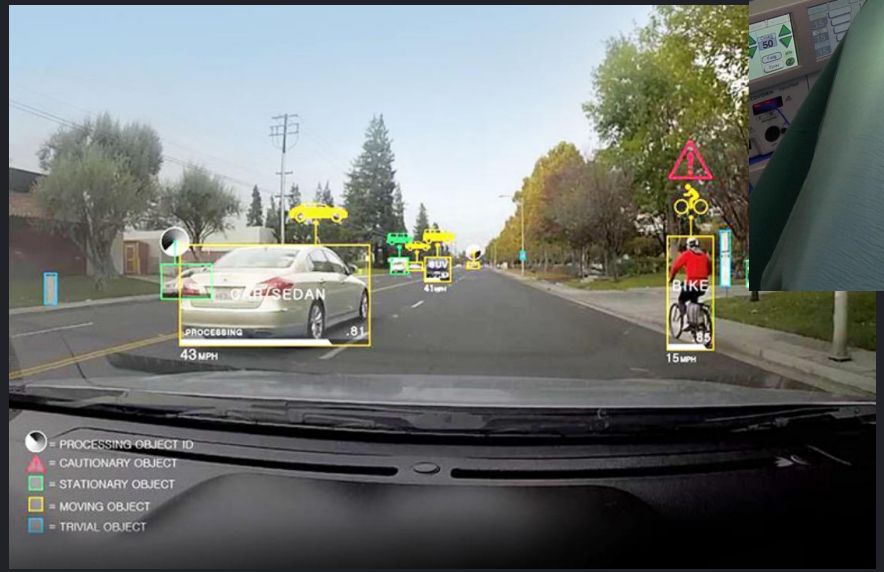
Poking holes in your deep learning vision model

Irina Vidal Migallón

2019.05.02 PyData Hamburg

Source: microsoftinsider.es, Hospital G. Marañón (Madrid, Spain) & Exovite



Source: NVIDIA

TESLA & UBER

2016: White truck on white

2018: Fire trucks

2018: Dismissed pedestrian

210,000,000 km driven by the time of the accident
[source]

A BIT ABOUT ME

Applied research (INRIA)

Startups:

- MedTechs in Madrid, Paris & Berlin
- AR/MR in Berlin

Siemens Mobility

ROBUSTNESS?

ROBUSTNESS?

