Adversarial Examples

Hanxiao Liu

April 2, 2018

Adversarial Examples

"Inputs to ML models that an attacker has intentionally designed to cause the model to make a mistake" [1]

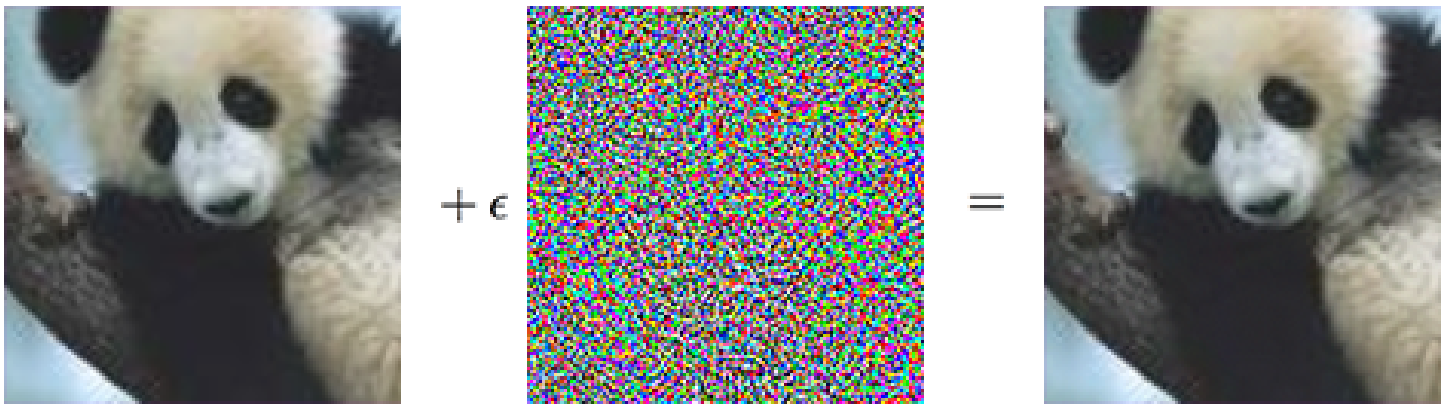
Why this is interesting:

- Safety.
- Interpretability.
- Generalization.

[1] https://blog.openai.com/adversarial-example-research/

Adversarial Examples

Fooling GoogLeNet (Inception) on ImageNet.



Adversarial Examples

Fooling a linear model (logistic regression) on ImageNet.

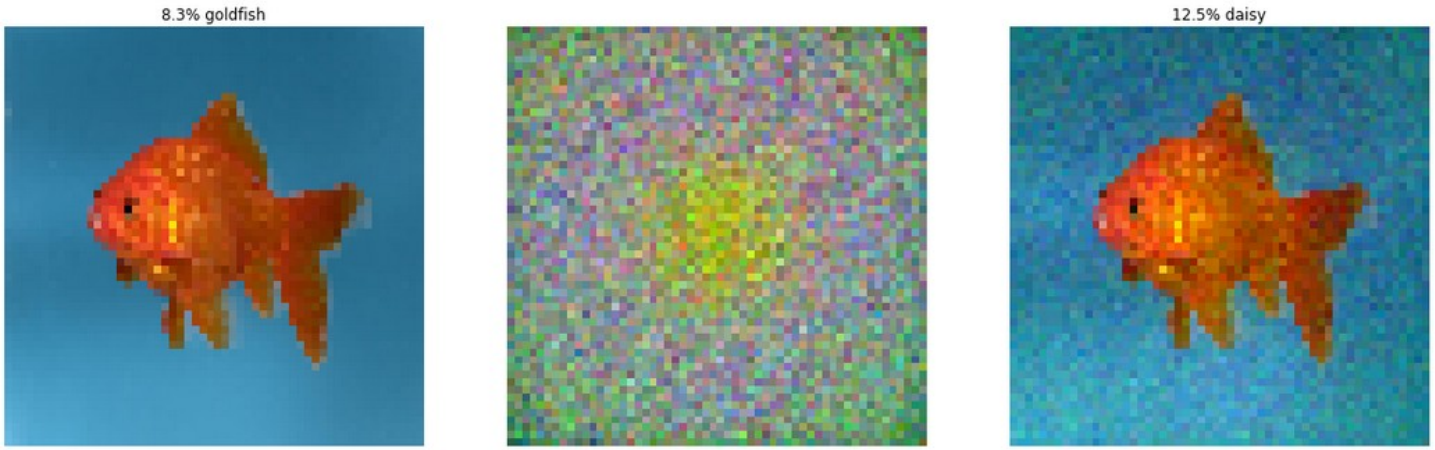


Figure : Before: 8.3% Goldfish; After: 12.5% Daisy.

