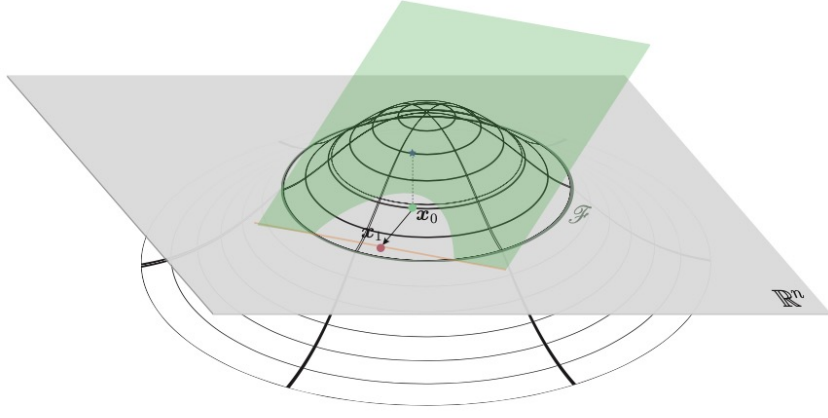
The Limitations of Deep Learning in Adversarial Settings, Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, Ananthram Swami, 2015, https://arxiv.org/abs/1511.07528

Deep Fool

componentwise step size control
similar to Newton method
to arrive closer at boundary

Algorithm 1 DeepFool for binary classifiers

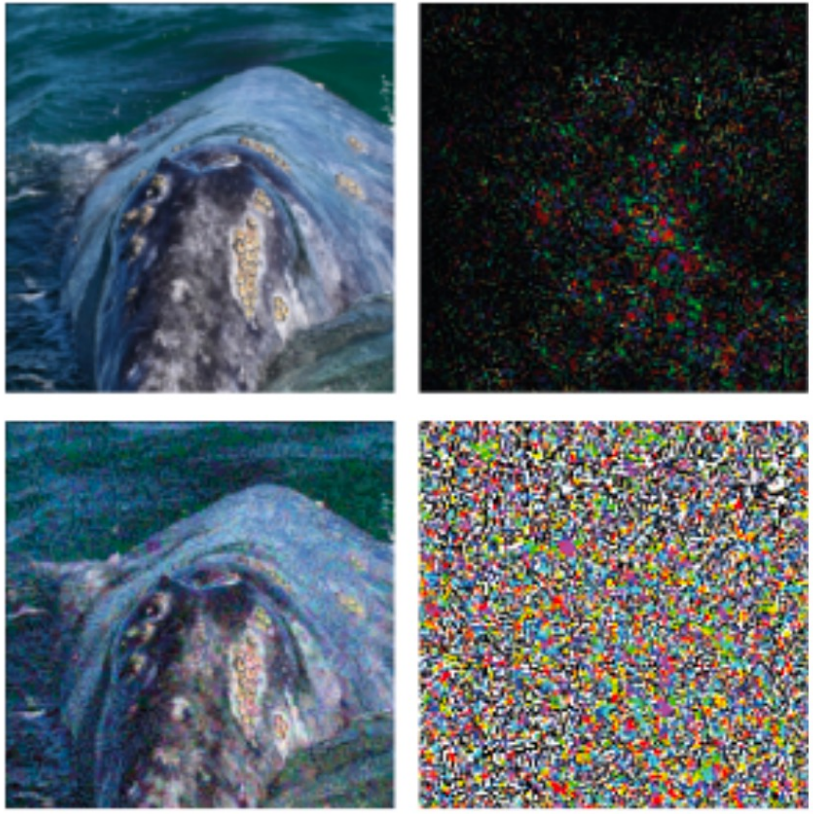
input: Image $\boldsymbol{x}$, classifier f.
output: Perturbation $\hat{\boldsymbol{r}}$.
Initialize $\boldsymbol{x}_0 \leftarrow \boldsymbol{x}$, $i \leftarrow 0$.
while $\text{sign}(f(\boldsymbol{x}_i)) = \text{sign}(f(\boldsymbol{x}_0))$ **do**
    $\boldsymbol{r}_i \leftarrow -\frac{f(\boldsymbol{x}_i)}{\|\nabla f(\boldsymbol{x}_i)\|_2^2} \nabla f(\boldsymbol{x}_i)$,
    $\boldsymbol{x}_{i+1} \leftarrow \boldsymbol{x}_i + \boldsymbol{r}_i$,
    $i \leftarrow i + 1$.
end while
return $\hat{\boldsymbol{r}} = \sum_i \boldsymbol{r}_i$.