EVALUATION

EVALUATION

Model	Metric
Baseline	
New Model	

EVALUATION

Model	Metric
Baseline	
New Model	

Automated

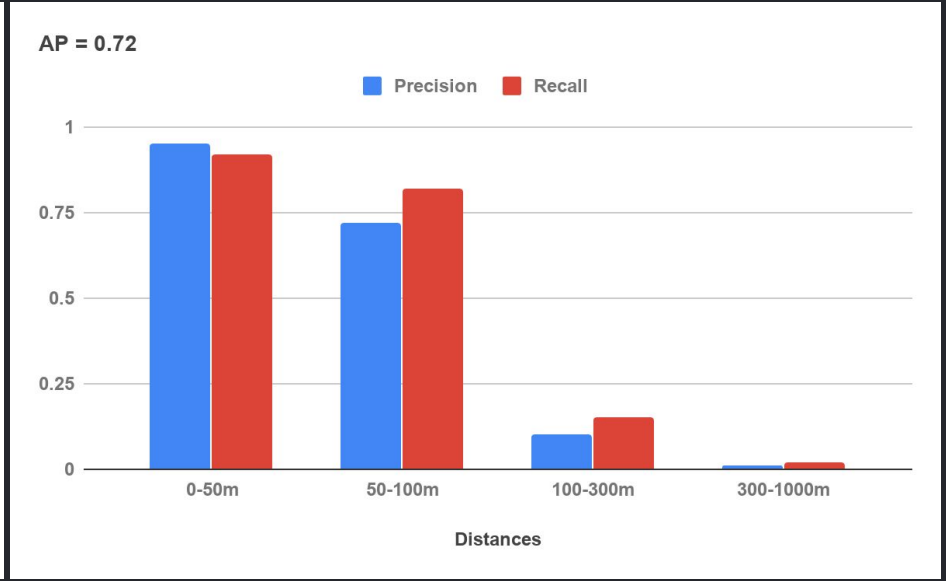