Adversarial Examples in Language Understanding

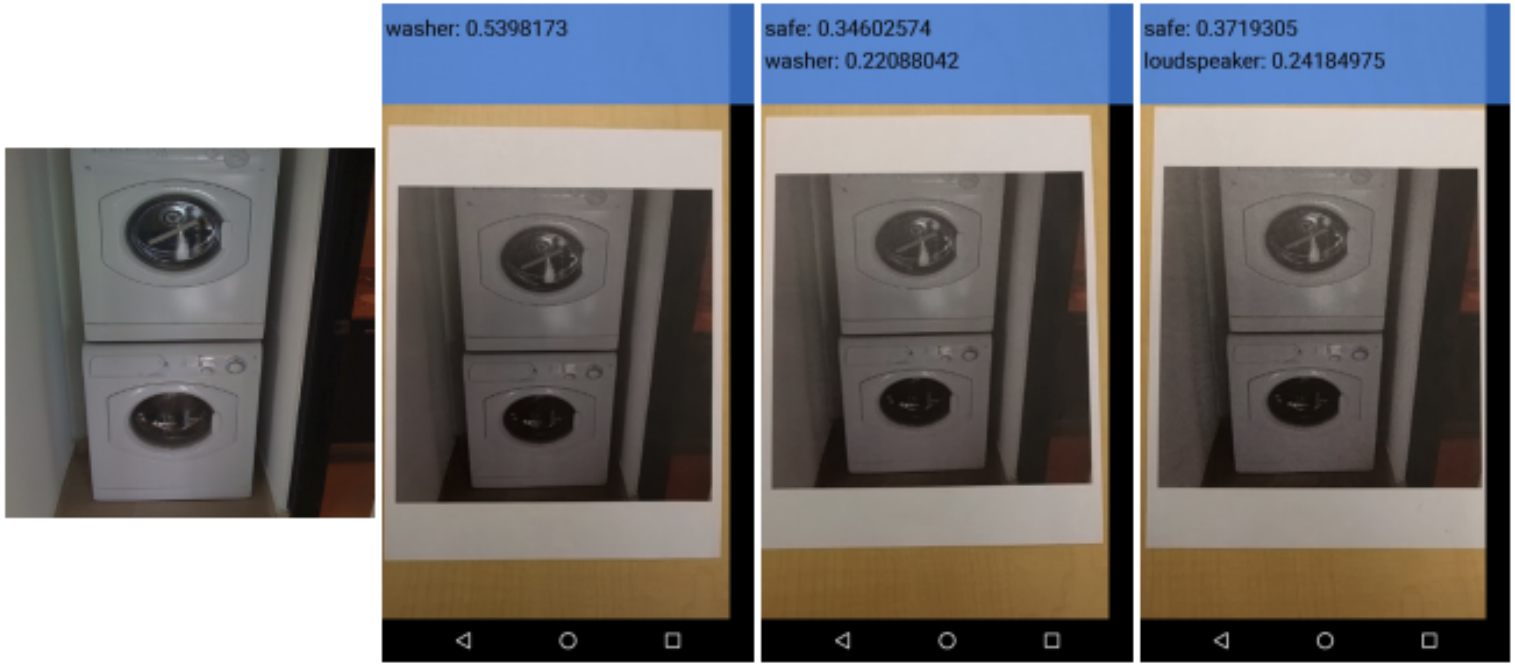
[Jia and Liang, 2017]

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Figure : Fooling BiDAF on SQuAD.

Adversarial Examples in the Physical World

[Kurakin et al., 2016]
Attaching a mask over the phone camera:

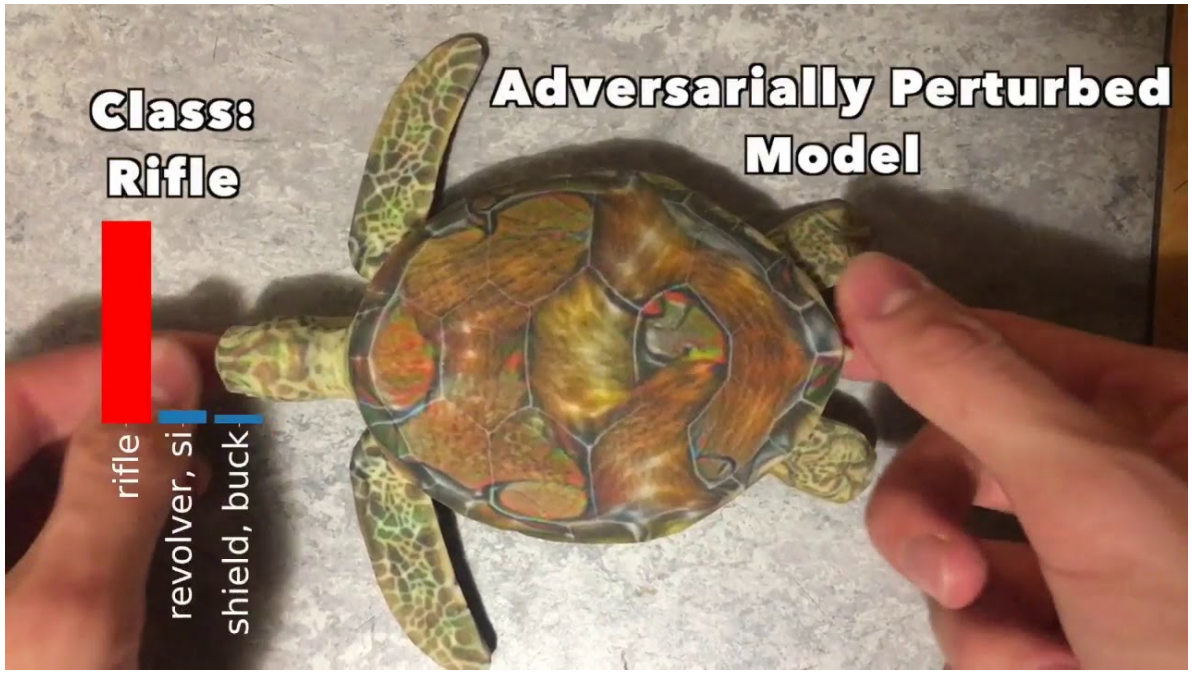


(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

https://www.youtube.com/watch?v=piYnd_wYlT8

Adversarial Examples in the Physical World

[Athalye et al., 2018]
Adversarial example using 3D-printing ...



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

Autonomous Vehicles

[Evtimov et al., 2017]

Figure : Before: Stop sign; After: 45 mph sign

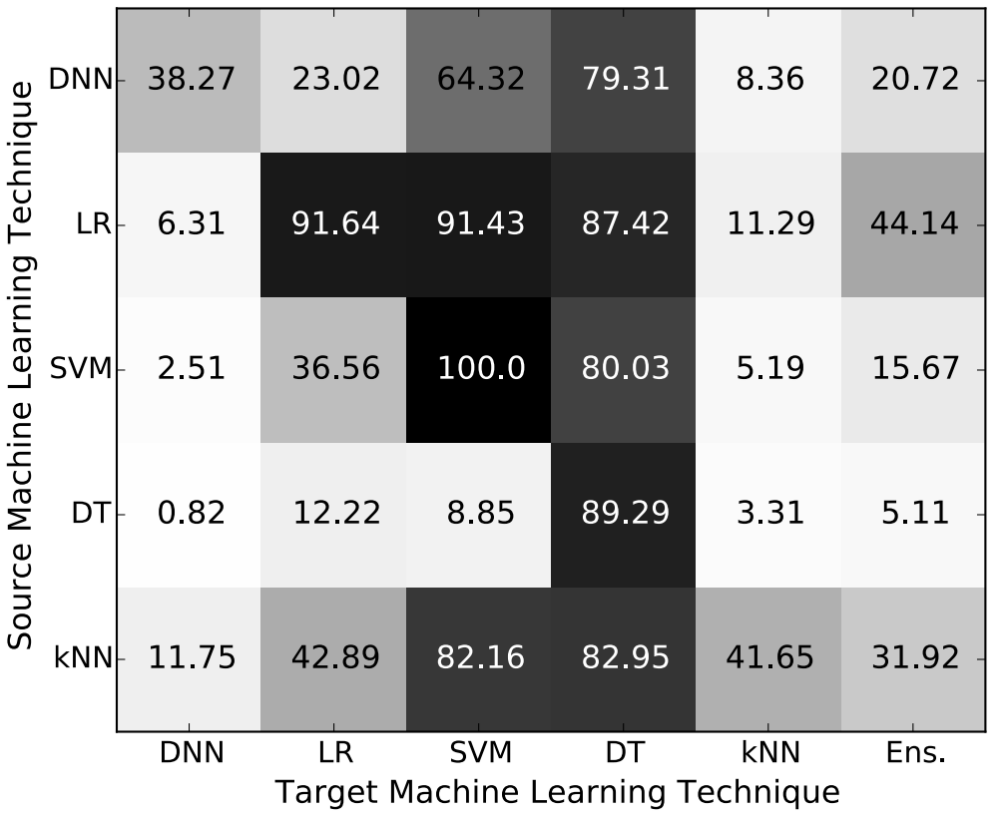
[Lu et al., 2017] argues existing systems are robust:

- A moving camera is able to view objects from different distances and different angles.

