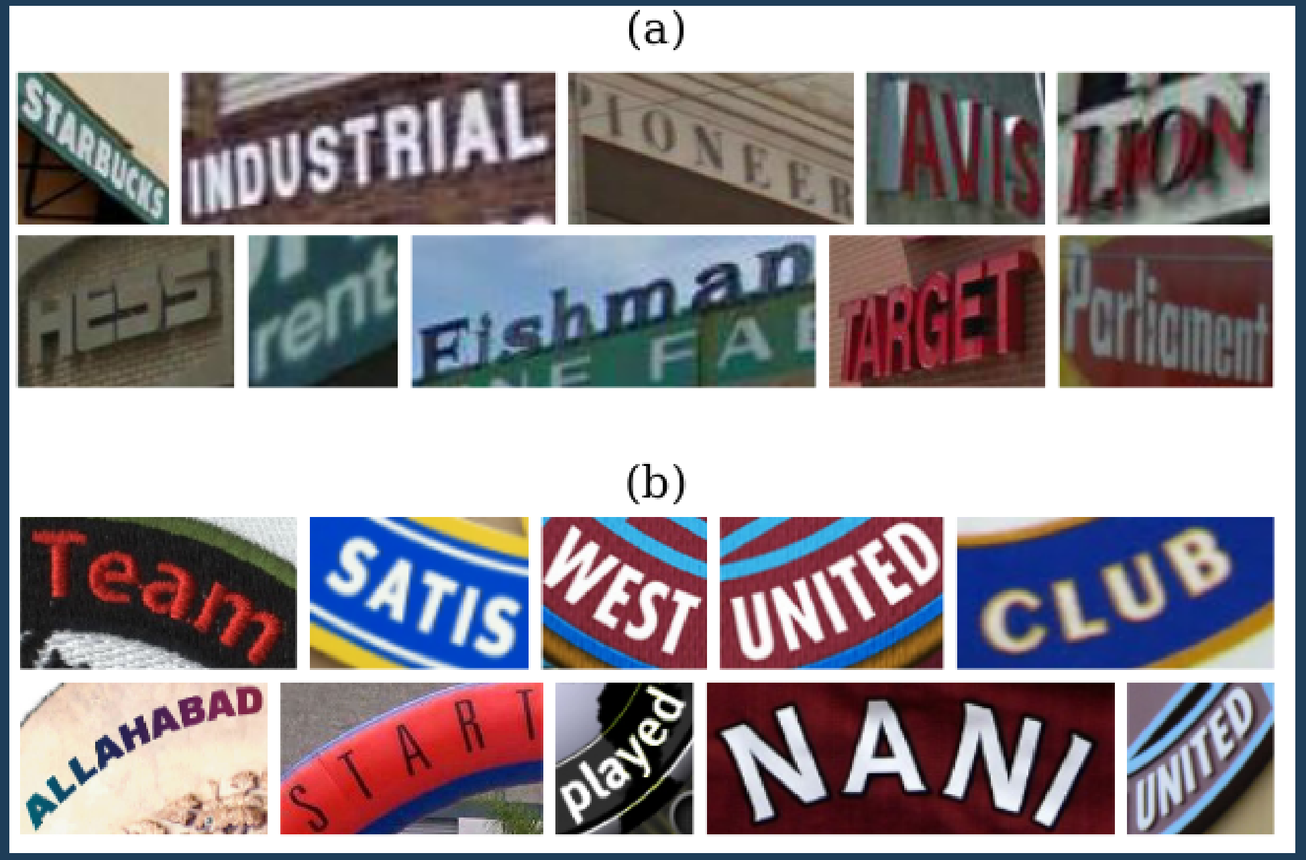
DEEP LEARNING

ROBUSTNESS OF NEURAL NETWORKS