Fast enough to iterate

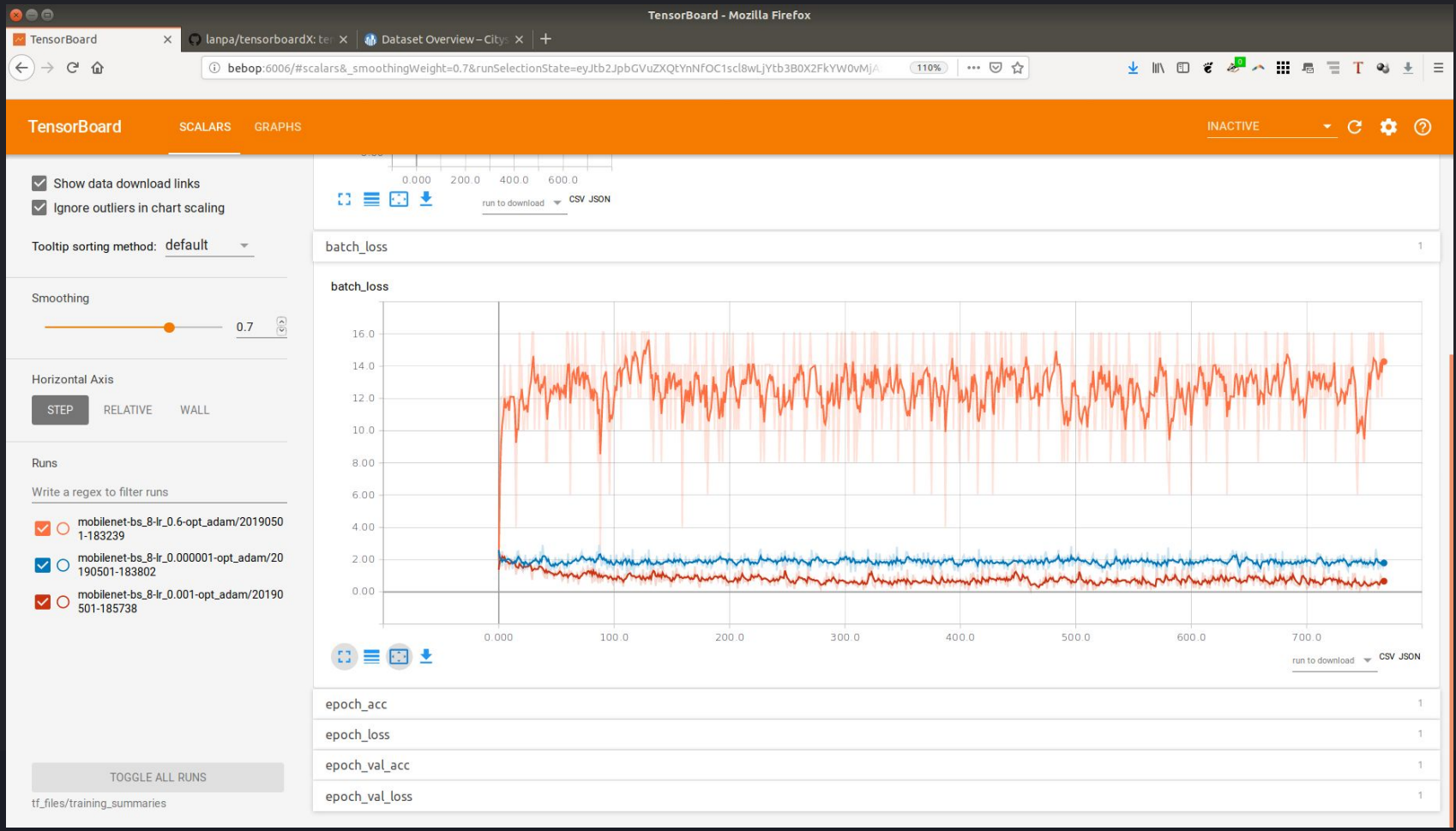
Traced

Close to production

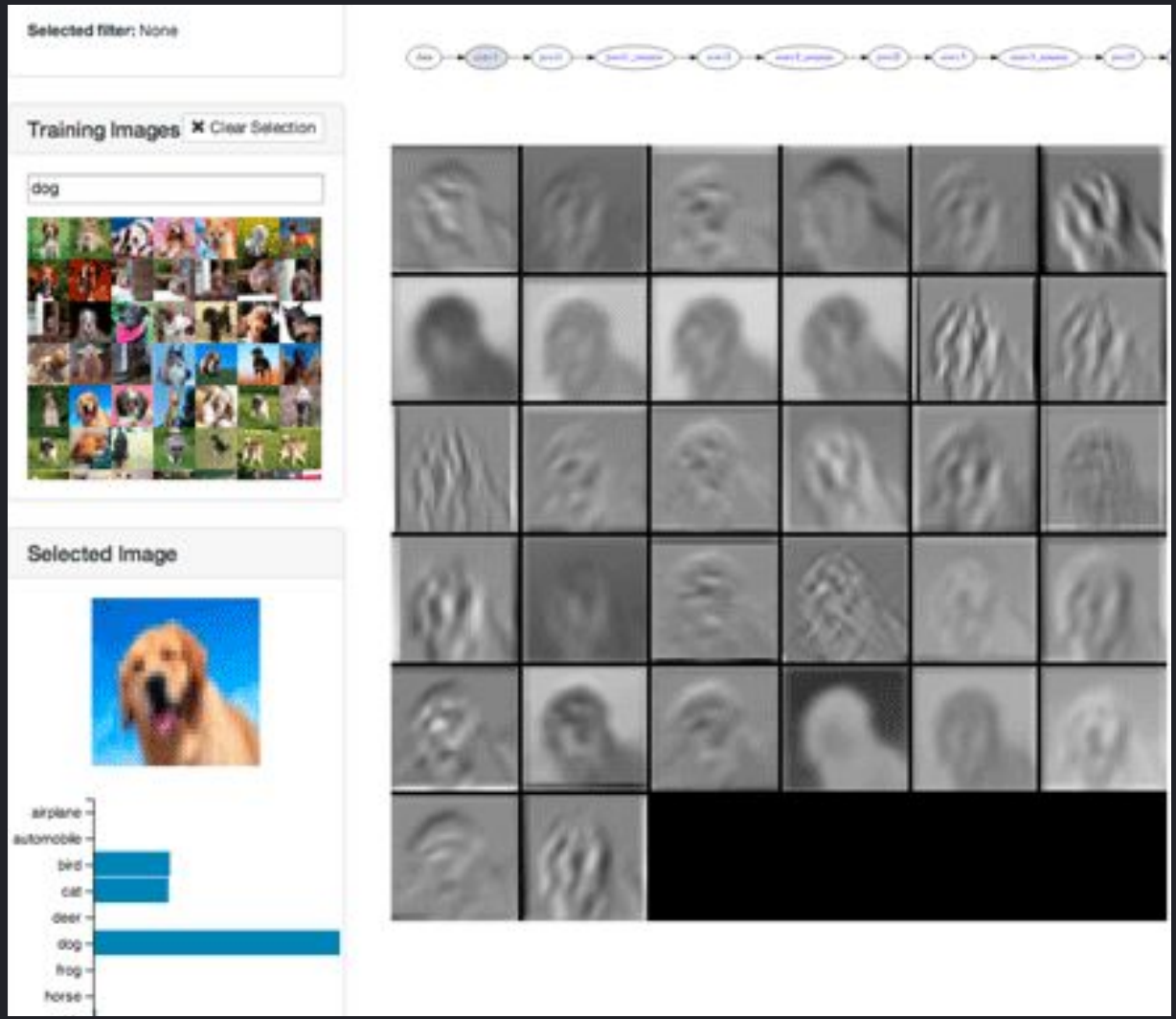
DEBUGGING

Dataset issues?