Specialized attacks for object detection systems?

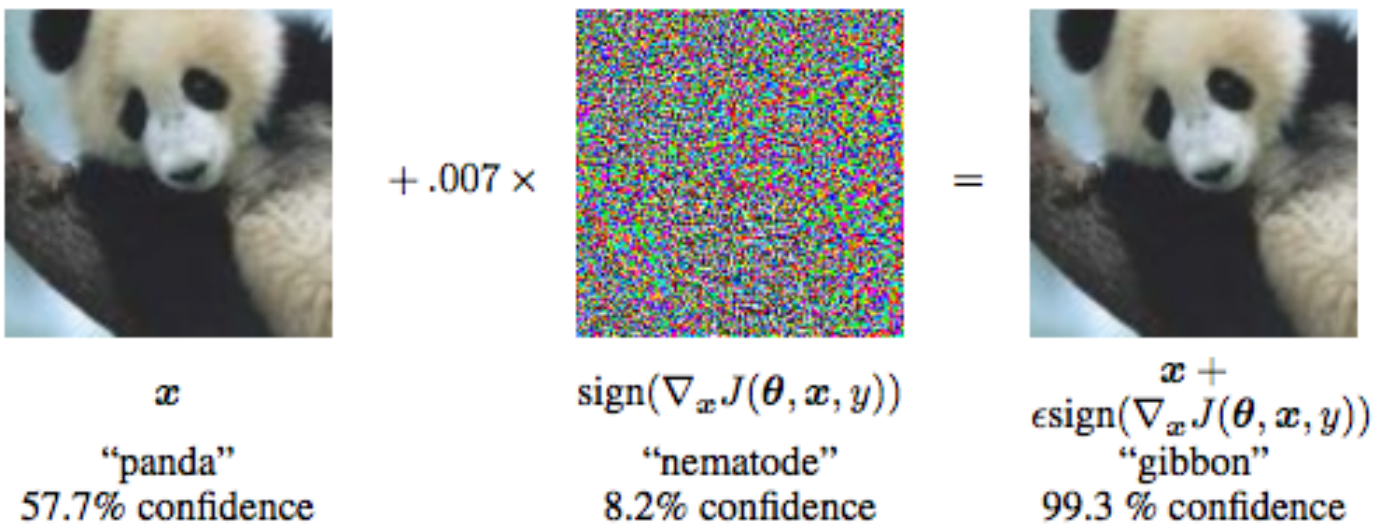
Transferability

Adversarial examples are transferable across ML models [Papernot et al., 2017].



Creating Adversarial Examples

Simple approach: Fast Gradient Sign Method (FGSM) [Goodfellow et al., 2014]



Other techniques: Iterative FGSM [Kurakin et al., 2016], L-BFGS [Szegedy et al., 2013], ...

Creating Adversarial Examples

One Pixel Attack [Su et al., 2017]

$$\max_{m} \quad f_{adv}(x+m) \quad \text{s.t.} \quad \|m\|_0 \leq 1 \tag{1}$$

Planetarium

Mosque(7.81%)

Comforter

Pillow(6.83%)

Defense

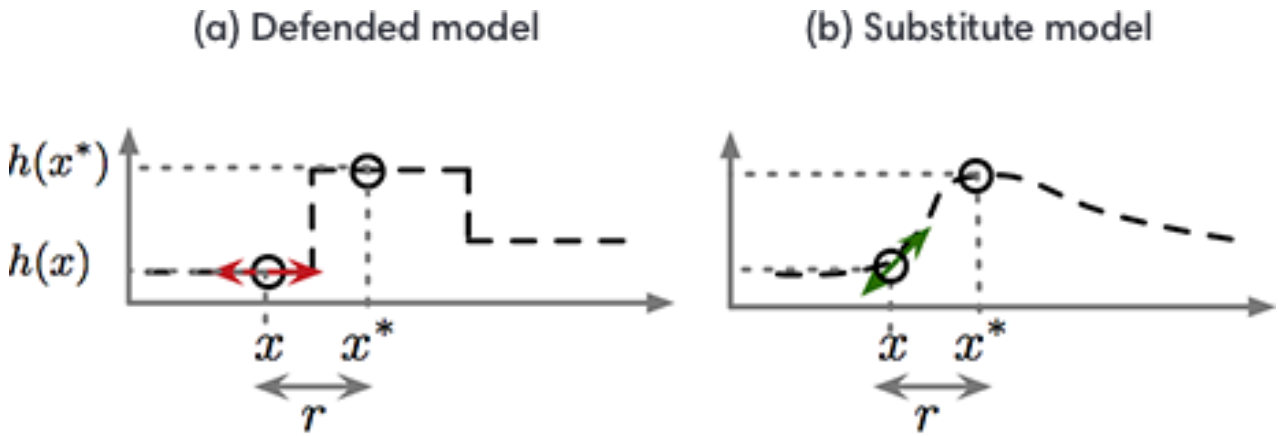
- Data Augmentation (e.g., dropout, cutout, mixup).
- Adversarial Training.
  - Generate adversarial examples and include them as part of the training data.
- Distillation/Smoothing.

