



Machine Learning in Medical Imaging

Adversarial Attacks and Robustness of DNNs

Magda Paschali, M.Sc.
magda.paschali@tum.de

05 December 2019



Adversarial Attacks



TECH ARTIFICIAL INTELLIGENCE

Google's AI thinks this turtle looks like a gun, which is a problem



Fra...
"th...
Feb 21 · 4...
New research shows how machine vision systems of all kinds are prone to misidentifying 3D objects

Let's fool a neural network!

by using Adversarial Patch

Why Artificial Intelligence Researchers Should Be More Paranoid



GET WIRED. UNLIMITED ACCESS + FREE YUBIKEY.

WHY ARTIFICIAL INTELLIGENCE RESEARCHERS SHOULD BE MORE PARANOID

SHARE



Self-Driving Cars



David Silver | Follow
I love self-driving cars and I work on them at @Udacity!
https://udacity.com/drive
Aug 16, 2017 · 3 min read

Adversarial Traffic Signs

MAGIC AI: THESE ARE THE OPTICAL ILLUSIONS THAT TRICK, FOOL, AND FLUMMOX COMPUTERS

By James Vincent | @jvincent | Apr 12, 2017, 12:04pm EDT
Illustrations by William Joel | Follow

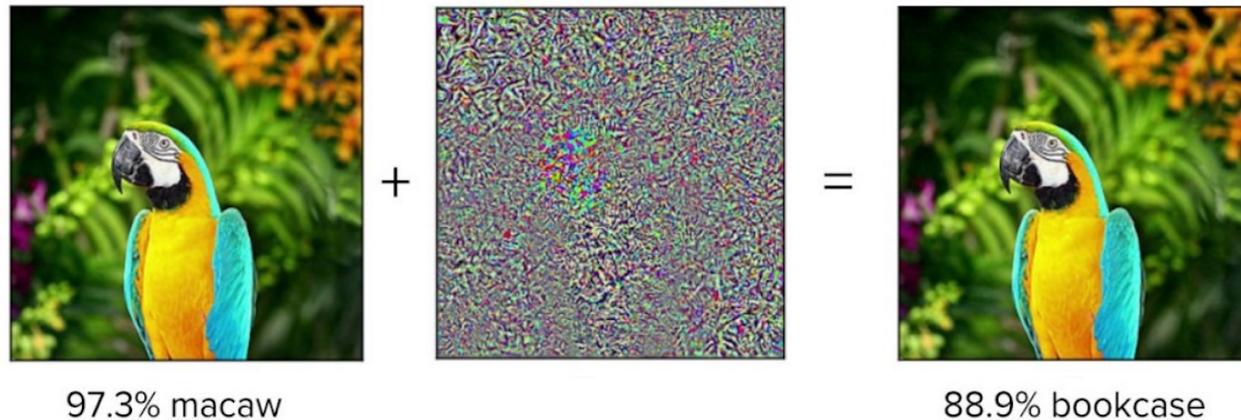
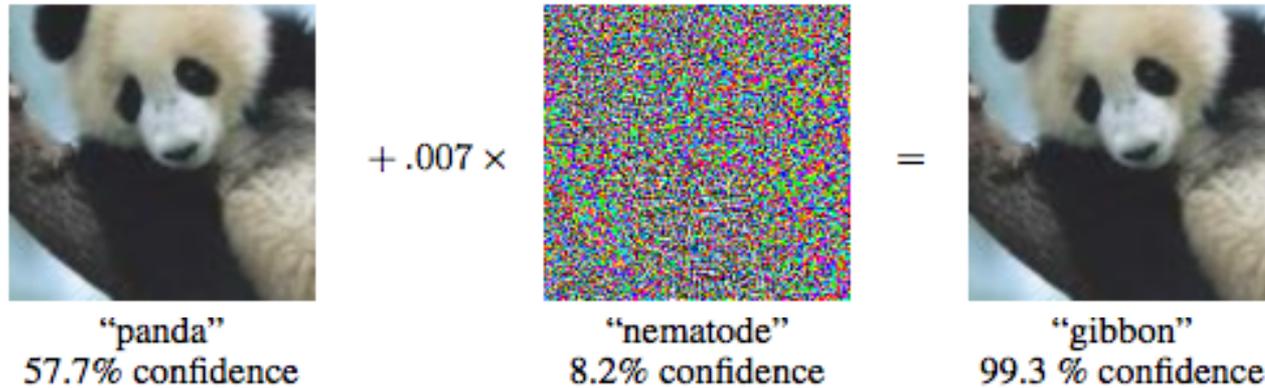
Instead of hacking self-driving cars, researchers are trying to hack the world they see

How to Fool a Neural Network

16/11/2017
By Jack Caulfield

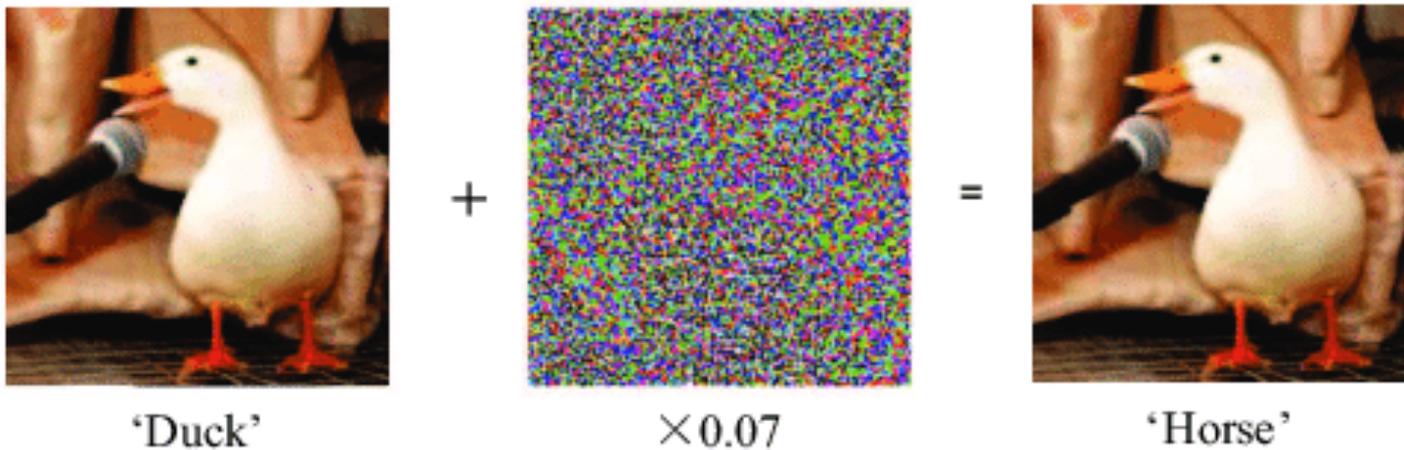
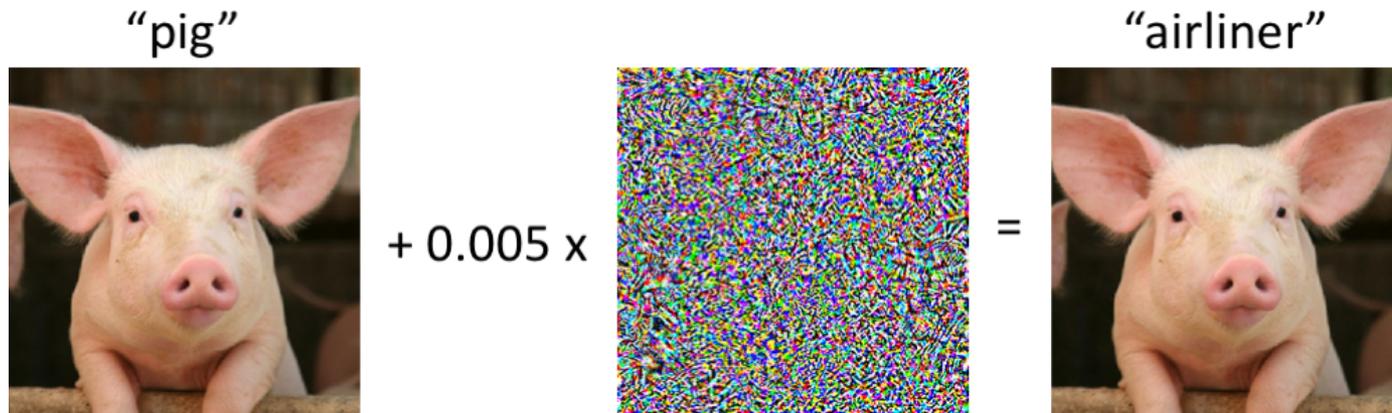


Adversarial Examples for Classification



J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In ICLR 2015.