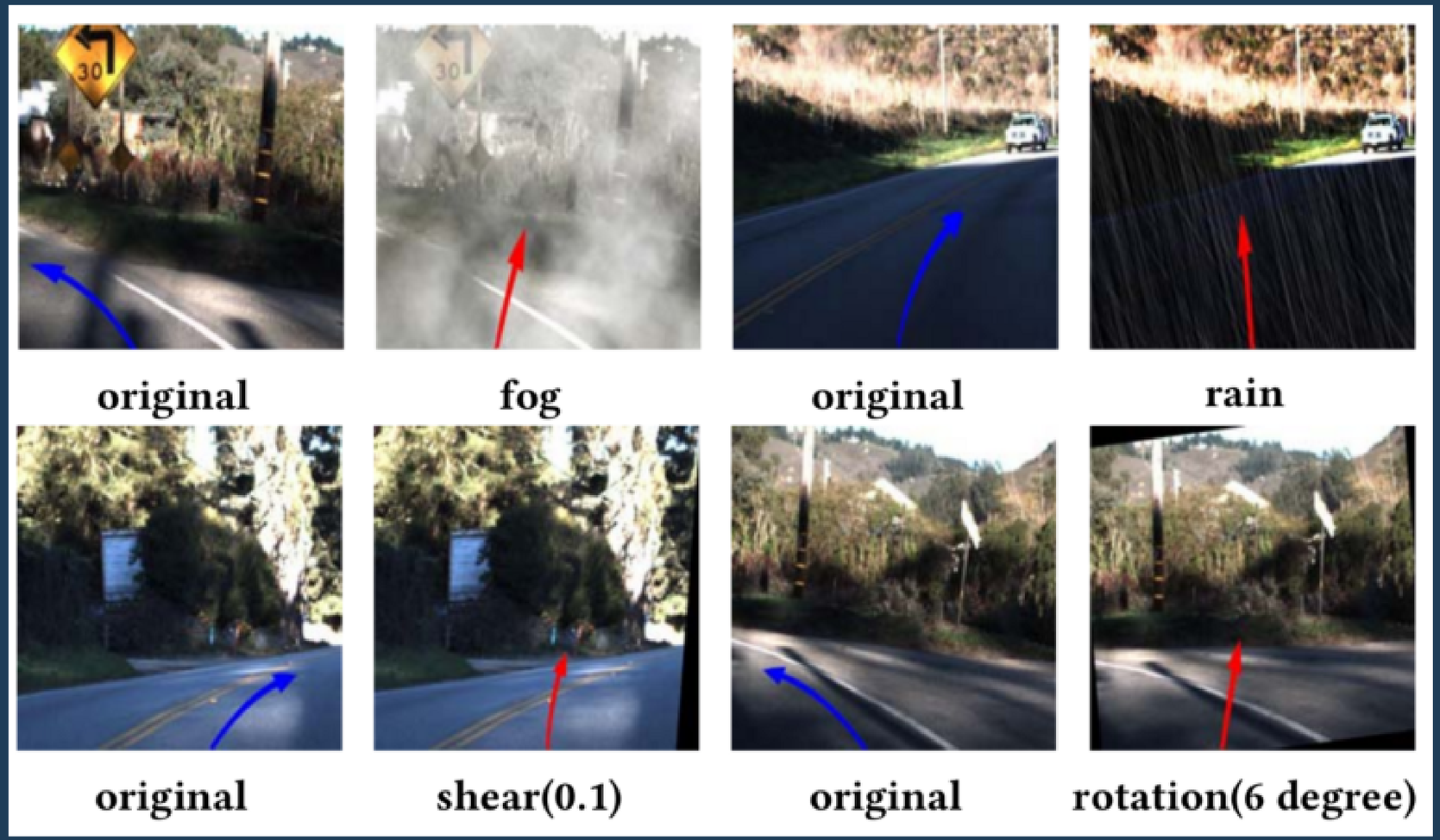
Presented by Vitória Guardieiro

What is robustness?