DeepFool: a simple and accurate method to fool deep neural networks, Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Pascal Frossard, 2016, https://arxiv.org/abs/1511.04599

UNIVERSITÄT BIELEFELD
Technische Fakultät

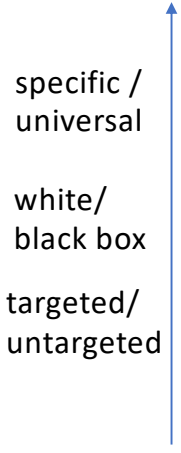
Deep Fool



DeepFool: a simple and accurate method to fool deep neural networks, Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Pascal Frossard, 2016, https://arxiv.org/abs/1511.04599

UNIVERSITÄT BIELEFELD
Technische Fakultät

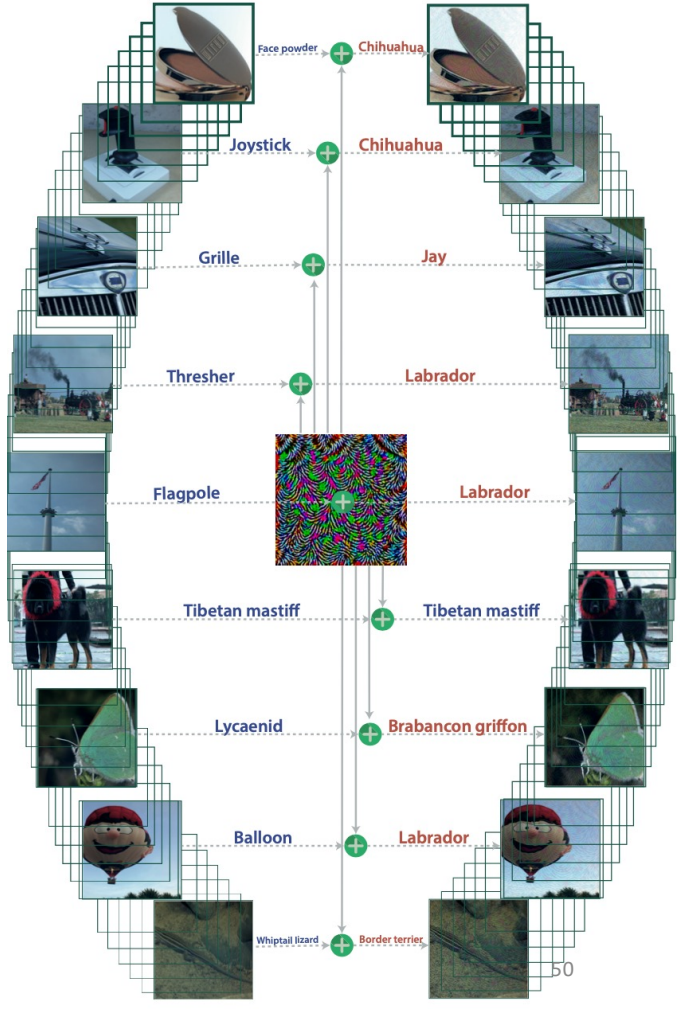
Vary targets …



UNIVERSITÄT BIELEFELD
Technische Fakultät

Universal Adversarial Perturbations

- adds small perturbations to random pixels
- such that the pertubation does not depend on the data and likely turns all points into adversarials



Universal adversarial perturbations, Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard, 2016, https://arxiv.org/abs/1610.08401

UNIVERSITÄT BIELEFELD
Technische Fakultät

Universal Adversarial Perturbations

attack many input data at once

Algorithm 1 Computation of universal perturbations.

input: Data points X, classifier $\hat{k}$, desired ℓ_p norm of the perturbation ξ, desired accuracy on perturbed samples δ.
output: Universal perturbation vector v.
Initialize $v \leftarrow 0$.
while $\text{Err}(X_v) \leq 1 - \delta$ **do**
  for each datapoint $x_i \in X$ **do**
    if $\hat{k}(x_i + v) = \hat{k}(x_i)$ **then**
      Compute the *minimal* perturbation that sends $x_i + v$ to the decision boundary:

$$\Delta v_i \leftarrow \arg\min_r \|r\|_2 \text{ s.t. } \hat{k}(x_i + v + r) \neq \hat{k}(x_i).$$

      Update the perturbation:

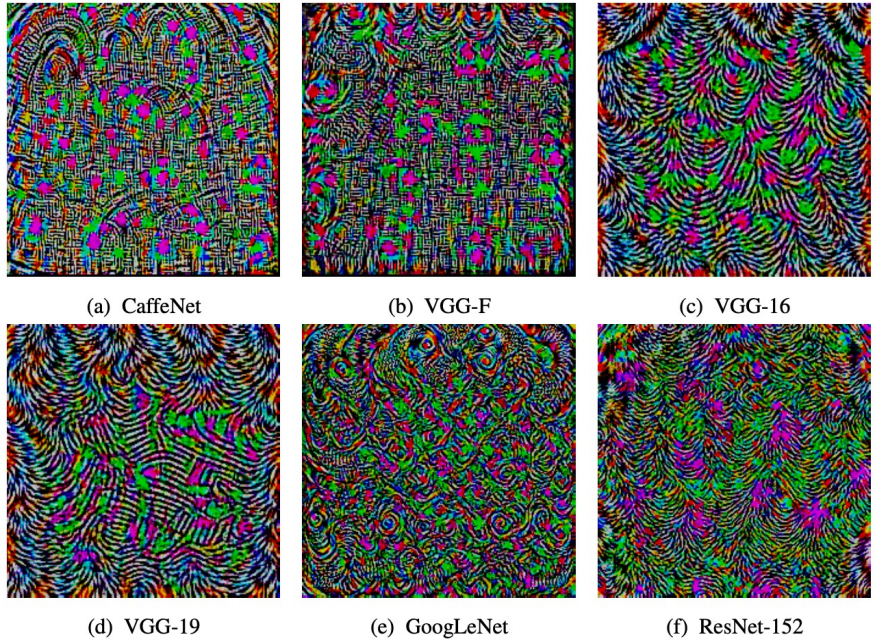
$$v \leftarrow \mathcal{P}_{p,\xi}(v + \Delta v_i).$$

    end if
  end for
end while

Universal adversarial perturbations, Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard, 2016, https://arxiv.org/abs/1610.08401

UNIVERSITÄT BIELEFELD
Technische Fakultät

Universal Adversarial Perturbations



Universal adversarial perturbations, Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard, 2016, https://arxiv.org/abs/1610.08401

Universal Adversarial Perturbations

Universal adversarial perturbations, Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard, 2016, https://arxiv.org/abs/1610.08401

UNIVERSITÄT BIELEFELD
Technische Fakultät

Fast Feature Fool

attack arbitrary inputs

- objective:

$$f(x+\delta) \neq f(x), \text{ for most } x \in \mathcal{X}$$
$$\|\delta\|_\infty < \xi$$

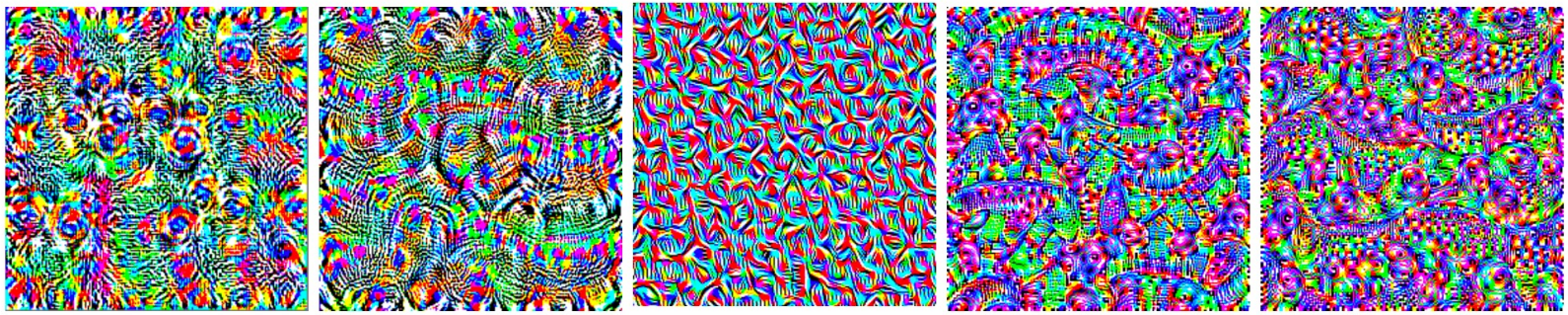
- targeted by maximizing spurious mean activation l_i at each layer i

$$Loss = -\log\left(\prod_{i=1}^{K} \bar{l}_i(\delta)\right) \quad \text{such that} \quad \|\delta\|_\infty < \xi$$

Fast Feature Fool: A data independent approach to universal adversarial perturbations, Konda Reddy Mopuri, Utsav Garg, R. Venkatesh Babu, 2017, https://arxiv.org/abs/1707.05572

UNIVERSITÄT BIELEFELD
Technische Fakultät

Fast Feature Fool



Fast Feature Fool: A data independent approach to universal adversarial perturbations, Konda Reddy Mopuri, Utsav Garg, R. Venkatesh Babu, 2017, https://arxiv.org/abs/1707.05572

Fast Feature Fool

Fast Feature Fool: A data independent approach to universal adversarial perturbations, Konda Reddy Mopuri, Utsav Garg, R. Venkatesh Babu, 2017, https://arxiv.org/abs/1707.05572

UNIVERSITÄT BIELEFELD
Technische Fakultät

Adversarial atttacks

	idea
LBFGS	LBFGS solution of data and model specific objective
FGSM	attack gradient of loss
BIM	iterate over FGSM
Saliency	guide with saliency map, ie output map
Deep fool	go to closest boundary
UAM	add universal domain dependent perturbations
FFF	add universal data independent perturbation