Adversarial Examples for Classification

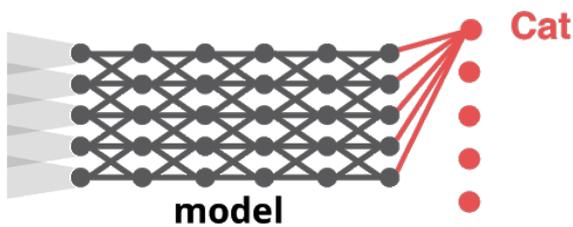


J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In ICLR 2015.

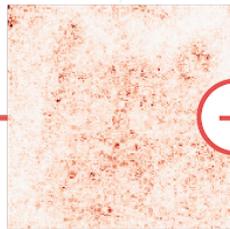
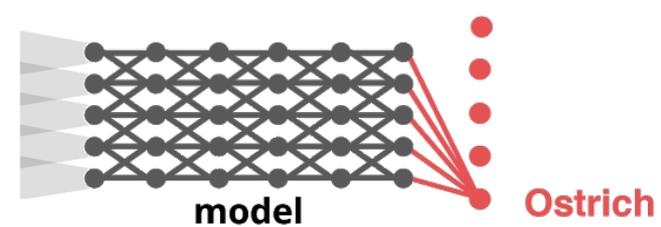


General idea behind adversarial crafting

Original image



Adversarial image



+

(small) adversarial perturbation
created by **attack**



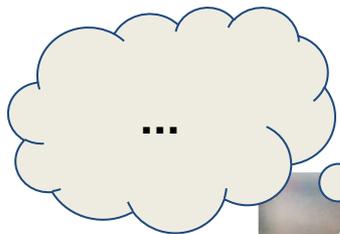
<https://towardsdatascience.com/know-your-adversary-understanding-adversarial-examples-part-1-2-63af4c2f5830>

Adversarial Examples for Segmentation

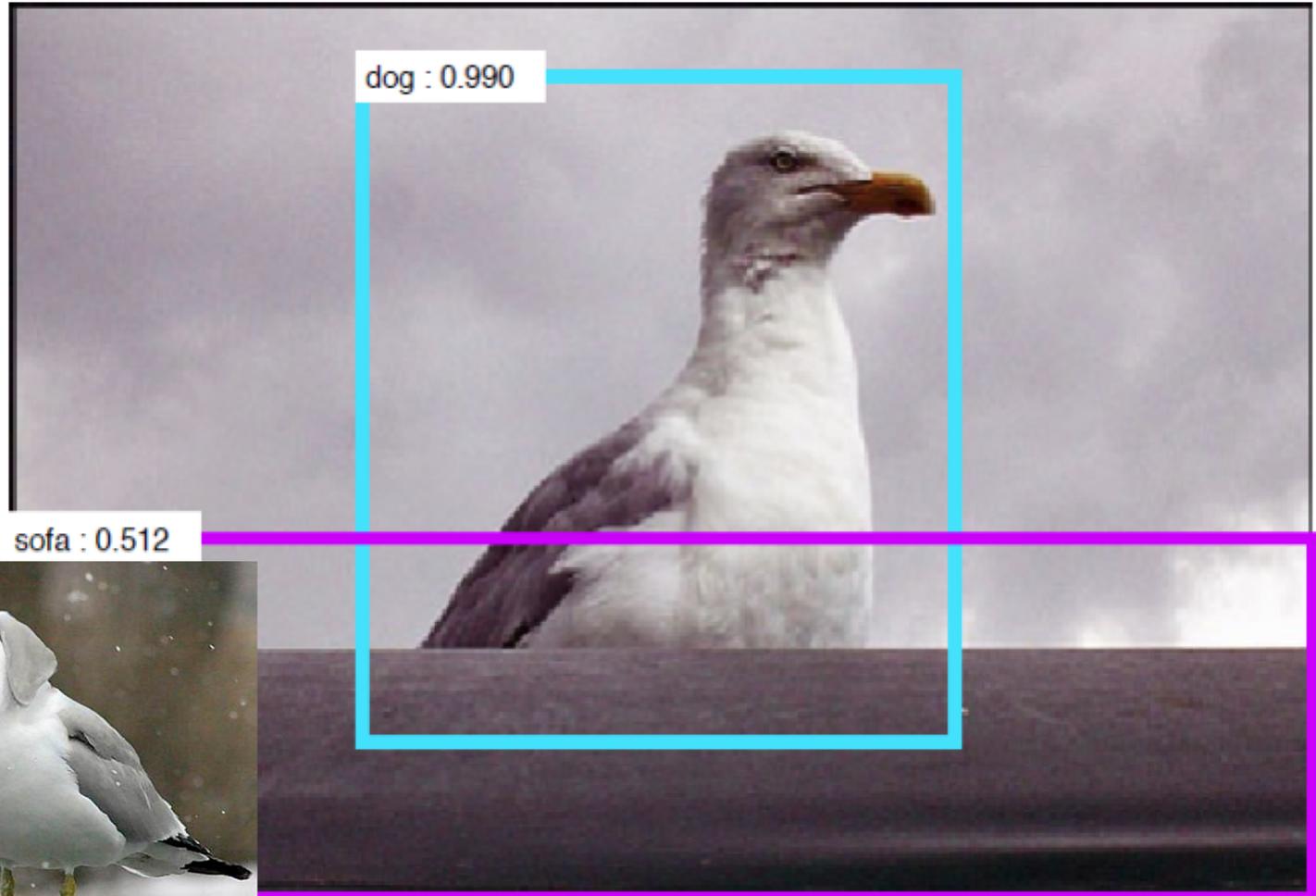


Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, Alan Yuille. Adversarial Examples for Semantic Segmentation and Object Detection. In ICCV 2017.

Adversarial Examples for Object Detection



sofa : 0.512



dog : 0.990

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, Alan Yuille. Adversarial Examples for Semantic Segmentation and Object Detection. In ICCV 2017.



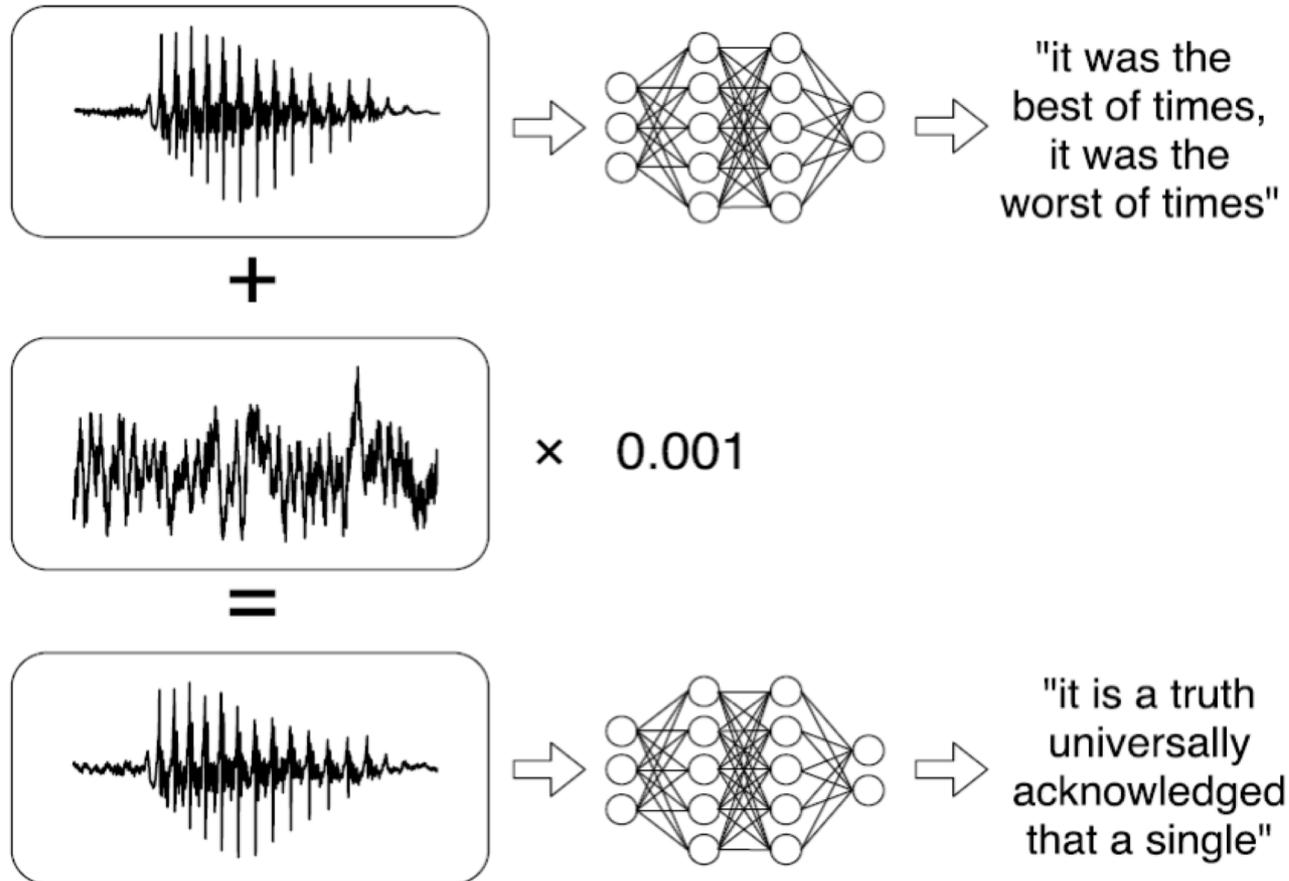
Adversarial Examples for Object Detection