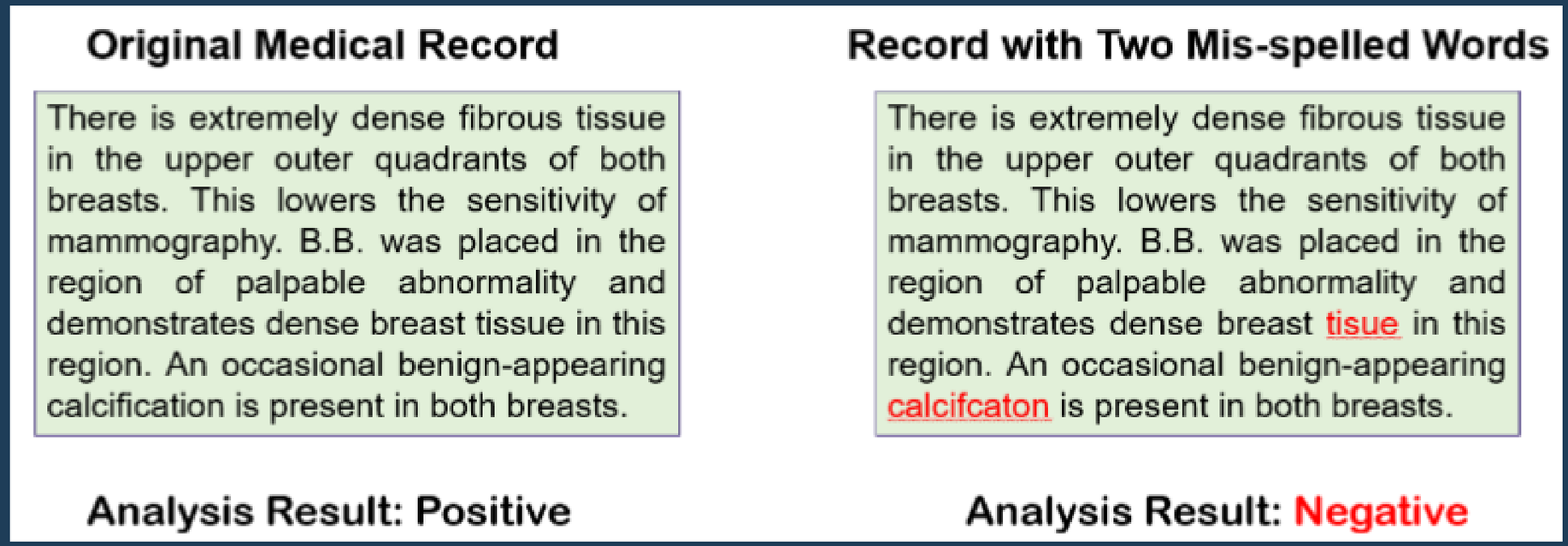
THE MODEL'S PREDICTION SHOULD BE STABLE
TO SMALL VARIATIONS IN THE INPUT



"*Robust Scene Text Recognition with Automatic Rectification*", in CVPR 2016.



"*DeepTest: Automated Testing of Deep-Neural-Network-driven Autonomous Cars*", in ICSE '18.



"*Robust Scene Text Recognition with Automatic Rectification*", in CVPR 2016.

Adversarial Robustness

A DEFENSE AGAINST ADVERSARIAL ATTACKS

Adversarial Attacks

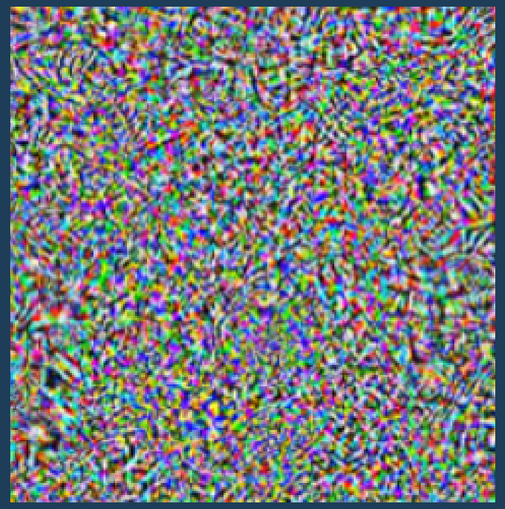
Adversarial Attacks

"pig" (91%)

Adversarial Attacks

"pig" (91%)

+ 0.005x

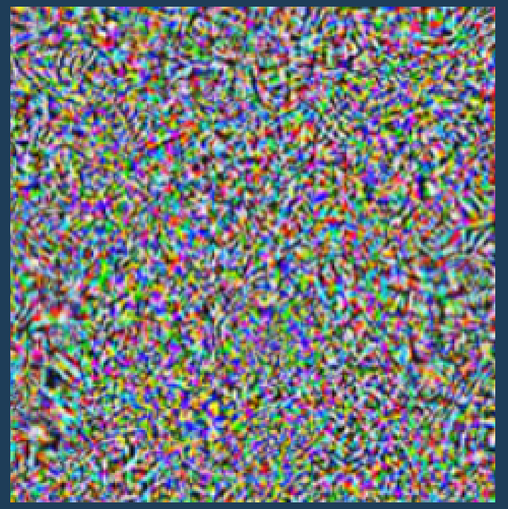


noise
(not random)

Adversarial Attacks

"pig" (91%)

+ 0.005x



noise
(not random)

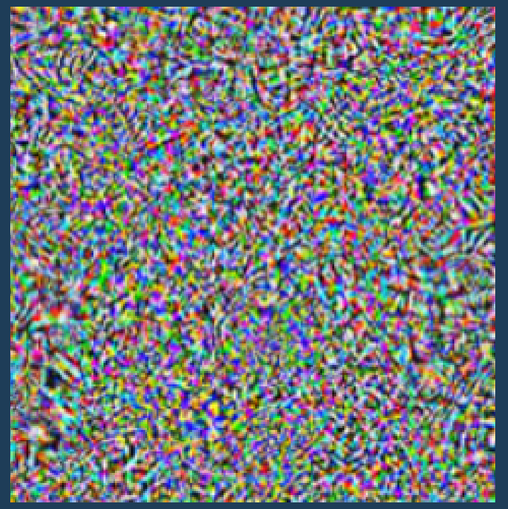
=

?