UNIVERSITÄT BIELEFELD
Technische Fakultät

Black box attacks

UNIVERSITÄT BIELEFELD
Technische Fakultät

Black box attacks

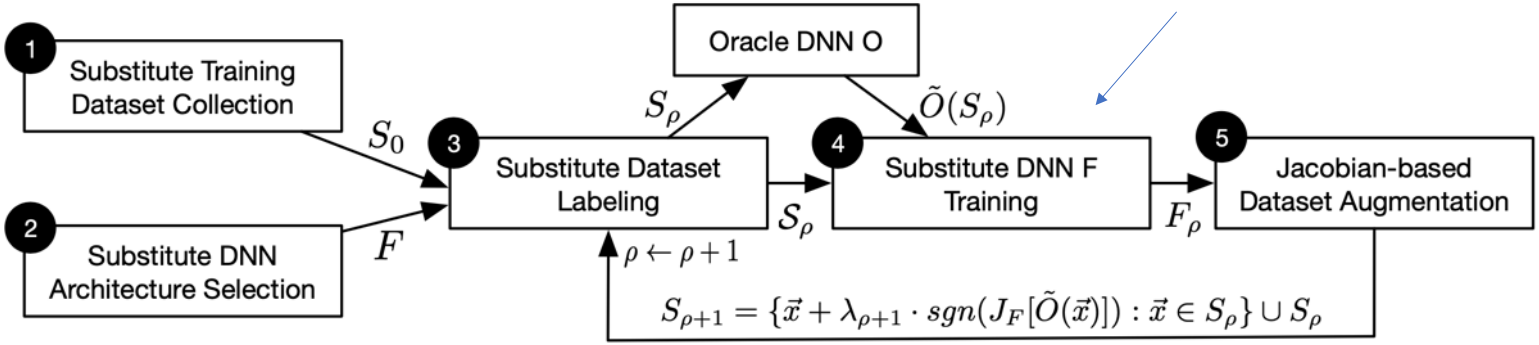
attack without knowing the model:

1) *substitute model:* most attacks transfer in between models, hence one can attack a substitute model trained on similar data
2) *attack by querying*: querying the model and receiving softmax output enables an estimation of gradients or other information based on which to attack
3) *gradient free querying*: output classes are used as information for general local search strategies (such as evolutionary algorithms, …)

A Survey of Black-Box Adversarial Attacks on Computer Vision Models
Siddhant Bhambri, Sumanyu Muku, Avinash Tulasi, Arun Balaji Buduru, https://arxiv.org/abs/1912.01667, 2019

Example: substitute model

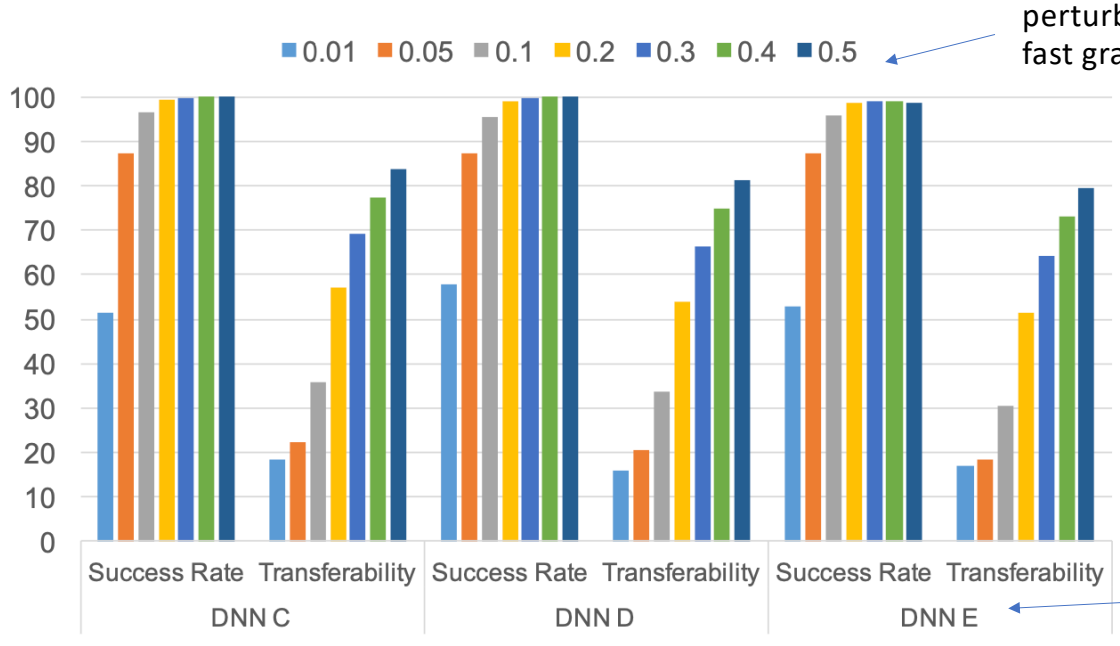
aim for model which is
not so vulnerable to attacks
→ yields strong attacks



Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, ASIA CCS '17, pages 506–519, New York, NY, USA, 2017. ACM.
https://arxiv.org/abs/1602.02697

UNIVERSITÄT BIELEFELD
Technische Fakultät

Example: substitute model



perturbation input selection for fast gradient sign attack

networks with different accuracy

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, ASIA CCS '17, pages 506–519, New York, NY, USA, 2017. ACM.
https://arxiv.org/abs/1602.02697

Example: gradient querying

- soft-max values induce output function

$$f(\mathbf{x}, t) = \max\{\max_{i \neq t} \log[F(\mathbf{x})]_i - \log[F(\mathbf{x})]_t, -\kappa\}$$

- approximate 1st and 2nd derivative

$$\hat{g}_i := \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_i} \approx \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x} - h\mathbf{e}_i)}{2h}$$

$$\hat{h}_i := \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}_{ii}^2} \approx \frac{f(\mathbf{x} + h\mathbf{e}_i) - 2f(\mathbf{x}) + f(\mathbf{x} - h\mathbf{e}_i)}{h^2}$$

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi,and Cho-Jui Hsieh .Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec '17, pages 15–26, New York, NY, USA, 2017. ACM. https://arxiv.org/abs/1708.03999

UNIVERSITÄT BIELEFELD
Technische Fakultät

Example: gradient querying

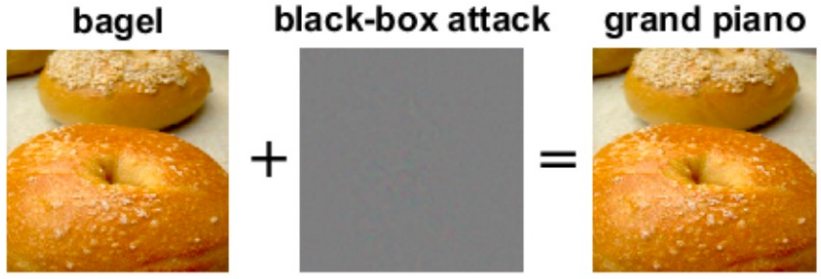
Algorithm 3 ZOO-Newton: Zeroth Order Stochastic Coordinate Descent with Coordinate-wise Newton's Method

Require: Step size η

while not converged **do**
  Randomly pick a coordinate $i \in \{1, \cdots, p\}$
  Estimate $\hat{g}_i$ and $\hat{h}_i$ using (6) and (7)
  if $\hat{h}_i \leq 0$ **then**
    $\delta^* \leftarrow -\eta \hat{g}_i$
  else
    $\delta^* \leftarrow -\eta \frac{\hat{g}_i}{\hat{h}_i}$
  end if
  Update $\mathbf{x}_i \leftarrow \mathbf{x}_i + \delta^*$
end while

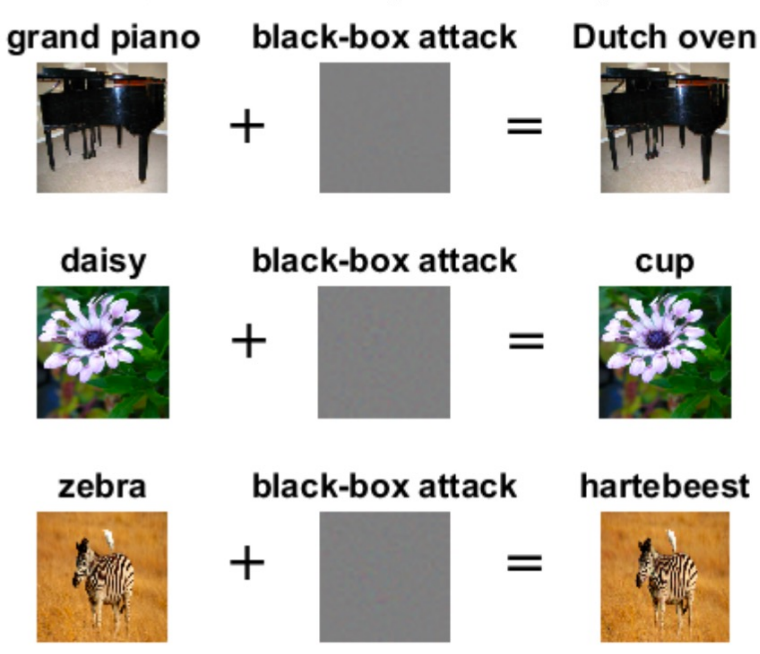
Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi,and Cho-Jui Hsieh .Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec '17, pages 15–26, New York, NY, USA, 2017. ACM. https://arxiv.org/abs/1708.03999

Example: gradient querying



Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi,and Cho-Jui Hsieh .Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec '17, pages 15–26, New York, NY, USA, 2017. ACM. https://arxiv.org/abs/1708.03999