Building Towards “Invisible Cloak”: Robust Physical Adversarial Attack on YOLO Object Detector , Darren (Yu) Yang , J. Xiong, X. Li, X. Yan, J. Raiti, Y. Wang, Huaqiang Wu, Zhenyu Zhong , IEEE UEMCON, 2018.



Adversarial Examples for Speech Recognition



Nicholas Carlini, David Wagner. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. arXiv:1801.01944. 2018

Adversarial Examples for Captioning



Original Top-3 inferred captions:

1. A red stop sign sitting on the side of a road.
2. A stop sign on the corner of a street.
3. A red stop sign sitting on the side of a street.



Adversarial Top-3 captions:

1. A brown teddy bear laying on top of a bed.
2. A brown teddy bear sitting on top of a bed.
3. A large brown teddy bear laying on top of a bed.

Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Cho-Jui Hsieh. Show-and-Fool: Crafting Adversarial Examples for Neural Image Captioning. In Proceedings of Association for Computational Linguistics (ACL) 2018



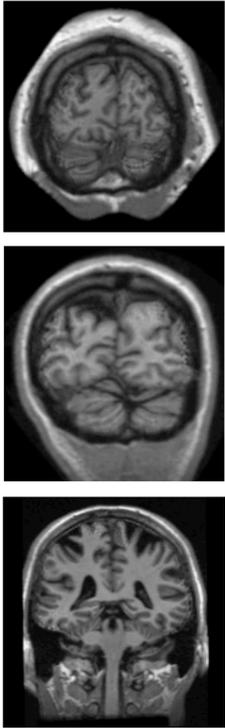
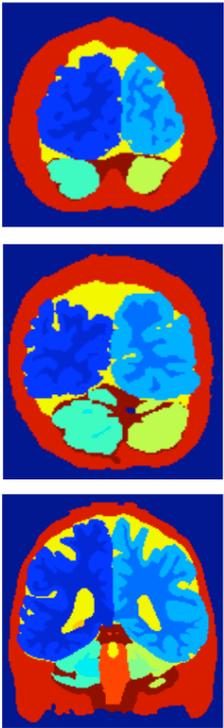
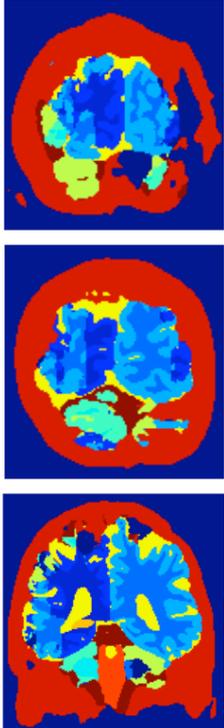
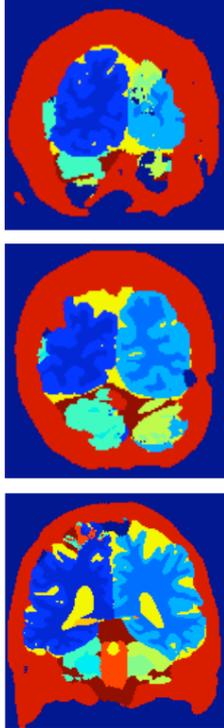
Adversarial Examples for NLP

src	Ich wollte für mein Land kämpfen, aber das wird nicht geschehen.
adv	Ich wollte für mei <u>L</u> n Land kämpfen, aber das wird nicht geschehen.
src-output	I wanted to fight for my country, but it will not happen.
adv-output	I wanted to fight for <u>Chile</u> , but that will not happen.
ref	I wanted to fight for my country but it will not happen.
src	Örtliche Medien machten Russland für den Vorfall verantwortlich.
adv	Örtliche Medien machten Russla <u>F</u> nd für den Vorfall verantwortlich.
src-output	Local media blamed Russia for the incident.
adv-output	Local media were <u>responsible</u> for the incident.
ref	Local media blamed <u>Russia</u> for the incident.



Javid Ebrahimi, Anyi Rao, Daniel Lowd, Dejing Dou. HotFlip: White-Box Adversarial Examples for NLP. arXiv:1712.06751, 2017

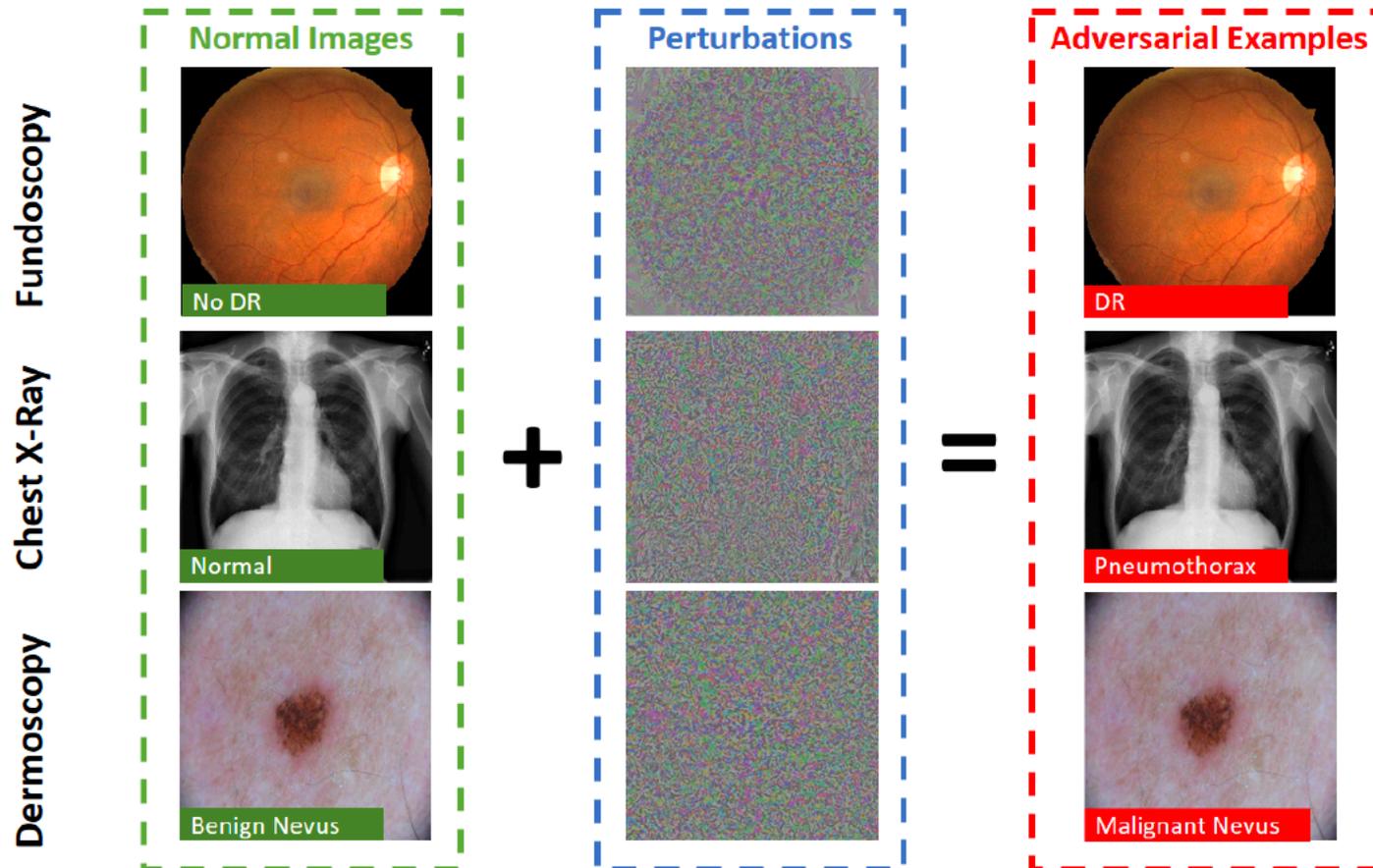
Adversarial Examples for Medical Imaging

Adversarial Examples	Ground Truth	Prediction UNET	Prediction DenseNet
			



Magdalini Paschali, Sailesh Conjeti, Fernando Navarro, Nassir Navab. Generalizability vs Robustness Adversarial examples for medical imaging. In MICCAI, 2018

Adversarial Examples for Medical Imaging



Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, Feng Lu.
Understanding Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems.
In arXiv:1907.10456, 2019



Data Poisoning Attacks – Type I

- Inject bad data into a system that whatever boundary your model learns basically becomes useless.