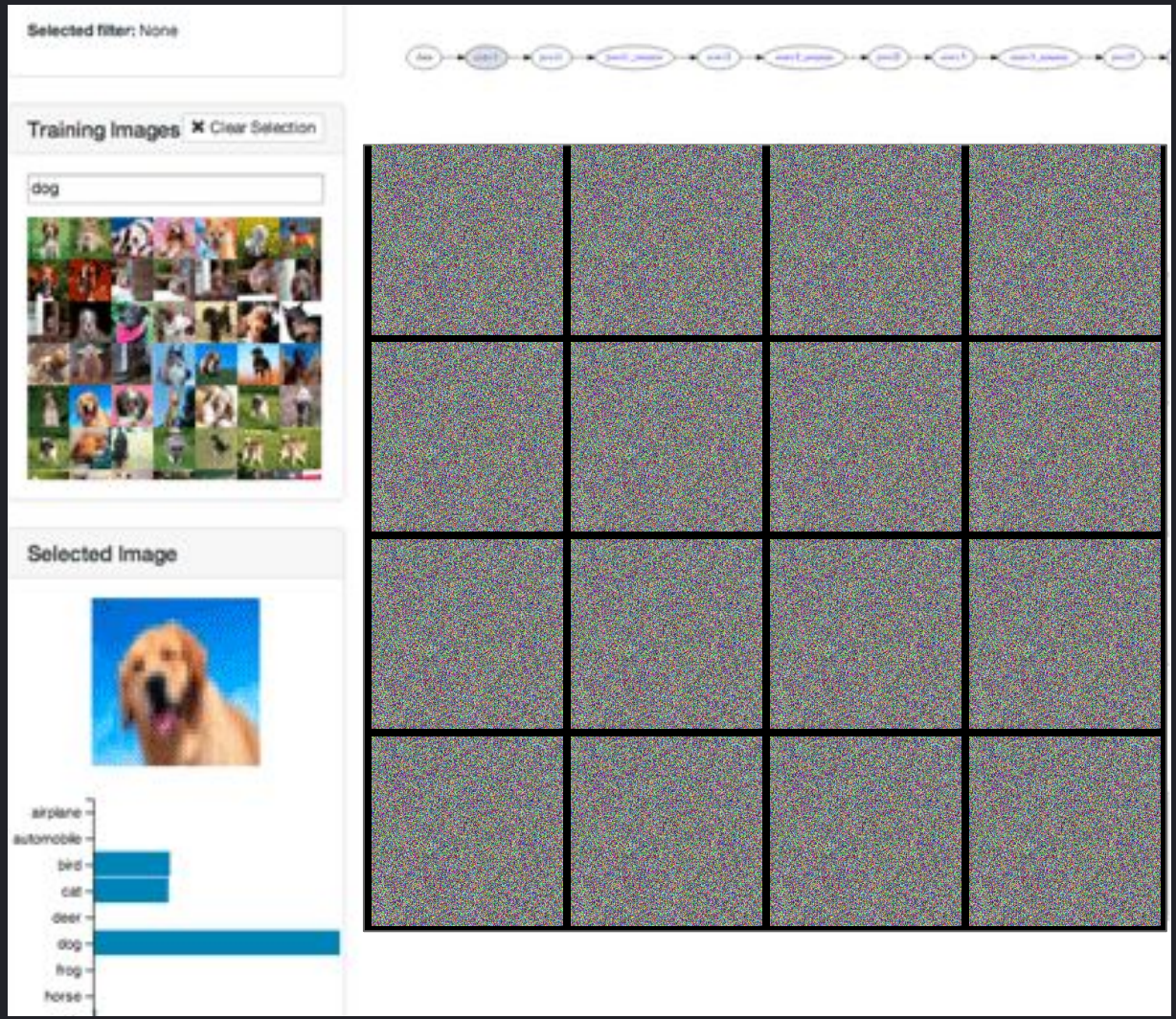


github.com/ividal/



https://github.com/bruckner/deepViz
http://yosinski.com/deepvis

https://github.com/bruckner/deepViz
http://yosinski.com/deepvis



INTERPRETABILITY

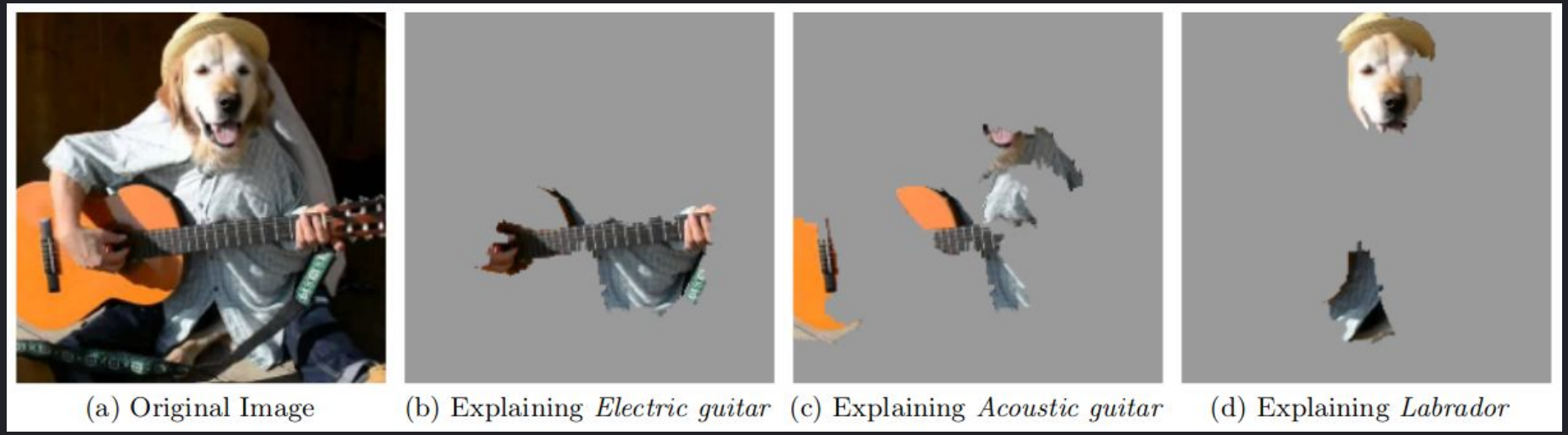
WHY CARE?

To…

- Debug
- Explain
- Trust
- Generalize
- Optimize

To…

Avoid silent failure



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

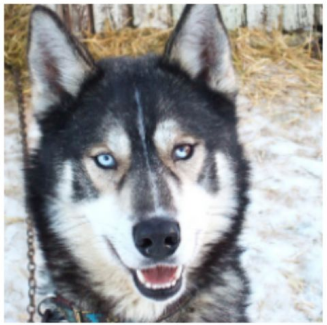
"Why Should I Trust You?"Explaining the Predictions of Any Classifier", Ribeiro, M.T. et al.