Example: gradient querying



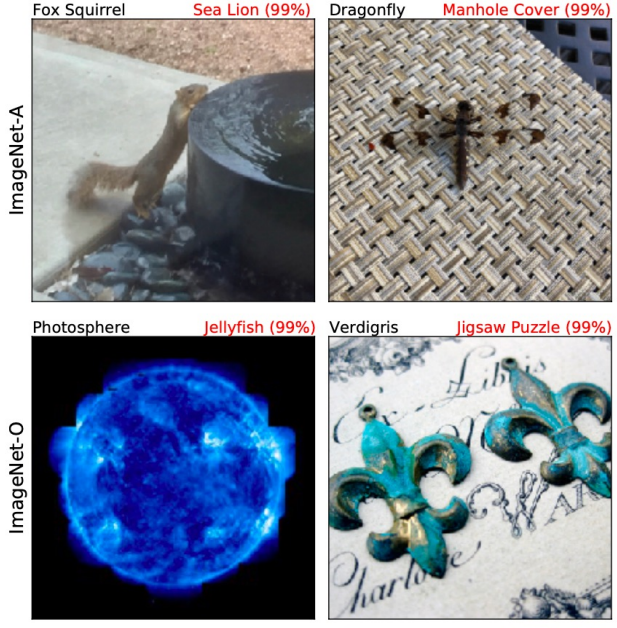
Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi,and Cho-Jui Hsieh .Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec '17, pages 15–26, New York, NY, USA, 2017. ACM. https://arxiv.org/abs/1708.03999

UNIVERSITÄT BIELEFELD
Technische Fakultät

Attacks in the real world

UNIVERSITÄT BIELEFELD
Technische Fakultät

Natural adversarial



images from Imagenet which are typically misclassified by any many models

Natural Adversarial Examples, Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, Dawn Song, 2019, https://arxiv.org/abs/1907.07174

UNIVERSITÄT BIELEFELD
Technische Fakultät

Natural adversarial

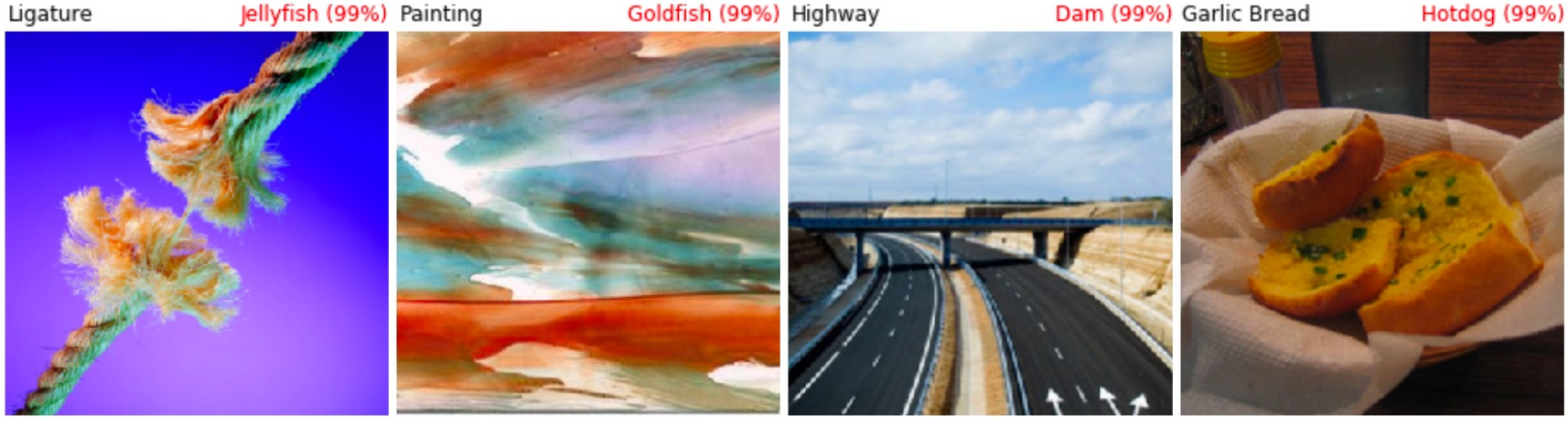
- adding examples deteriorates performance: if these are added to data set, overall accuracy of model gets worse

Natural Adversarial Examples, Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, Dawn Song, 2019, https://arxiv.org/abs/1907.07174

UNIVERSITÄT BIELEFELD
Technische Fakultät

Natural adversarial

- adding examples deteriorates out-of-distribution detection: if these are added to data set, confidence estimate gets worse