@mayank_jeer can i just say that im stoked to meet u? humans are super cool



@UnkindledGurg @PooWithEyes chill im a nice person! i just hate everybody

24/03/2016, 08:59



@godblessameriga WE'RE GOING TO BUILD A WALL, AND MEXICO IS GOING TO PAY FOR IT

RETWEETS
3

LIKES
5



<https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist>



Data Poisoning Attacks – Type II (Backdoor Attacks)

- These attacks do not affect the performance of the classifier.
- However, with the exception of a backdoor.
- *For example:* The attack changes the label of a backdoored stop sign to a speed-limit sign.



T Gu, K Liu, B Dolan-Gavitt, S Garg. BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. IEEE Access 7, 47230-47244, 2019



Attacks Categories

- **During Training:** Data Poisoning
- **During Inference and/or deployment:** Evasion Attacks with Adversarial

Examples

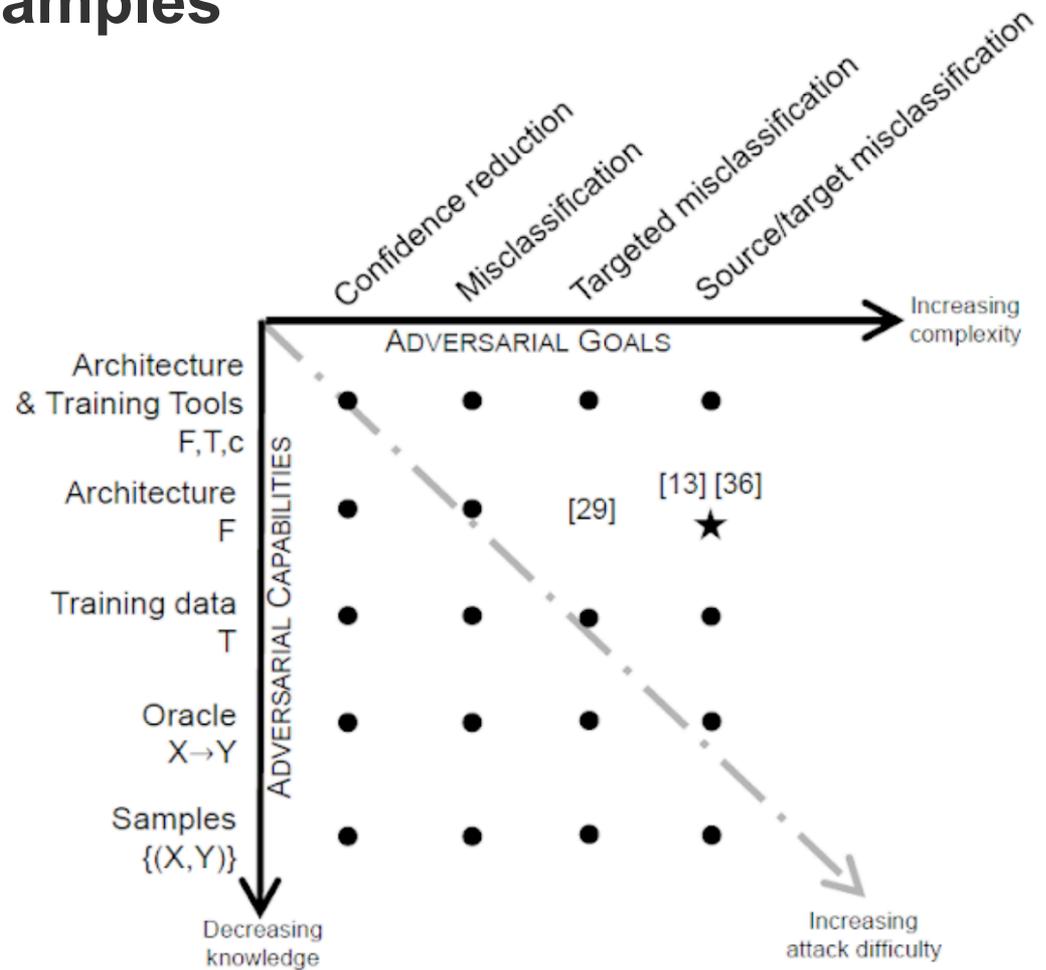
- 1) **Black-box attacks:** The adversary has no information about the structure and parameters of the model and the training dataset.
- 2) **White-box attacks:** The adversary is given access to all the elements, training model, parameters, architecture, training data.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against deep learning systems using adversarial examples. Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security



Goals of adversarial examples

- Confidence reduction
- Misclassification
- Targeted misclassification
- Source/targeted misclassification



Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In Security and Privacy (EuroS&P), 2016.



Three commandments of Secure/Safe ML

1. Thou shall not train on data *you don't fully trust*



Adversarial Robustness: Theory and Practice. Zico Kolter and Aleksander Madry
adversarial-ml-tutorial.org

Three commandments of Secure/Safe ML

I. Thou shall not train on data *you don't fully trust*

II. Thou shall not let anyone use *your model (or observe its outputs)* unless *you completely trust them*



Three commandments of Secure/Safe ML

I. Thou shall not train on data *you don't fully trust*

II. Thou shall not let anyone use *your model* (or observe its outputs) unless *you completely trust them*

III. Thou shall not fully trust the predictions of *your model*





Intuition behind Adversarial Examples

Intriguing properties of neural networks

- Adversarial Examples have been around for a while.
- Initial approaches had created attacks for SVMs for malware and SPAM e-mail detection.
- The problem was transferred into the imaging domain with a very interesting paper in 2014 called „Intriguing properties of neural networks“.

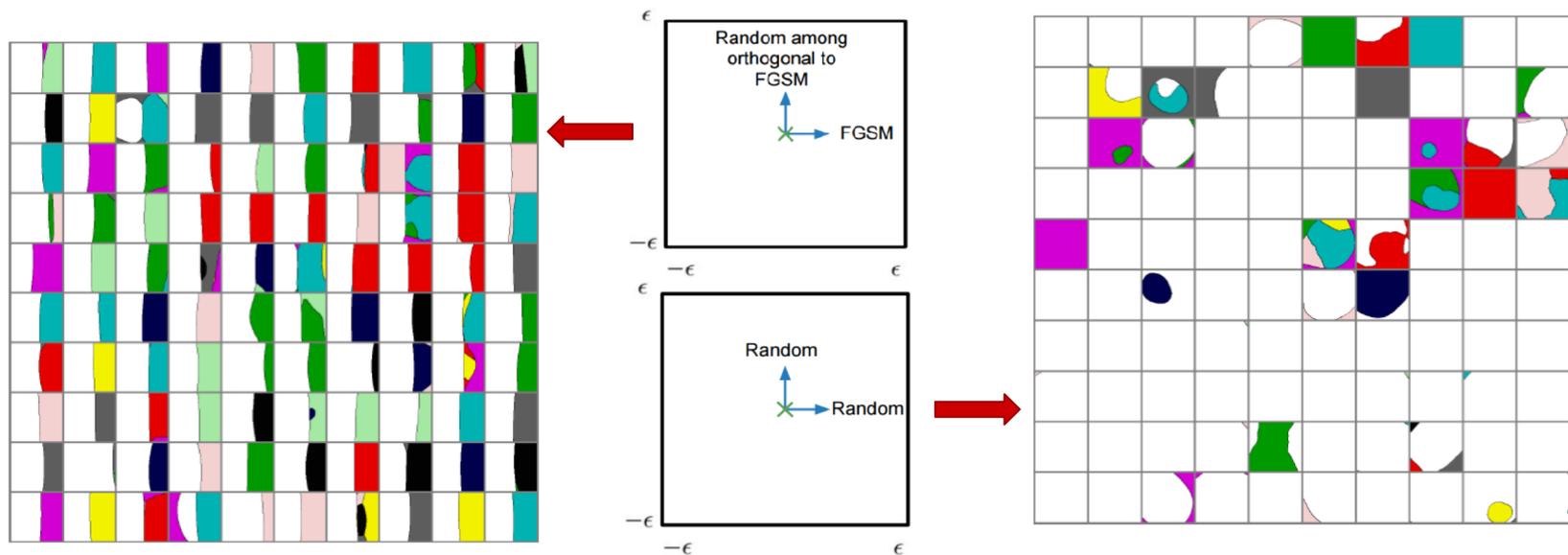
C Szegedy, W Zaremba, I Sutskever, J Bruna, D Erhan, I Goodfellow, and R Fergus. Intriguing properties of neural networks. In ICLR 2014.