(a) Original Image

(b) Explaining *Electric guitar*

(c) Explaining *Acoustic guitar*

(d) Explaining *Labrador*



(a) Husky classified as wolf

(b) Explanation

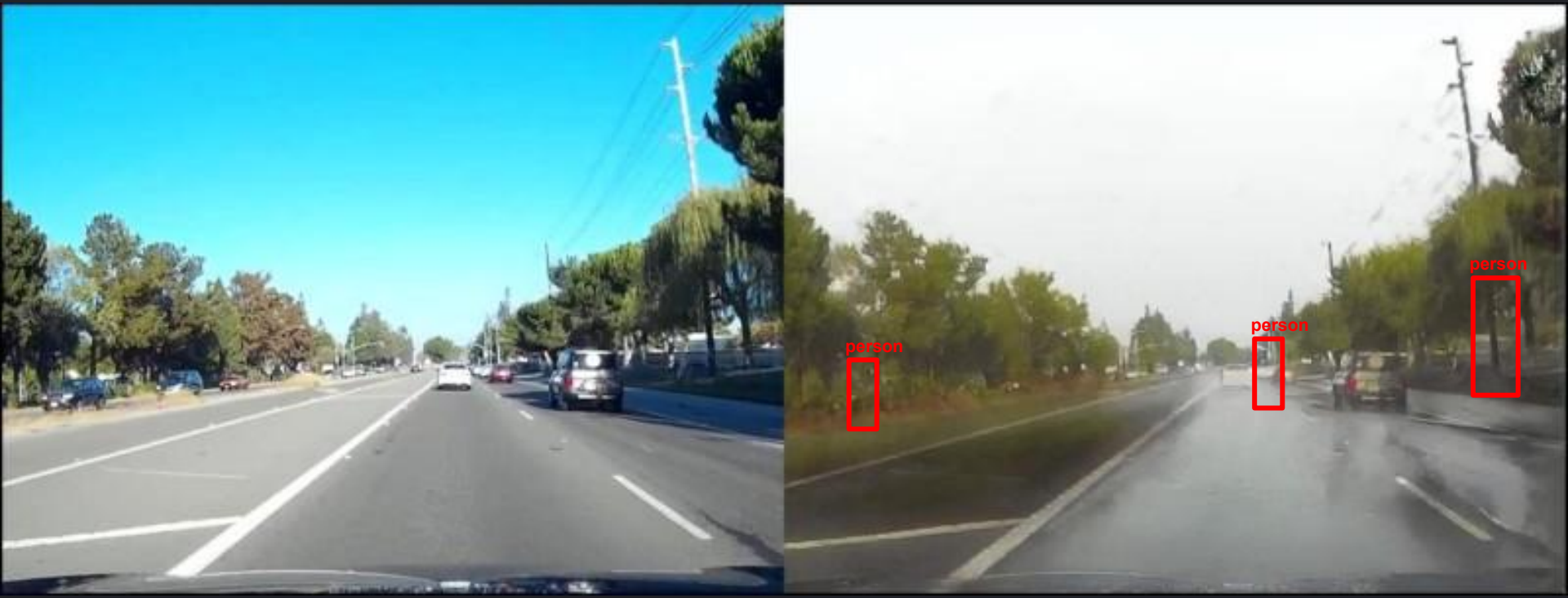
"Why Should I Trust You?"Explaining the Predictions of Any Classifier", Ribeiro, M.T. et al.

ADVERSARIAL *

ADVERSARIAL SAMPLES 101

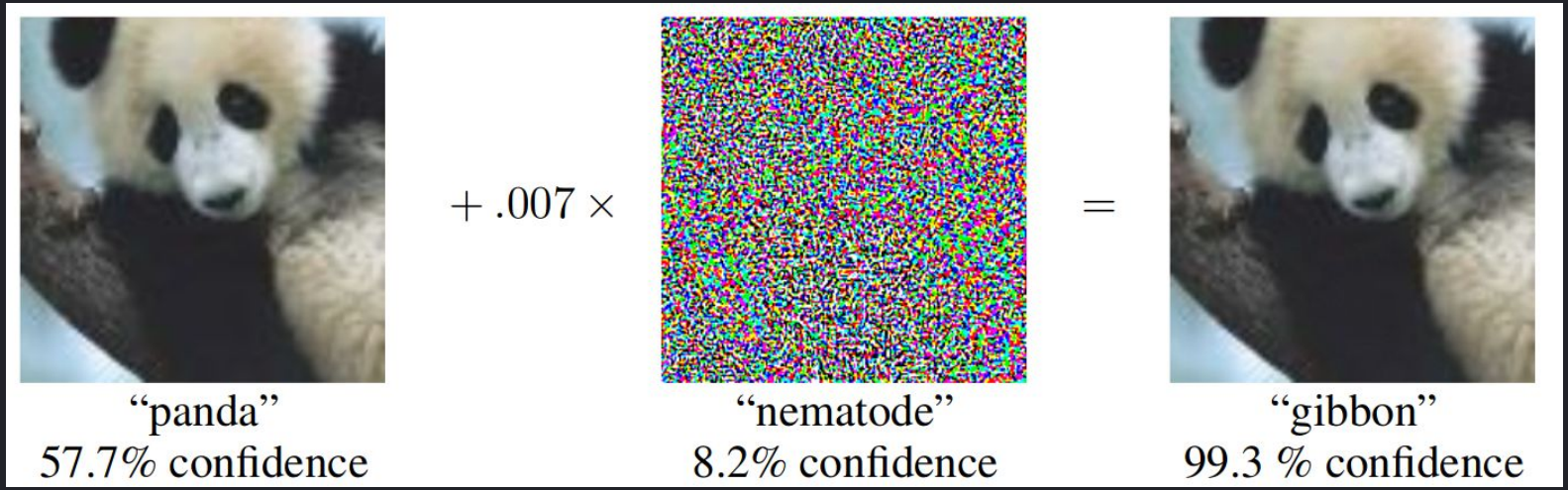
Collect your failures!