Adversarial Attacks

"pig" (91%)

+ 0.005x



noise
(not random)

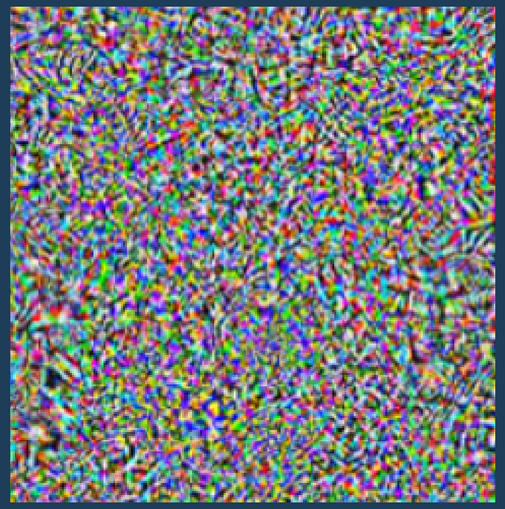
=

"airliner" (99%)

Adversarial Robustness [3]

"pig" (91%)

+ 0.005x



noise
(not random)

=

~~"airliner" (99%)~~

"pig"

MAIN AREAS IN ADVERSARIAL ROBUSTNESS

ATTACKS

How to generate adversarial perturbation effectively and efficiently

DEFENSES

Methods to improve model robustness and avoid adversarial attacks

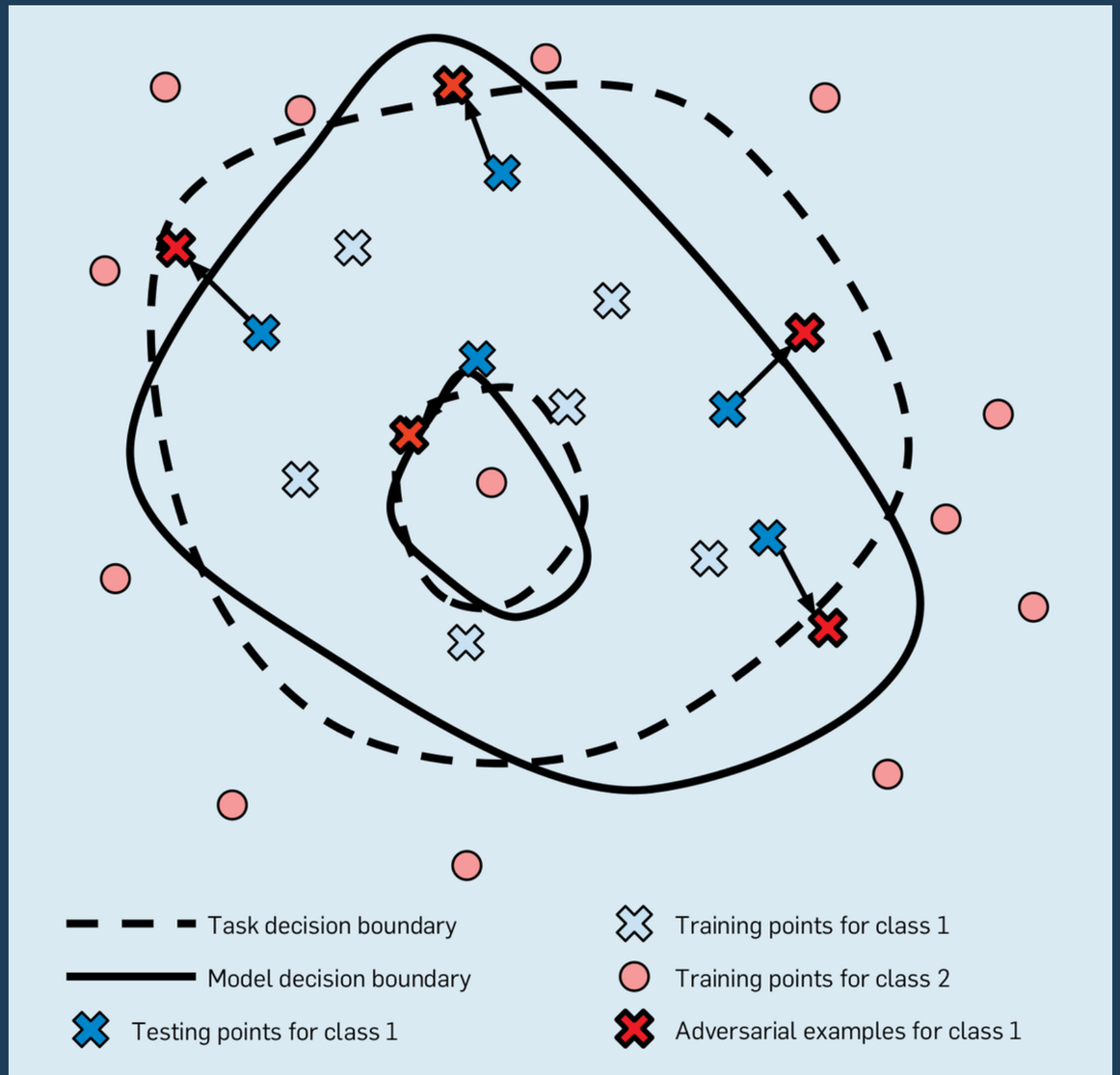
VERIFICATION

Techniques for checking whether a convolutional neural network is robust

Adversarial Attacks

Why are there adversarial attacks?

Why are there adversarial attacks?



The adversarial model looks for mismatch between the actual decision boundary and the model's learned decision boundary [4]

CATEGORIES OF ADVERSARIAL ATTACKS [5]