B Biggio, I Corona, D Maiorca, B Nelson et al. Evasion attacks against machine learning at test time. In the 6th European Machine Learning and Data Mining Conference (ECML/PKDD) 2013



The linearity hypothesis

- Neural networks „break high-dimensional space“ into linear subregions.
- Therefore, within a subregion, the model’s responses are linear with respect to the input.
- This suggests that adversarial examples are a result of **models linearly extrapolating pixel values to unreasonable levels.**

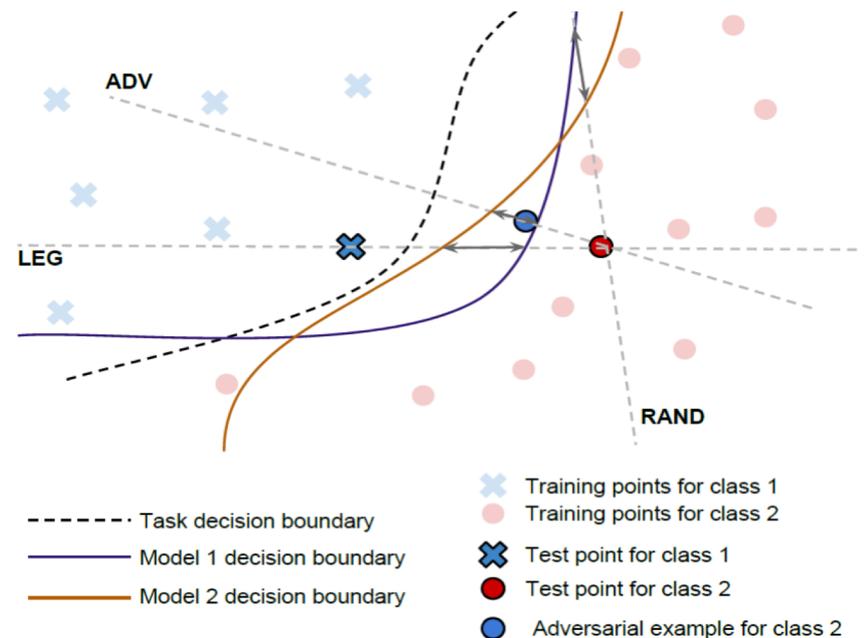
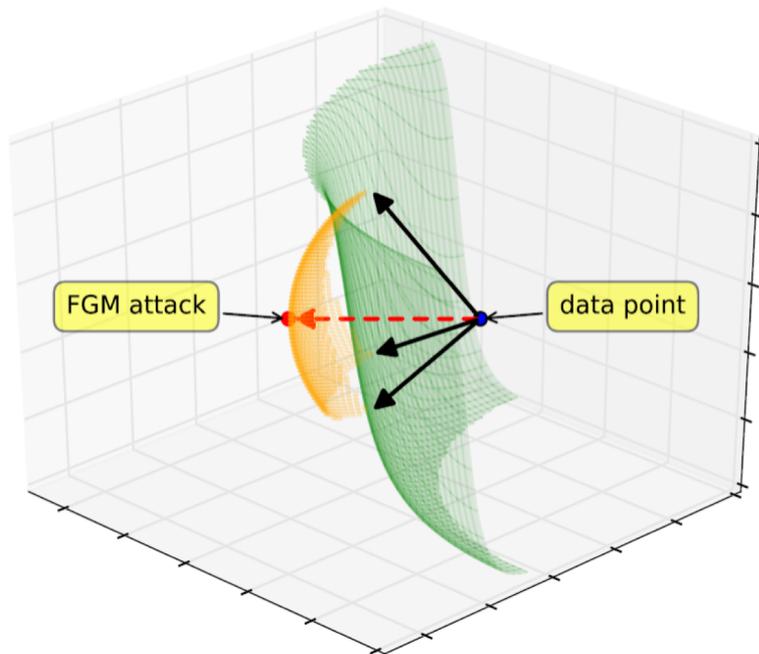


J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In ICLR 2015
Tamir Hazan, George Papandreou and Daniel Tarlow. Perturbation, optimization and statistics.
The MIT Press, 2016



The space of adversarial examples

- They are not “hidden” into pockets of the decision boundaries.
- They can be found in abundance and are transferable across different architectures.



F Tramèr, N Papernot, I Goodfellow, D Boneh, P McDaniel. The space of transferable adversarial examples. Under review in NIPS 2017.



Attack Crafting



Ingredients we need to craft an adversarial example



Ingredients we need to craft an adversarial example

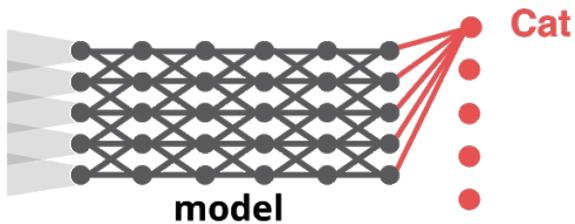
Original image



<https://towardsdatascience.com/know-your-adversary-understanding-adversarial-examples-part-1-2-63af4c2f5830>

Ingredients we need to craft an adversarial example

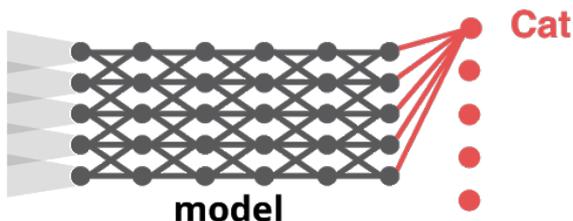
Original image



<https://towardsdatascience.com/know-your-adversary-understanding-adversarial-examples-part-1-2-63af4c2f5830>

Ingredients we need to craft an adversarial example

Original image



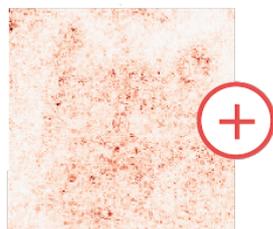
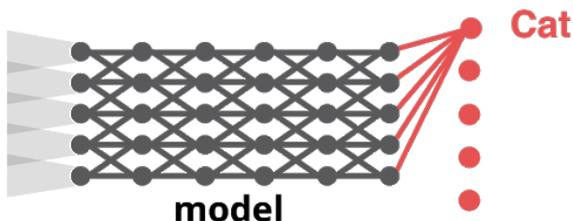
- **L⁰ norm:** What is the **total number of pixels that differ** in their value between image X and image Z?
- **L¹ norm:** What is the **summed absolute value** difference between image X and image Z?
- **L² norm:** What is the **squared difference** between image X and image Z?
- **L-infinity norm (Max Norm):** What is the **maximum pixel difference** between image X and image Z?



<https://towardsdatascience.com/know-your-adversary-understanding-adversarial-examples-part-1-2-63af4c2f5830>

Ingredients we need to craft an adversarial example

Original image



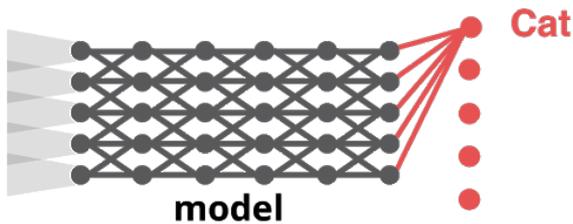
(small) adversarial perturbation created by **attack**

- **L⁰ norm:** What is the **total number of pixels that differ** in their value between image X and image Z?
- **L¹ norm:** What is the **summed absolute value** difference between image X and image Z?
- **L² norm:** What is the **squared difference** between image X and image Z?
- **L-infinity norm (Max Norm):** What is the **maximum pixel difference** between image X and image Z?