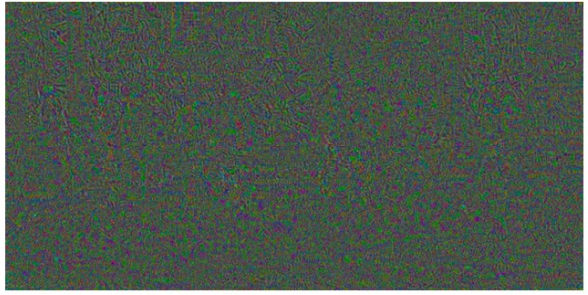
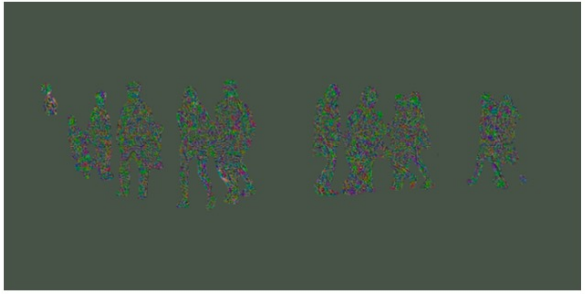
Natural Adversarial Examples, Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, Dawn Song, 2019, https://arxiv.org/abs/1907.07174

UNIVERSITÄT BIELEFELD
Technische Fakultät

Attacks on Image segmentation

- attack all pixels of a certain region at once using BIM

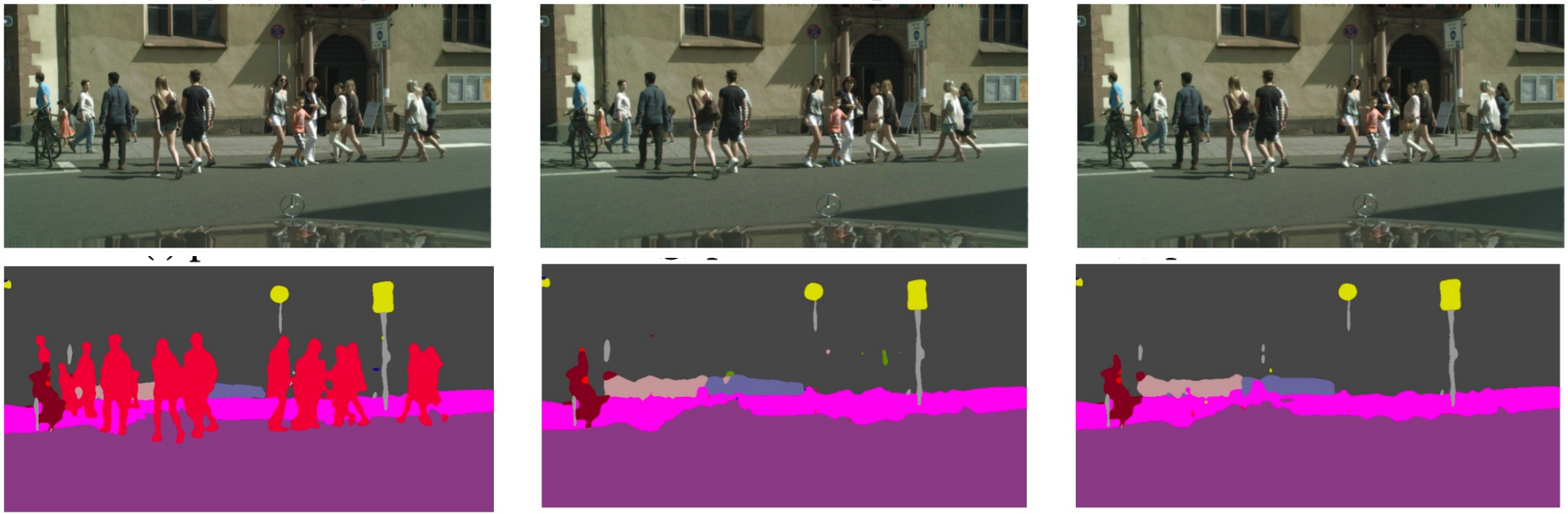


Adversarial Examples for Semantic Image Segmentation, Volker Fischer, Mummadi Chaithanya Kumar, Jan Hendrik Metzen, Thomas Brox, 2017, https://arxiv.org/abs/1703.01101

UNIVERSITÄT BIELEFELD
Technische Fakultät

Attacks on Image segmentation



Adversarial Examples for Semantic Image Segmentation, Volker Fischer, Mummadi Chaithanya Kumar, Jan Hendrik Metzen, Thomas Brox, 2017, https://arxiv.org/abs/1703.01101

One pixel attack

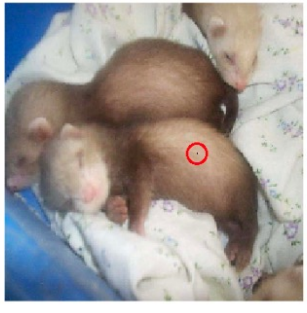
use differential evolutionary methods to optimize

$$\begin{aligned} \underset{e(\mathbf{x})^*}{\text{maximize}} \quad & f_{adv}(\mathbf{x} + e(\mathbf{x})) \\ \text{subject to} \quad & \|e(\mathbf{x})\|_0 \leq d, \end{aligned}$$