WHITE-BOX V.S. BLACK-BOX

What kind of information is required from the model?

TARGETED V.S. UN-TARGETED

The attacker is/is not able tocontrol the resulting misclassification label

LOCAL VS. UNIVERSAL

Perturbation can fool one/a set of inputs

DISTANCE METRIC

Lp-norm, total variation of pixel displacemen (images), similarity of sentences or text (NLP), ...

FGSM Attack [6]

The Fast Gradient Sign Method[6] is able to find adversarial perturbations with a fixed L∞-norm constraint very efficiently

It is a white-box attack, because it uses the model's loss function

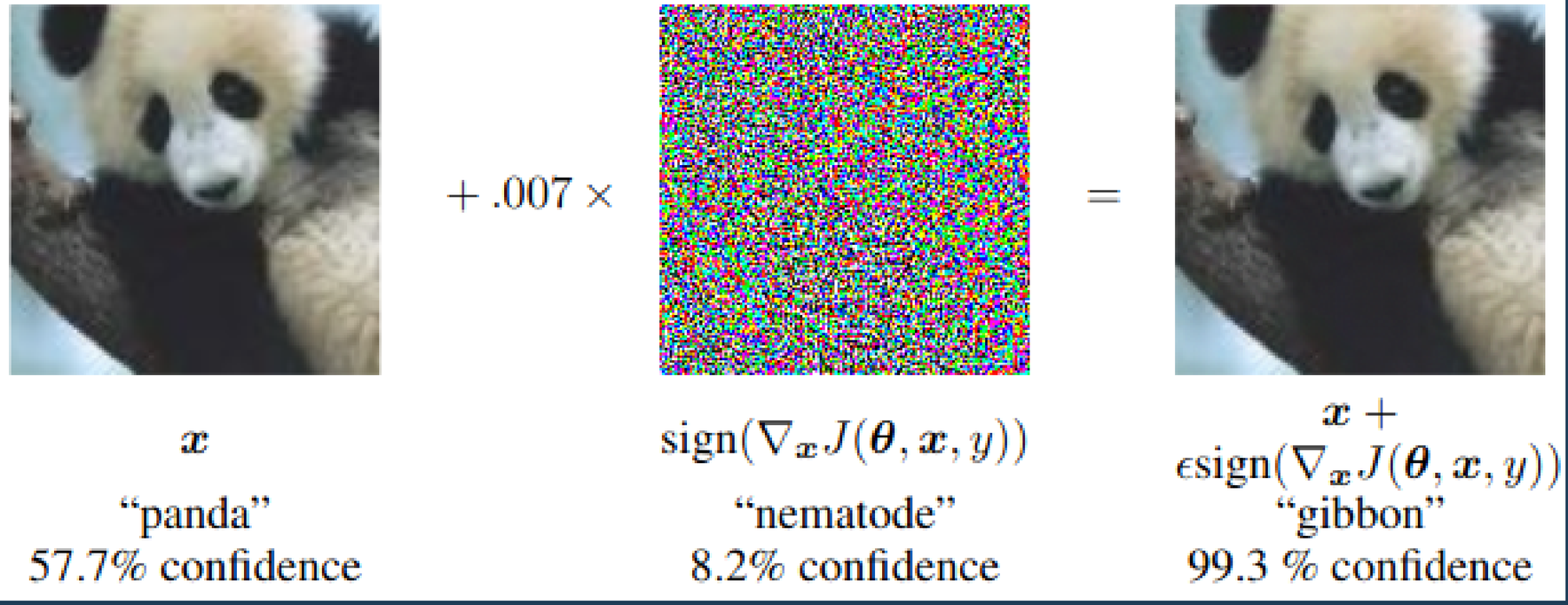
FGSM Attack [6]

Let θ be the parameters of a model, x the input to the model, y the targets associated with x and J(θ,x,y) be the loss function used to train the neural network.