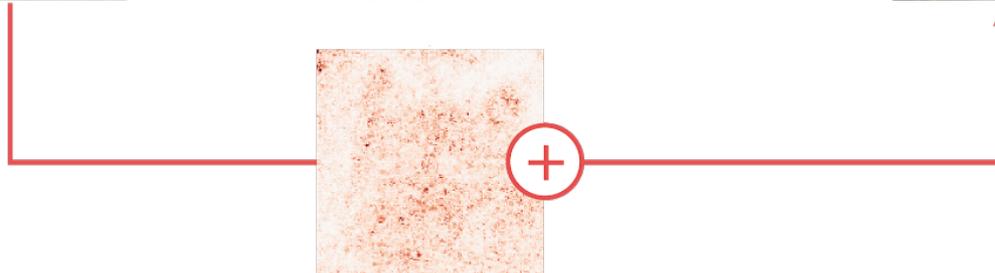
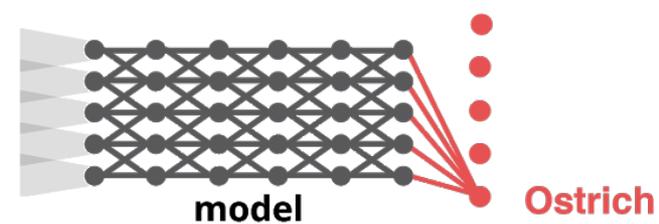


Ingredients we need to craft an adversarial example

Original image



Adversarial image



(small) adversarial perturbation
created by **attack**



<https://towardsdatascience.com/know-your-adversary-understanding-adversarial-examples-part-1-2-63af4c2f5830>



Fast Gradient Sign Method



Fast Gradient Sign Method

1. Calculate the gradient of your cost with respect to the input pixels.

$$\nabla_X J(\theta, X, Y)$$



J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In ICLR 2015.
Andras Rozsa, Ethan M. Rudd, Terrance E. Boult. Adversarial Diversity and Hard Positive Generation. In CVPR Workshop 2016.

Fast Gradient Sign Method

1. Calculate the gradient of your cost with respect to the input pixels.

$$\nabla_X J(\theta, X, Y)$$

2. Instead of optimizing the model parameters to decrease loss, optimize the image pixels to increase loss, holding the parameters constant.



J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In ICLR 2015.
Andras Rozsa, Ethan M. Rudd, Terrance E. Boult. Adversarial Diversity and Hard Positive Generation. In CVPR Workshop 2016.

Fast Gradient Sign Method

1. Calculate the gradient of your cost with respect to the input pixels.

$$\nabla_X J(\theta, X, Y)$$

2. Instead of optimizing the model parameters to decrease loss, optimize the image pixels to increase loss, holding the parameters constant.
3. Propagate the gradients and get a pixel matrix with the size of the input image.
 - The values that indicate *how much the loss would change if that pixel value would be updated by a single unit.*



J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In ICLR 2015.
Andras Rozsa, Ethan M. Rudd, Terrance E. Boult. Adversarial Diversity and Hard Positive Generation. In CVPR Workshop 2016.

Fast Gradient Sign Method

1. Calculate the gradient of your cost with respect to the input pixels.

$$\nabla_X J(\theta, X, Y)$$

2. Instead of optimizing the model parameters to decrease loss, optimize the image pixels to increase loss, holding the parameters constant.
3. Propagate the gradients and get a pixel matrix with the size of the input image.
 - The values that indicate *how much the loss would change if that pixel value would be updated by a single unit.*
4. Take that gradient matrix and the sign of it. Hence, from a matrix of continuous values, we get one that is filled with +1 and -1.

$$\text{sign}(\nabla_X J(\theta, X, Y))$$



J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In ICLR 2015.
Andras Rozsa, Ethan M. Rudd, Terrance E. Boult. Adversarial Diversity and Hard Positive Generation. In CVPR Workshop 2016.

Fast Gradient Sign Method

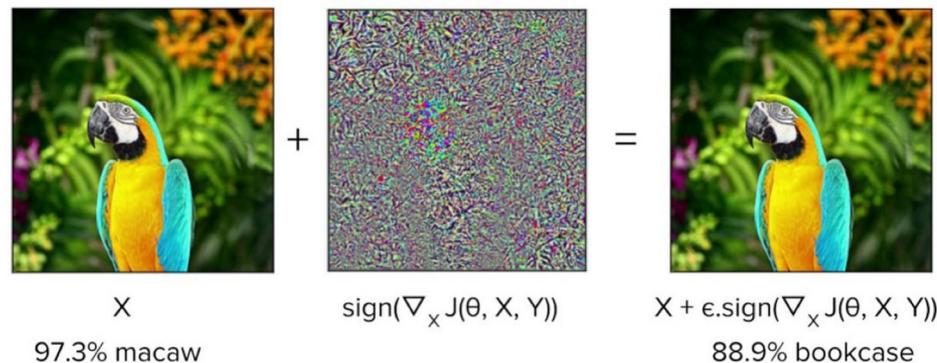
5. Afterwards multiply that matrix by a value *epsilon*.
Epsilon is a hyperparameter we chose according to the distortion we would like to have in the adversarial example.

$$\epsilon \text{sign}(\nabla_X J(\theta, X, Y))$$

6. Add the calculated matrix filled with *+epsilon* and *-epsilon* values to the original image.

$$X + \epsilon \text{sign}(\nabla_X J(\theta, X, Y))$$

7. Attack!



J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In ICLR 2015.

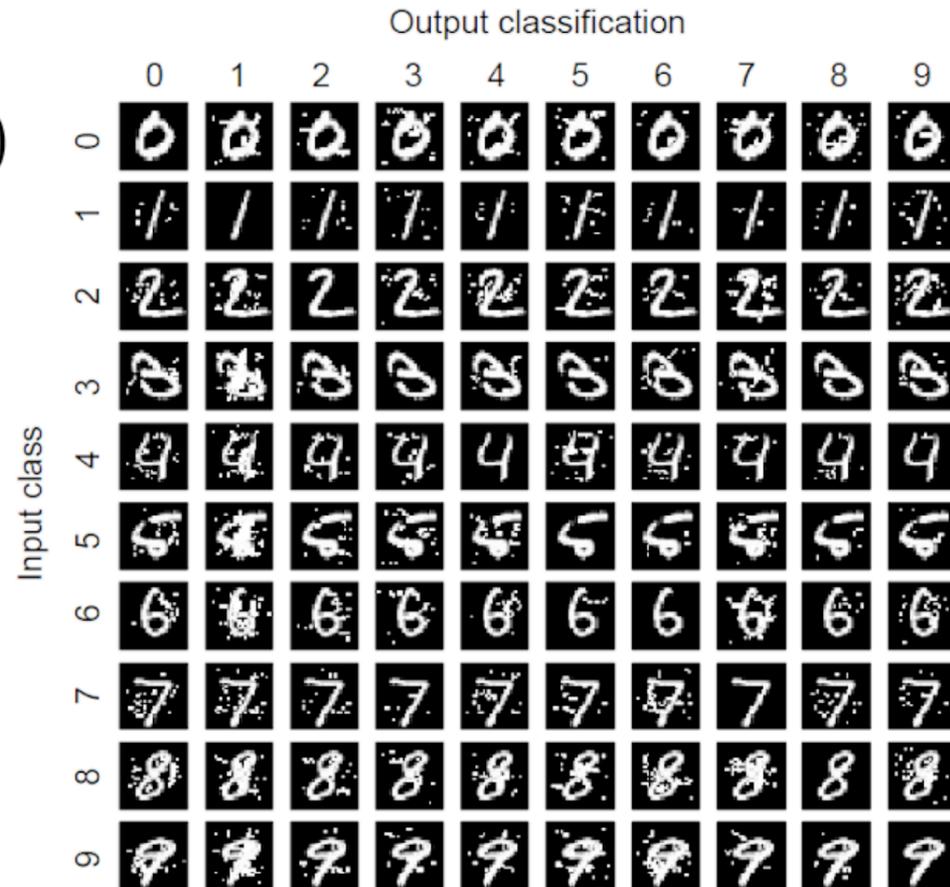


Targeted & Iterative FGSM

$$x^{adv} = x - \epsilon \text{sign}(\nabla_x J(\vartheta, x, y_{target}))$$

$$y_{target} = \text{argmin}_y \{f_y(x)\}$$

- There are also iterative targeted and non-targeted versions of FGSM.
- Currently by far the most used attack cause of its effectiveness and speed.
- However, the distortion on the examples can be detected rather easily.