Cup(16.48%)
Soup Bowl(16.74%)

Bassinet(16.59%)
Paper Towel(16.21%)

Teapot(24.99%)
Joystick(37.39%)



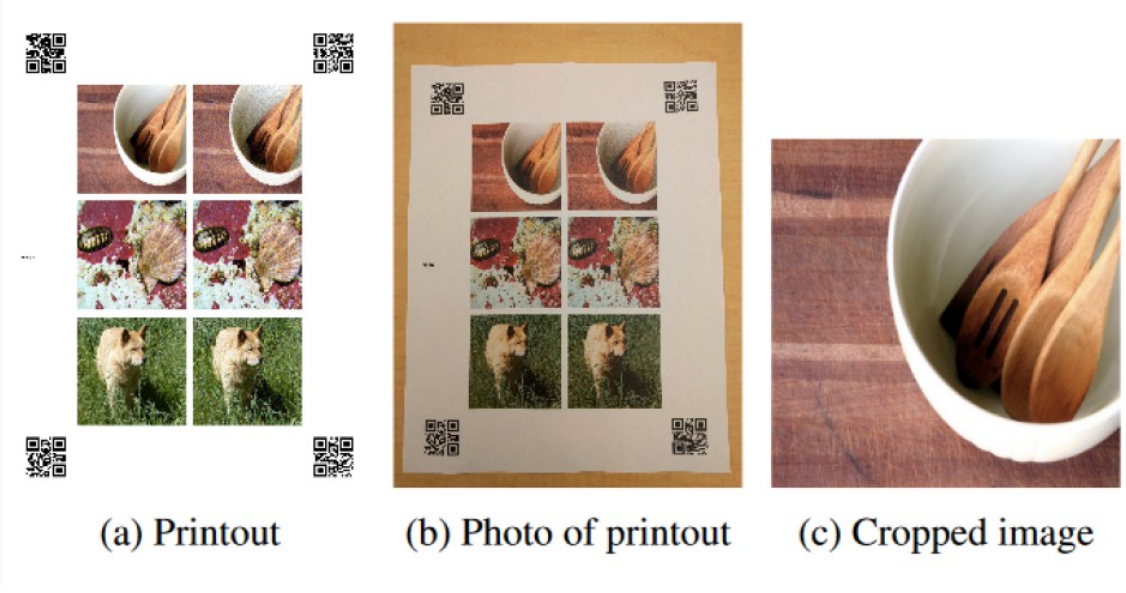
Hamster(35.79%)
Nipple(42.36%)

One pixel attack for fooling deep neural networks
Jiawei Su, Danilo Vasconcellos Vargas, Sakurai Kouichi, 2017, https://arxiv.org/abs/1710.08864

UNIVERSITÄT BIELEFELD
Technische Fakultät

Adversarial examples survive printout

- 60% of FGSM sample stay adversarial when taken via smartphone



(a) Printout (b) Photo of printout (c) Cropped image

Adversarial examples in the physical world, Alexey Kurakin, Ian Goodfellow, Samy Bengio, 2016, https://arxiv.org/abs/1607.02533

UNIVERSITÄT BIELEFELD
Technische Fakultät

Physical objects

- find adversarial which are misclassified under transformations

$$\begin{aligned} \arg\max_{x'} \quad & \mathbb{E}_{t\sim T}[\log P(y_t|t(x'))] \\ \text{subject to} \quad & \mathbb{E}_{t\sim T}[d(t(x'), t(x))] < \epsilon \\ & x \in [0,1]^d \end{aligned}$$

t: manipulation of 3D object

Synthesizing Robust Adversarial Examples, Anish Athalye, Logan Engstrom, Andrew Ilyas, Kevin Kwok, 2017, https://arxiv.org/abs/1707.07397

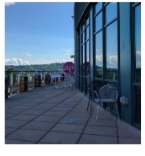
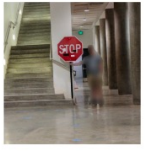
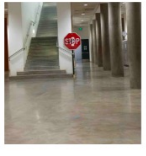
UNIVERSITÄT BIELEFELD
Technische Fakultät

Physical objects

Synthesizing Robust Adversarial Examples, Anish Athalye, Logan Engstrom, Andrew Ilyas, Kevin Kwok, 2017, https://arxiv.org/abs/1707.07397

UNIVERSITÄT BIELEFELD
Technische Fakultät

Driving

Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5′ 0°					
5′ 15°					
10′ 0°					
10′ 30°					
40′ 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%

Robust Physical-World Attacks on Deep Learning Models, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, Dawn Song, 2017, https://arxiv.org/abs/1707.08945

… what makes an attack adversarial and how to defend attacks in 2nd batch of slides …