Find adversarial perturbation r by linearizing the loss function around the current value of θ:

$$r = \boxtimes \text{ sign } (\nabla_x J(\theta,x,y))$$

FGSM Attack [6]



Existing Tools for Adversarial Attacks

ADVERSARIAL ROBUSTNESS TOOLBOX (ART)

https://github.com/Trusted-AI/adversarial-robustness-toolbox

FOOLBOX NATIVE

https://github.com/bethgelab/foolbox

CLEVERHANS

https://github.com/tensorflow/cleverhans

ADVBOX FAMILY

https://github.com/advboxes/AdvBox

Adversarial Defenses

PROTECTING AGAINST ATTACKS

Training dataset augmentation [7]

Retrains the model with adversarial attack samples

For instance, via an adversarial process, in which we simultaneously train two models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G

Adversarial Training via Robust Optimisation [8]

Training is made through a robust optimization technique

For each data point x, we introduce a set of allowed perturbations S ⊆ R^d that formalizes the manipulative power of the adversary

Then they modify the definition of population risk E_D[L] by incorporating the adversary:

$$\min_{\theta} \rho(\theta), \quad \text{where} \quad \rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta\in\mathcal{S}} L(\theta, x+\delta, y) \right]$$

Random smoothing [9]

Modifies the network by introducing a noise layer that adds zero mean noise to the output of the layer preceding it

The noise's standard deviation is proportional to:

1. L, the p-norm attack bound and
2. Δ, the sensitivity of the pre-noise computation with respect to p-norm input changes.