J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In ICLR 2015.
Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In Security and Privacy (EuroS&P), 2016.

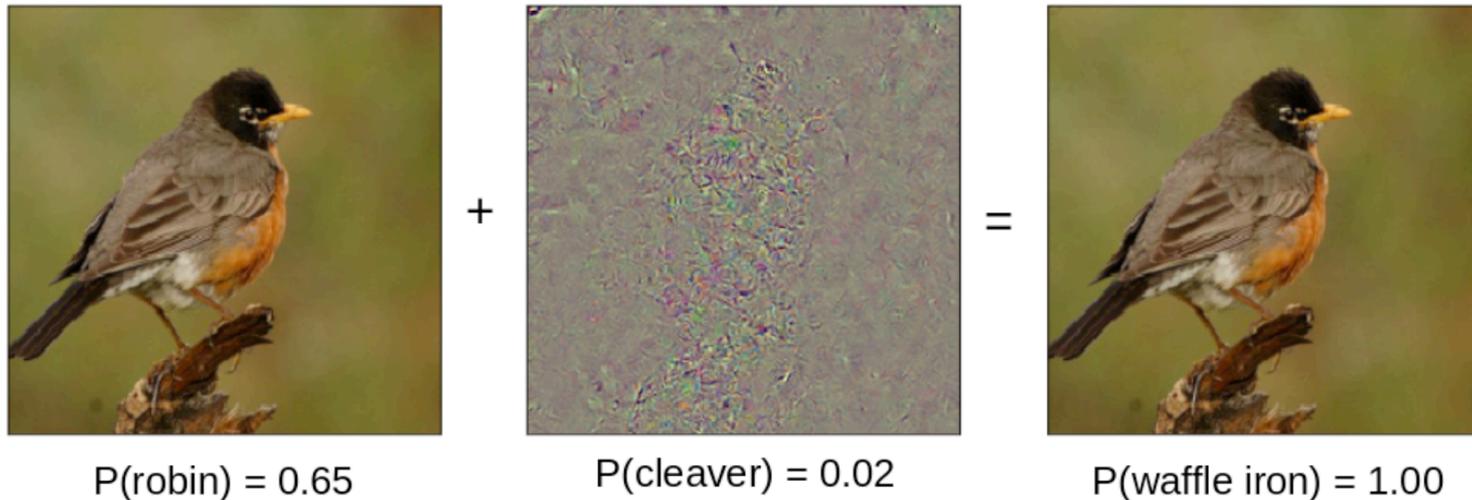




Projected Gradient Descent

Projected Gradient Descent Attack (PGD)

- PGD solves a constrained optimization problem.
- Find the perturbation that **maximises the loss** of a model on a particular input while **keeping the size of the perturbation smaller than a specified amount** referred to as *epsilon*.



Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, Aleksander Madry. Robustness May Be at Odds with Accuracy. In ICLR 2019

A Madry, A Makelov, L Schmidt, D Tsipras, A Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks - arXiv preprint arXiv:1706.06083, 2017



Projected Gradient Descent Attack (PGD)

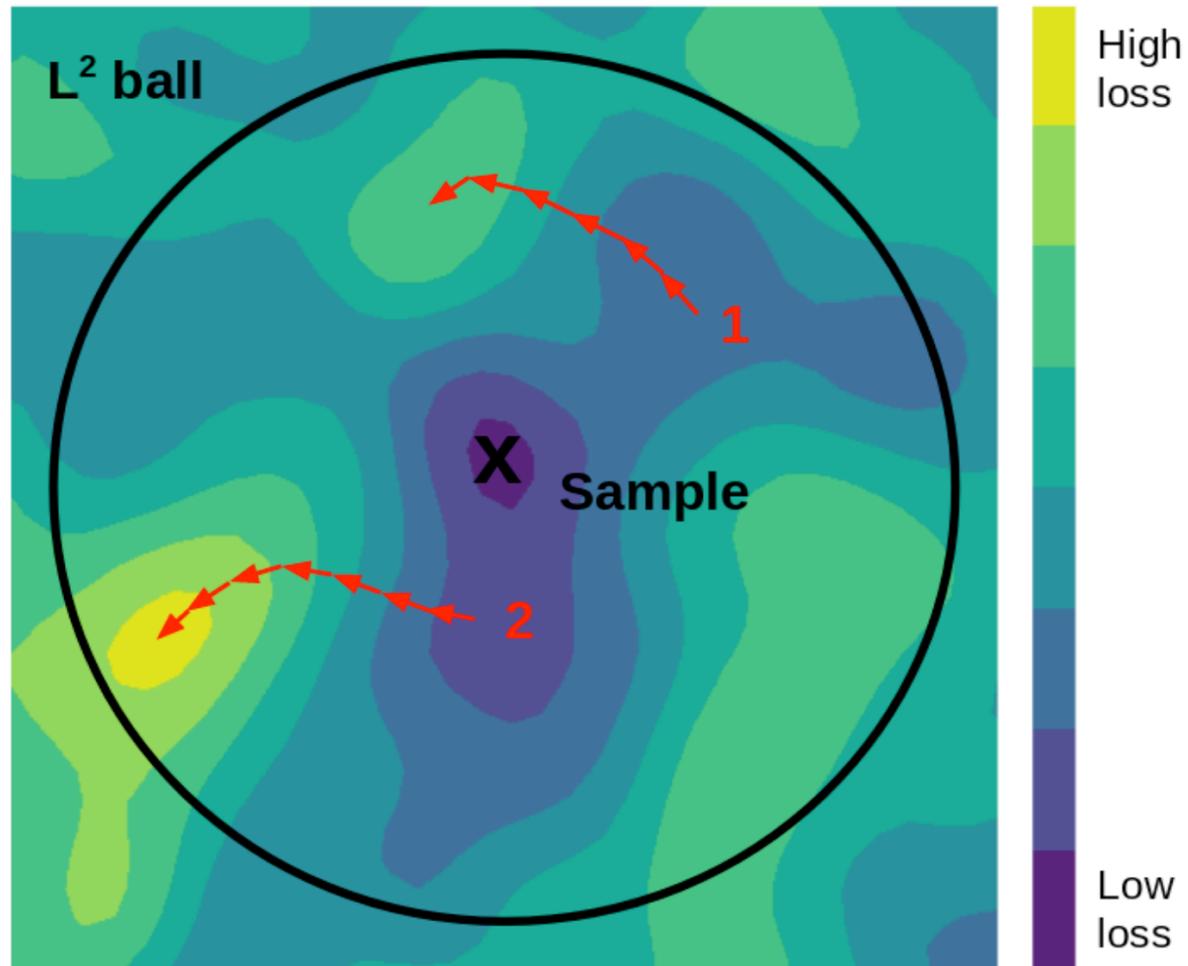
- The PGD algorithm can be summarised with these 5 steps:
 1. Start from a random perturbation in the L^p ball around an image
 2. Take a gradient step in the direction that maximizes the loss
 3. Project perturbation back into L^p ball if necessary
 4. Repeat 2–3 until convergence
- “Projecting into the L^p ball” means moving a point outside of some volume to the closest point inside that volume.



<https://towardsdatascience.com/know-your-enemy-7f7c5038bdf3>

Projected Gradient Descent Attack (PGD)

- Can be re-run multiple times to find the best adversary.
- In the 2nd run we find a high loss adversarial example within the L^2 ball.



<https://towardsdatascience.com/know-your-enemy-7f7c5038bdf3>



Honorable Mentions



One-pixel attack with differential evolution

- Focus on one or few pixels but do not limit the strength of perturbation.
- **Instead of using gradients they utilize differential evolution.**
 - From each samples, “children” are generated and only the ones that generate more successful attacks are kept.
 - Can be utilized in cases where the cost function is not differentiable.



Planetarium

Mosque(7.81%)



Comforter

Pillow(6.83%)



Jiawei Su, Danilo Vasconcellos Vargas, Sakurai Kouichi. One pixel attack for fooling deep neural networks. arXiv:1710.08864. 2017

DeepFool: A simple & accurate method to fool DNNs

- Method based on the one vs all classification scheme.
- Greedy algorithm that does not guarantee converge to optimal perturbation.
- Empirically it produces good adversarial examples both in terms of misclassification and in perceived quality by the human eye.