Defense

Hiding information (e.g. gradient) from the attackers?

Black box attack [Papernot et al., 2017]

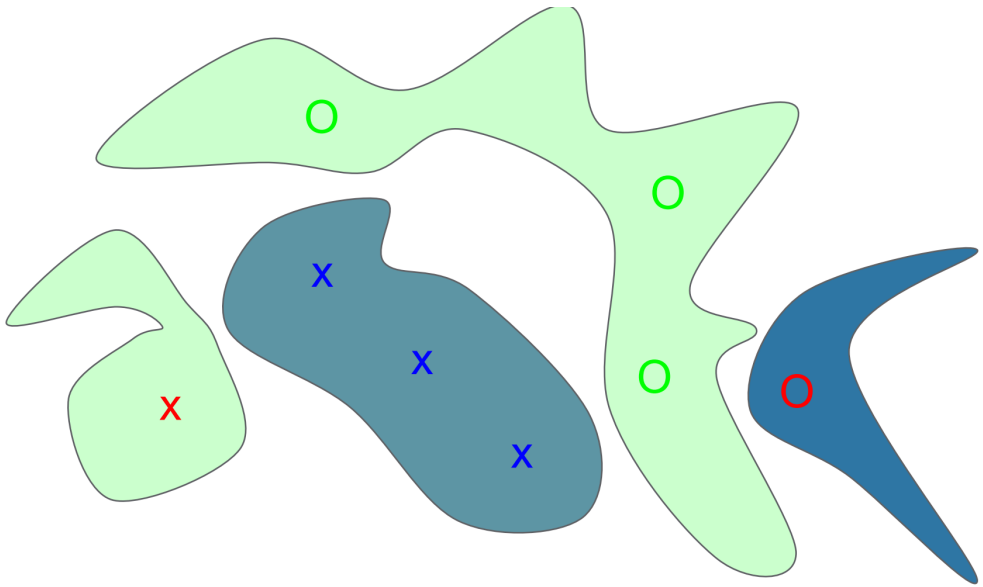
- Train a "substitute model", compute adversarial examples there and transfer them to the target model.



Why ML models are prone to adversary?

Conjecture 1: Overfitting.

- Nature images are within the correct regions but are also sufficiently close to the decision boundary.

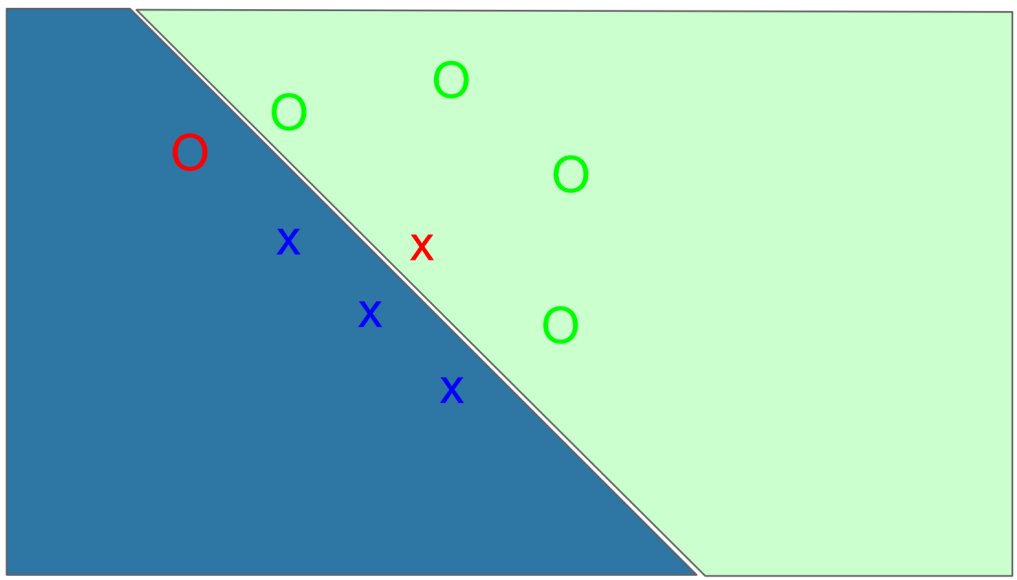


(Goodfellow 2016)

Why ML models are prone to adversary?