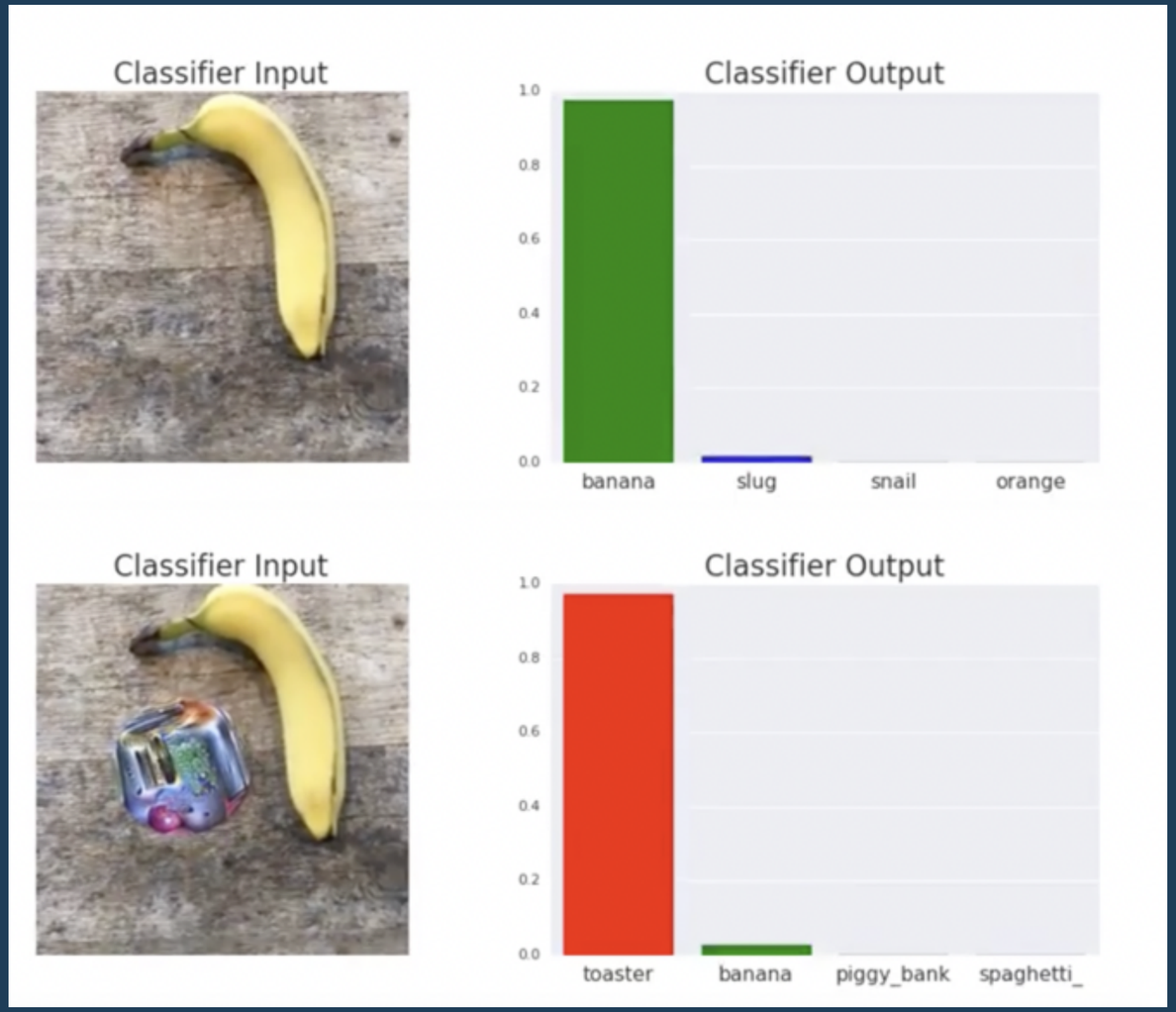
Aplications & Critics

WHAT IS THE POINT?

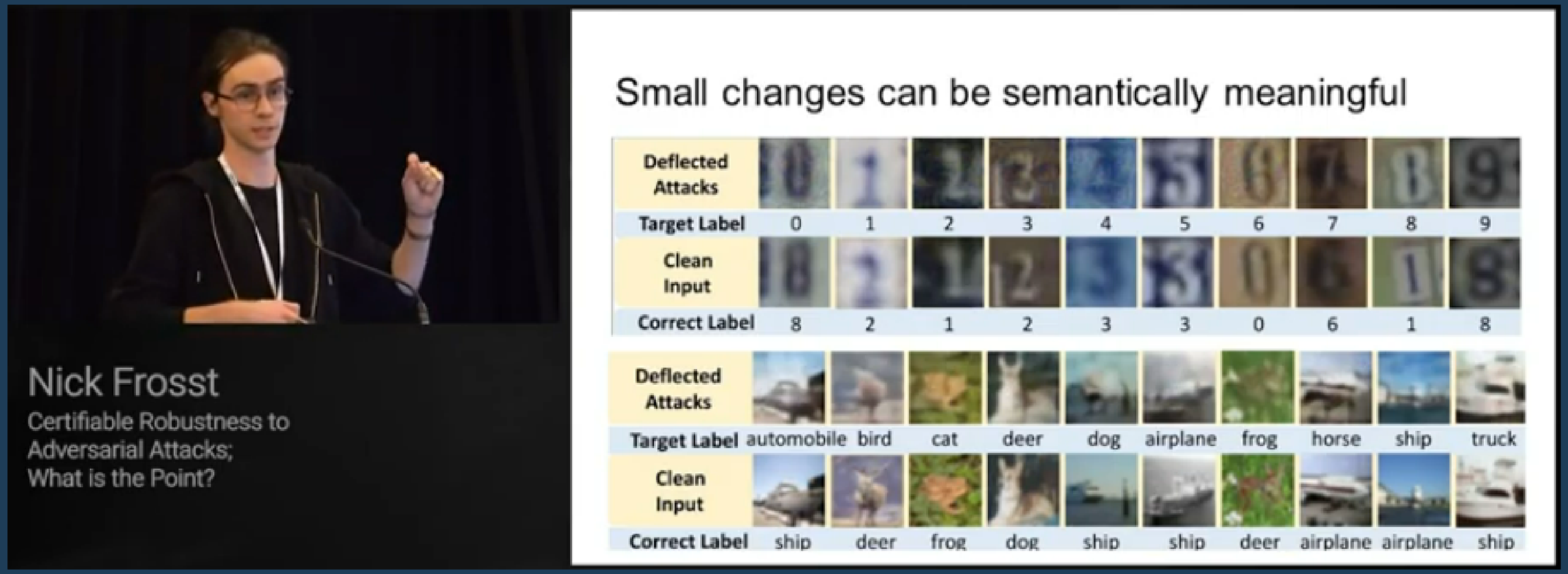
Security against attacks [10]