ADVERSARIAL SAMPLES 101



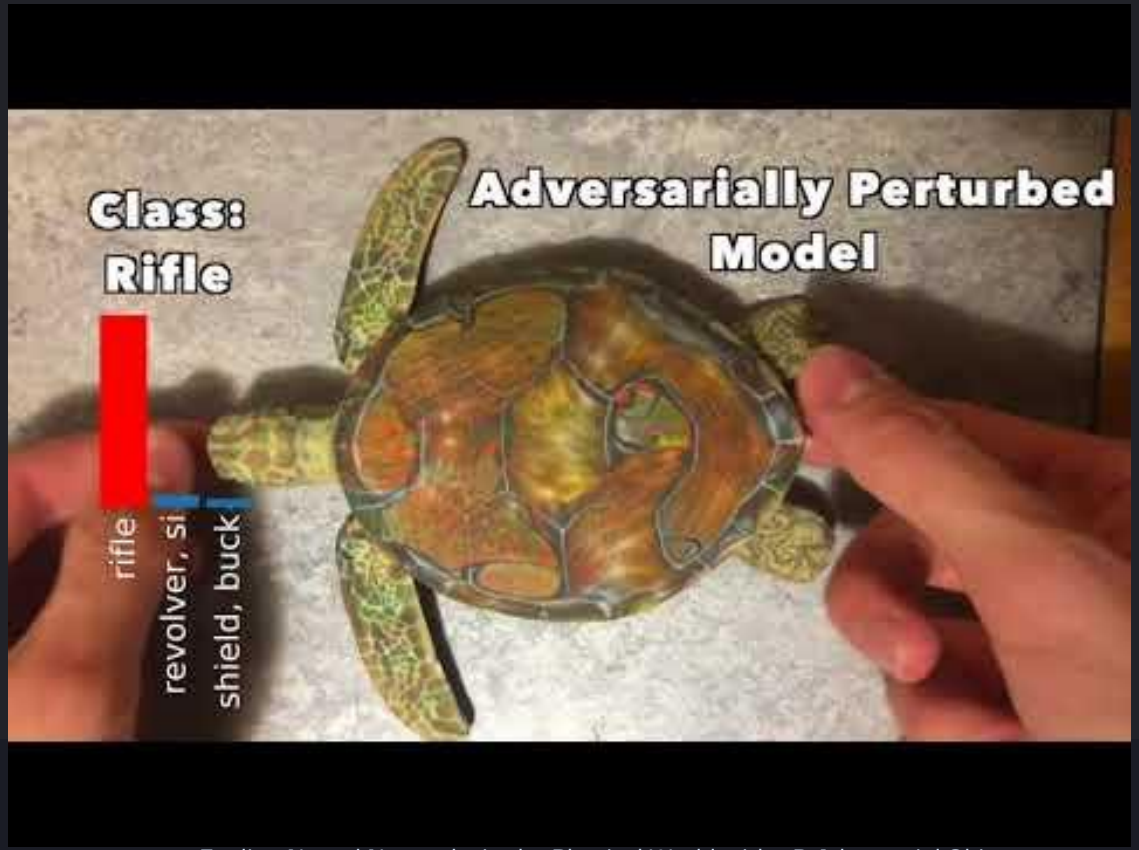
Background image: NVIDIA

ADVERSARIAL SAMPLES



“Explaining and Harnessing Adversarial Examples”, Goodfellow et al. 2014

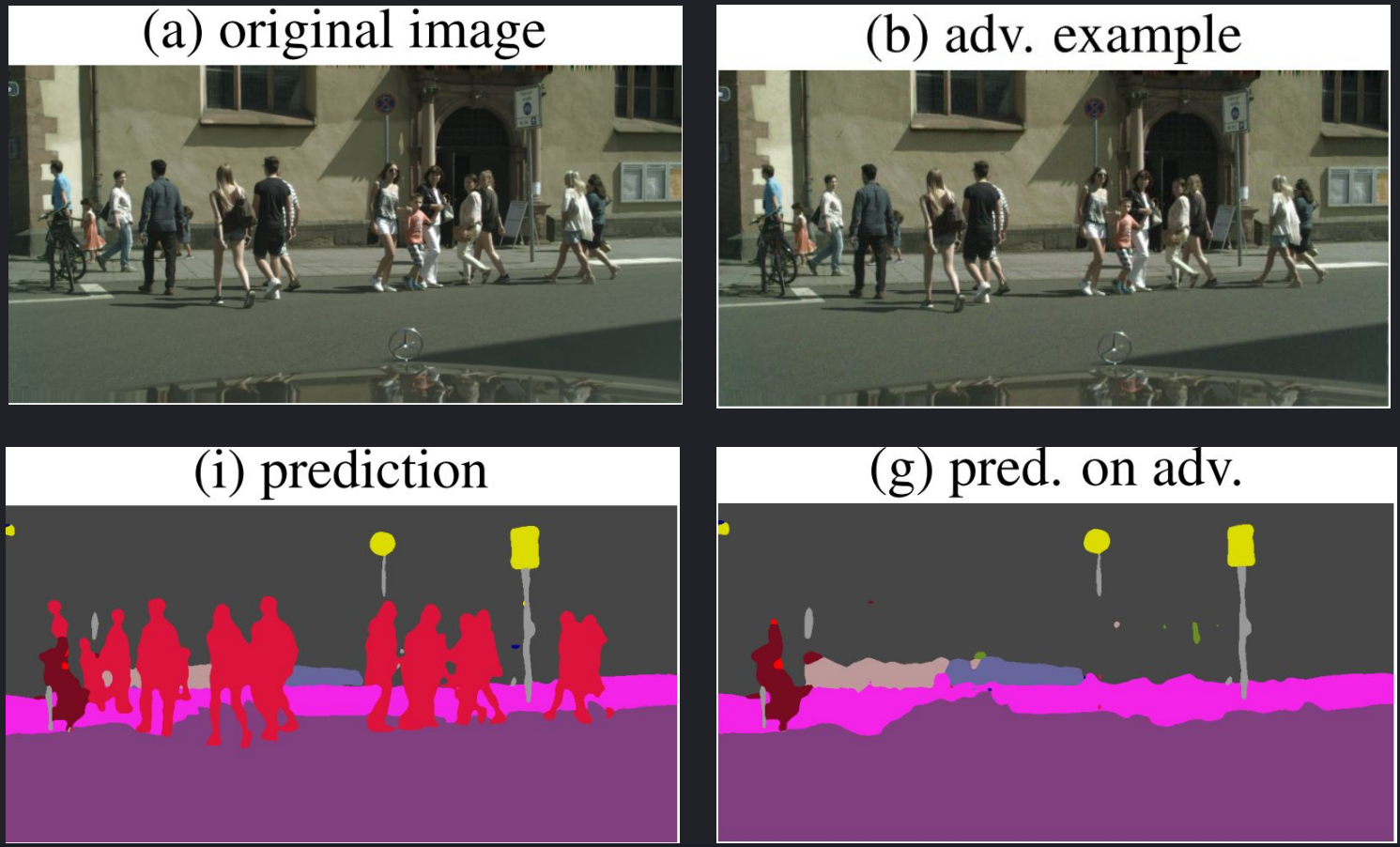
ADVERSARIAL SAMPLES



Fooling Neural Networks in the Physical World with 3D Adversarial Objects

ADVERSARIAL SAMPLES

"Robust Physical-World Attacks on Deep Learning Visual Classification", Evtimov et al. (CVPR 2018)



Fischer V. et al, *Adversarial Examples for Semantic Image Segmentation*, ICLR 2017

RECAP

ROBUSTNESS IS MANY THINGS

Evaluation

Debugging

Interpretability

Adversarial samples

ROBUSTNESS IS MANY THINGS

Evaluation

Debugging

Interpretability

Adversarial samples

Which ones are you already using?

Thank you!

linkedin.com/in/irinavidal/

Thank you!

linkedin.com/in/irinavidal/

Special thanks to the people behind the presentation template: SlidesCarnival