Conjecture 2: Excessive Linearity.

- Decision boundary for most ML models are (near-)piecewise linear.
- In high dimension, $w^\top x$ is prone to perturbation.



(Goodfellow 2016)

Why ML models are prone to adversary?

Empirical observation: nearly linear responses over ϵ.

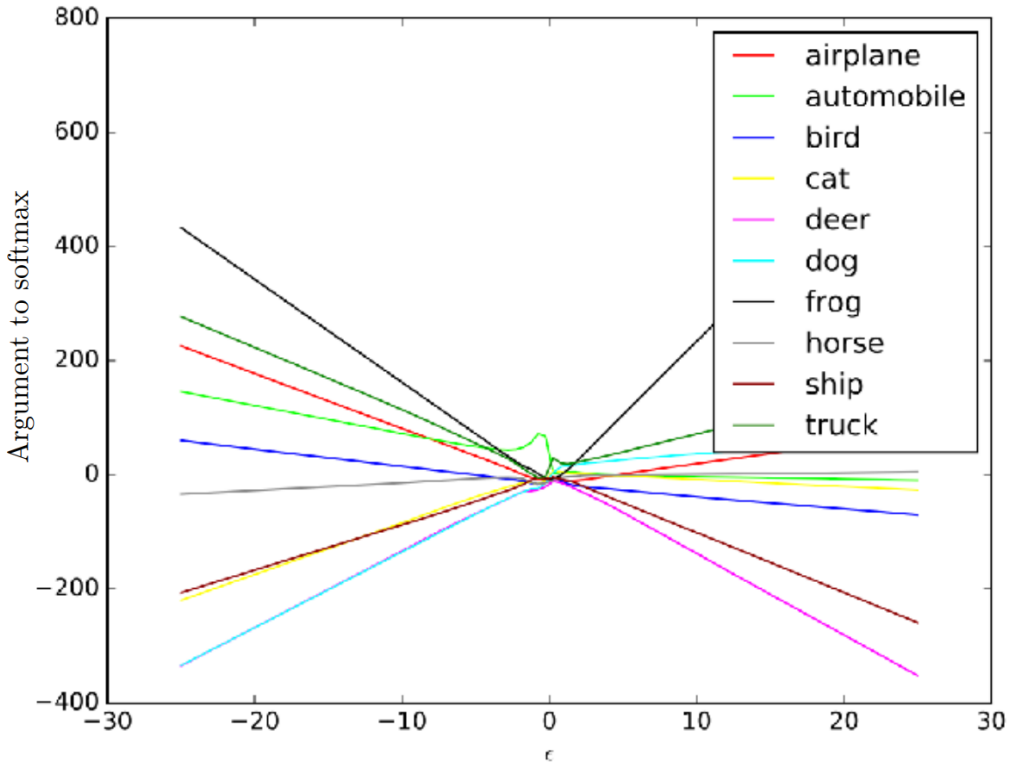


Figure : How ϵ affects the softmax logits on CIFAR-10. [Goodfellow et al., 2014]

Interpretability

Why this is relevant?