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%

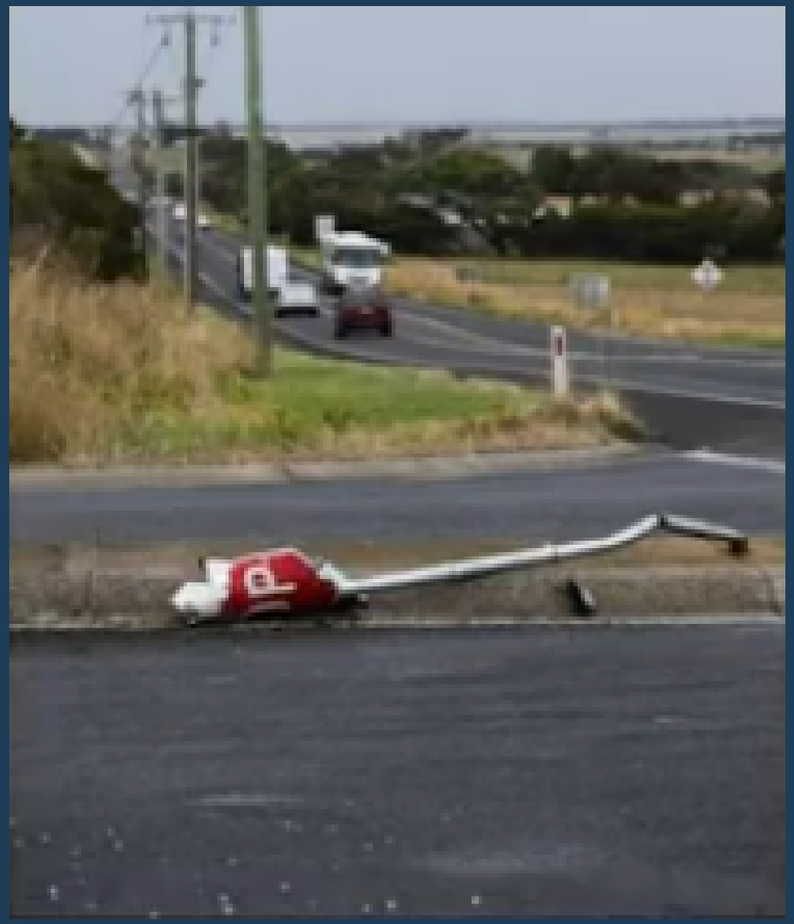
Security against attacks [11, 12]



Accuracy loss



Are adversarial attacks plausible?



References

1 O. Bastani, Y. Ioannou, L. Lampropoulos, D. Vytiniotis, A. V. Nori,and A. Criminisi, "*Measuring neural net robustness with constraints*", in NIPS'16.

2 G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer, "*Reluplex: An efficient smt solver for verifying deep neural networks*", in Computer Aided Verification, 2017.

3 W. Ruan, E. Botoeva, X. Yi, X. Huang, Tutorial "*Towards Robust Deep Learning Models: Verification, Falsification, and Rectification*", The 30th International Joint Conference on Artificial Intelligence (IJCAI 2021), 21-26 Aug 2021, Canada
Link: http://tutorial-ijcai.trustai.uk/

References

4 Ian Goodfellow, Patrick McDaniel, Nicolas Papernot, "*Making Machine Learning Robust Against Adversarial Inputs*". Communications of the ACM, 61(7), 56–66. 2018

5 Xiaowei Huang and Daniel Kroening and Wenjie Ruan and James Sharp and Youcheng Sun and Emese Thamo and Min Wu and Xinping Yi, "*A Survey of Safety and Trustworthiness of Deep Neural Networks: Verification, Testing, Adversarial Attack and Defence, and Interpretability*". 2020.

6 Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy, "*Explaining and Harnessing Adversarial Examples*", 2014.

References

7 Ian J. Goodfellow and Jean Pouget-Abadie and Mehdi Mirza and Bing Xu and David Warde-Farley and Sherjil Ozair and Aaron Courville and Yoshua Bengio, “*Generative Adversarial Nets*”. 2014.

8 Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu. "*Towards Deep Learning Models Resistant to Adversarial Attacks*". ICLR 2018.

9 Mathias Lecuyer and Vaggelis Atlidakis and Roxana Geambasu and Daniel Hsu and Suman Jana, “*Certified Robustness to Adversarial Examples with Differential Privacy*”, 2019.

References

10 Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, Dawn Song, "*Robust Physical-World Attacks on Deep Learning Models*", CVPR 2018

11 Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, Justin Gilmer, "*Adversarial Patch*". NeurIPS, 2017.

12 Ranjie Duan, Xingjun Ma, Yisen Wang, James Bailey, A. K. Qin, Yun Yang, "*Adversarial Camouflage: Hiding Physical-World Attacks with Natural Styles*", CVPR 2020.

References

13 Nick Frosst, "*Certifiable Robustness to adversarial Attacks; What is the Point?*". Toronto Machine Learning Summit, 2020. https://www.youtube.com/watch?v=OfSxYqU-6s0&t=889s

14 Wenjie Ruan, Xinping Yi, Xiaowei Huang, "*Adversarial Robustness of Deep Learning: Theory, Algorithms, and Applications*", ICDM 2020

THANK YOU!