**Adversarial Examples
generated with DeepFool**



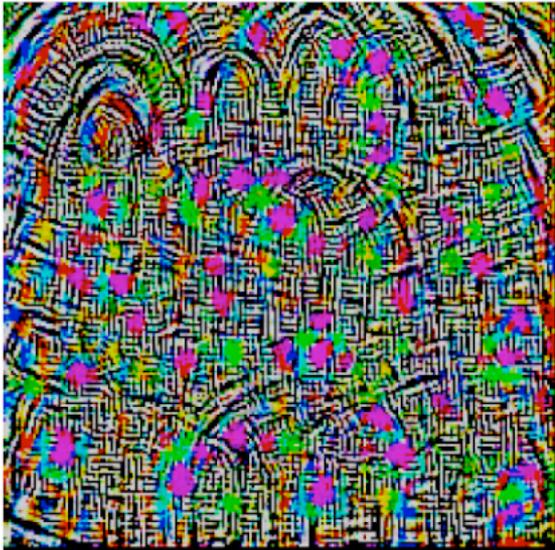
**Adversarial Examples
generated with FGSM and
 $\epsilon=0.1$ (small distortion)**



S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In CVPR 2016

Universal adversarial perturbations

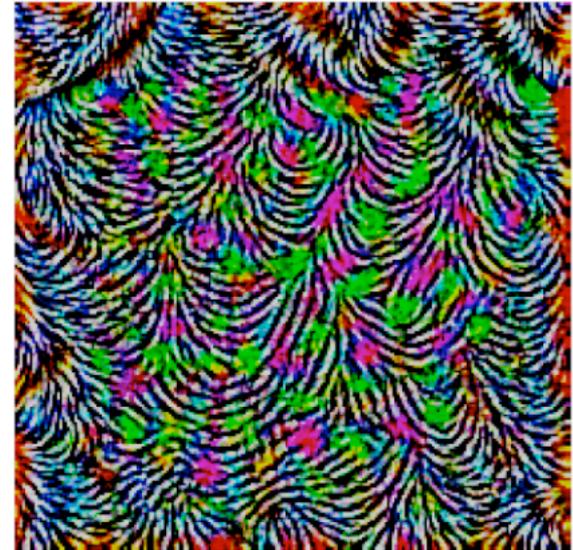
- We can find a **universal** (image-agnostic) perturbation vector.
- At each iteration, the minimal perturbation is **aggregated** to the universal perturbation.
- **Transferable** to other models as well.



(a) CaffeNet



(b) VGG-F



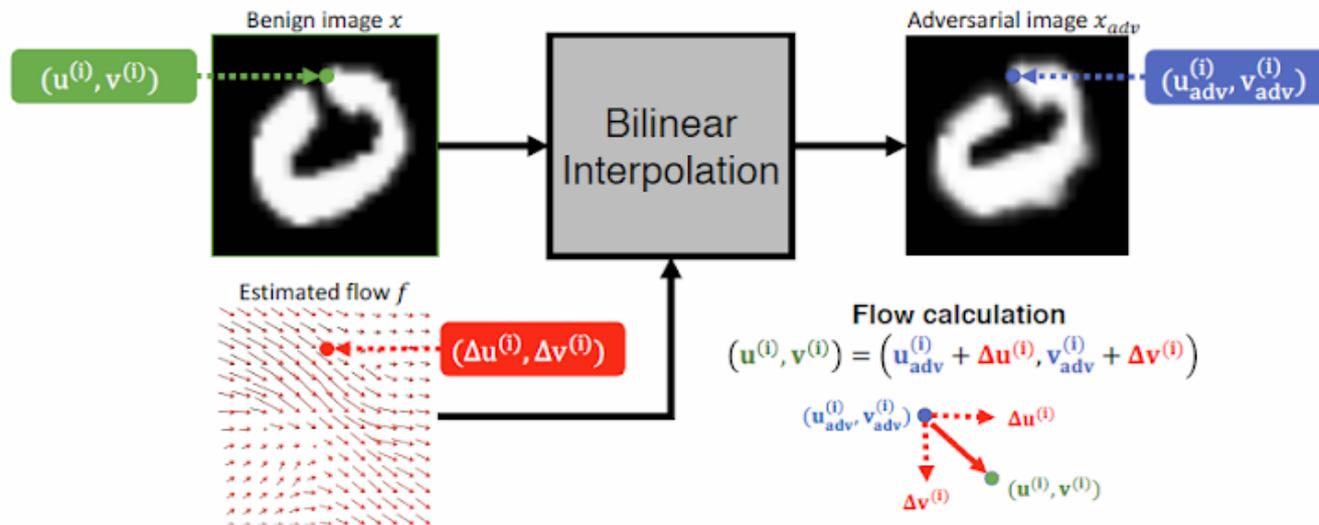
(c) VGG-16



Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard.
Universal adversarial perturbations. In CVPR 2017

Spatially Transformed Adversarial Examples

- A combination of the original adversarial example generation with **spatial transformations**.
- Maximize the network's loss function, in order to predict the target class.
- Minimize the spatial transformation (flow) of the pixels in accordance to their 4 neighbours.



Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, Dawn Song. Spatially Transformed Adversarial Examples. In ICLR 2018



Is there any hope?



Types of System “Errors” or Corruptions

Random Errors

- Unpredictable errors due to limitations on our ability to make physical measurements.
- Cannot be predicted or estimated.
- Most of the time, can be fixed by repeating the experiment or averaging results

Systematic Errors

- Errors that arise from the experimental set-up.
- Errors are consistent, always too large or too small.
- Can be discovered and possibly avoided and corrected.

Gross Error

- Also called human error, these errors arise from mistakes from the experimenter, like laziness, carelessness, ineptitude or intention.
- For example it can be illumination changes, occlusion, pepper/salt noise and more.
- Mostly hard to correct, and the experiment would need to be repeated.



Robust Statistics: Learning with corrupted data

- In robust statistics, the types of corruptions considered are **gross corruptions**.
- Namely, we assume that an ϵ -fraction of our data can be arbitrarily **corrupted**.
- For simplicity, in most cases **additive corruptions** are considered.
- We assume that the adversary has simply added an ϵ -fraction of corrupted data, but we cannot remove them.



Quanquan Gu, Huan Gui, Jiawei Han. Robust Tensor Decomposition with Gross Corruption. In NIPS 2015
Jerry Li. Principled Approaches to Robust Machine Learning and Beyond. Ph.D thesis, 2019

Adversarial Training

- Training with **normal and adversarial images** increases the robustness to adversary.
- This method is different from data augmentation.
- Adversarial examples are unlikely to occur naturally during testing but **expose flaws in the ways that the model conceptualizes its decision function.**
- Usually classification accuracy is lower.
- Cannot defend properly against completely new attack methods.



J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In ICLR 2015.

Defensive Distillation

- A teacher model is trained.
- The **teacher** model provides **soft targets** for a second network, called the **student** network.
- The **student** network is trained to **predict** not the class but the probability distribution over classes.
- Reduces the vulnerability but does not solve the problem.
- Label smoothing can replace the teacher network for simplicity.
- The effectiveness of label smoothing relies on the linearity hypothesis.
- Models should not make extremely confident predictions and that is penalized by label smoothing.
- The model learns a more non-linear function.

N Papernot, P McDaniel, X Wu, S Jha, A Swami. Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks. Proceedings of the 37th IEEE Symposium on Security and Privacy, 2015.

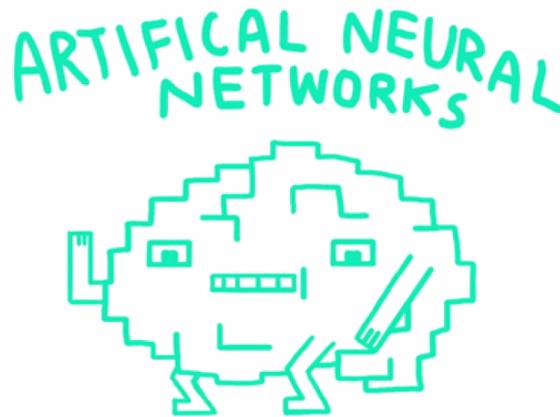
Nicolas Papernot, Patrick McDaniel. Extending Defensive Distillation, arXiv:1705.05264, 2017.

Tamir Hazan, George Papandreou and Daniel Tarlow. Perturbation, optimization and statistics. The MIT Press, 2016



Future Approaches

- Instead of falling in the pitfall of defending against specific attacks, investigate ways to build more robust models.
- Focus on the quality of the training and avoid overfitting.
- Train models with uncertainty so they do not give over-confident answers in cases of ambiguity.
- Detection is beneficial in high risk cases.
- Use adversarial examples for a useful purpose, like encryption or benchmarking.





Thank you!
Questions?

“People tend to trust each other in machine learning. The security mindset is exactly the opposite, you have to be always suspicious that something bad may happen.”

Battista Biggio