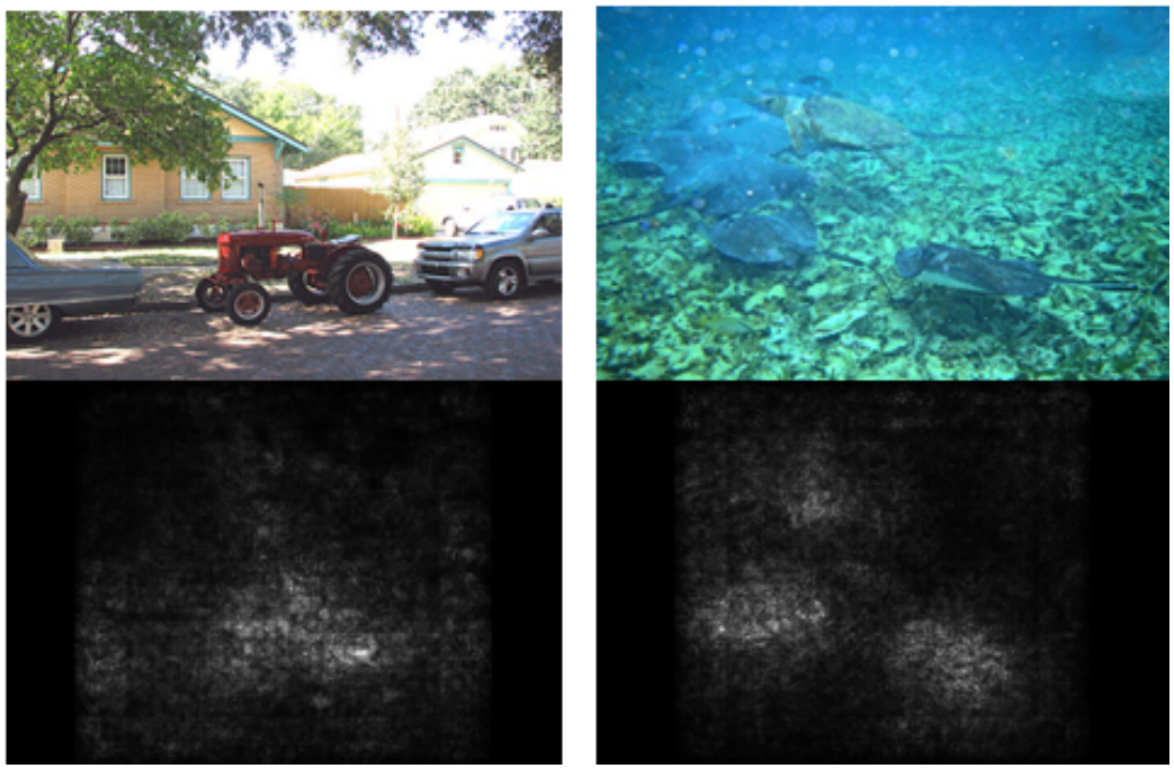
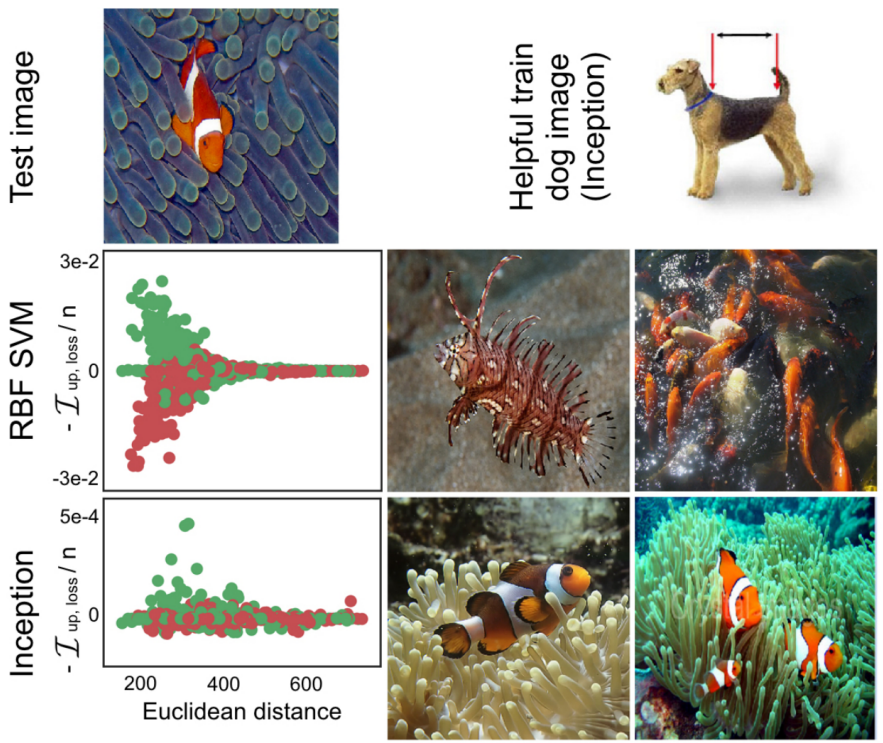


Figure : $\nabla_x f(x)$ reveals the salient features of x.
[Simonyan et al., 2013]

Interpretability via Influence Functions

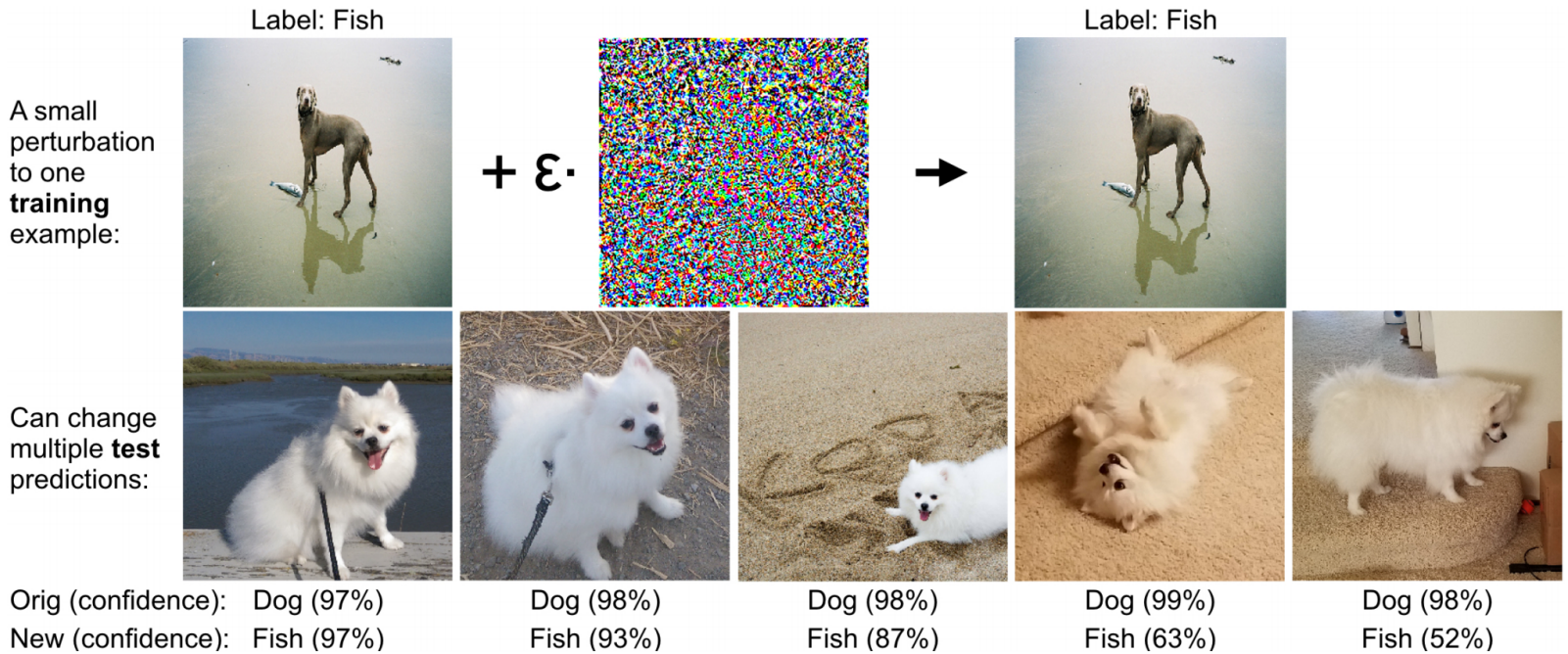
[Koh and Liang, 2017]: Identifying training points most responsible for a given prediction.

- How would the model's predictions change if we did not have this training point?



Interpretability via Influence Functions

[Koh and Liang, 2017]
The learned influence function allows us to create adversarial *training* (not testing!) examples.



Reference I


Athalye, A., Engstrom, L., Ilyas, A., and Kwok, K. (2018).
Synthesizing robust adversarial examples.


Evtimov, I., Eykholt, K., Fernandes, E., Kohno, T., Li, B., Prakash, A., Rahmati, A., and Song, D. (2017).
Robust physical-world attacks on deep learning models.
arXiv preprint arXiv:1707.08945, 1.


Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014).
Explaining and harnessing adversarial examples.
arXiv preprint arXiv:1412.6572.


Jia, R. and Liang, P. (2017).
Adversarial examples for evaluating reading comprehension systems.
arXiv preprint arXiv:1707.07328.


Koh, P. W. and Liang, P. (2017).
Understanding black-box predictions via influence functions.
arXiv preprint arXiv:1703.04730.


Kurakin, A., Goodfellow, I., and Bengio, S. (2016).
Adversarial examples in the physical world.
arXiv preprint arXiv:1607.02533.

Reference II


Lu, J., Sibai, H., Fabry, E., and Forsyth, D. (2017).
No need to worry about adversarial examples in object detection in autonomous vehicles.
arXiv preprint arXiv:1707.03501.


Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. (2017).
Practical black-box attacks against machine learning.
In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pages 506–519. ACM.


Simonyan, K., Vedaldi, A., and Zisserman, A. (2013).
Deep inside convolutional networks: Visualising image classification models and saliency maps.
arXiv preprint arXiv:1312.6034.


Su, J., Vargas, D. V., and Kouichi, S. (2017).
One pixel attack for fooling deep neural networks.
arXiv preprint arXiv:1710.08864.

Reference III


Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013).
Intriguing properties of neural networks.
arXiv preprint arXiv:1312.6199.