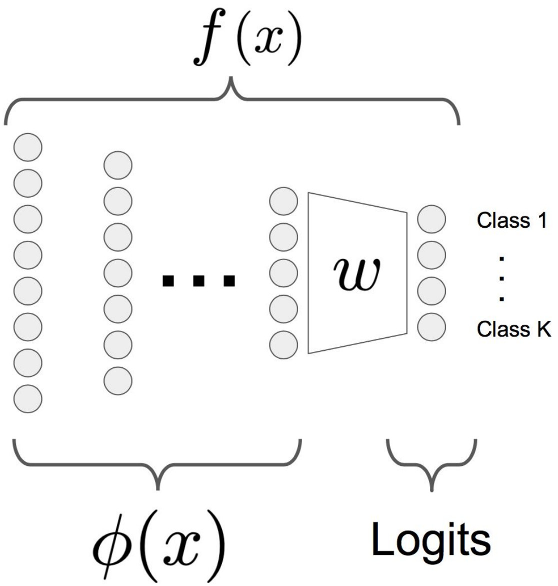
The Odds are Odd:
A Statistical Test for Detecting Adversarial Examples

Kevin Roth*, **Yannic Kilcher***, Thomas Hofmann

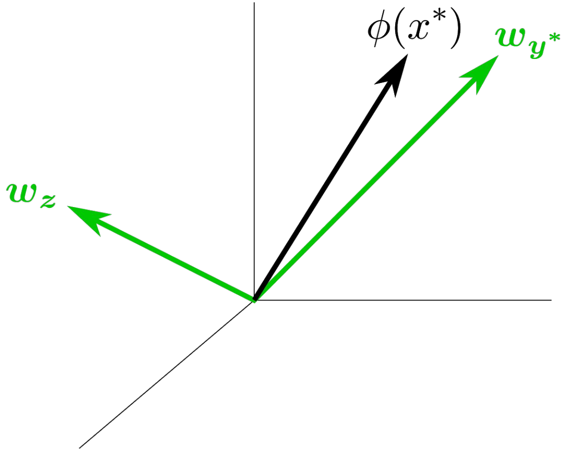
ETH Zürich

poster #62

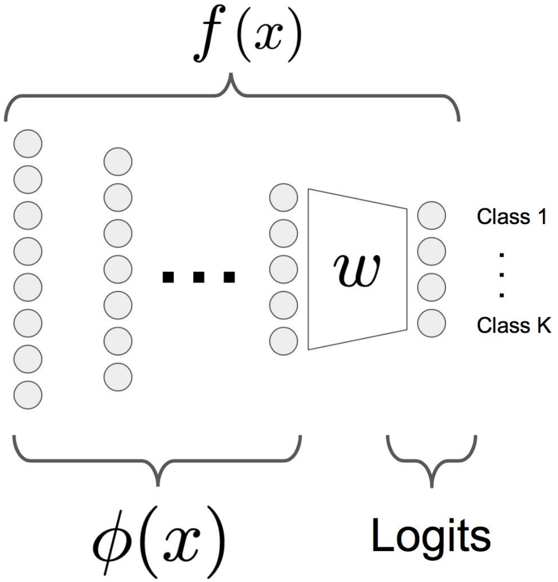
Log-Odds & Adversarial Examples



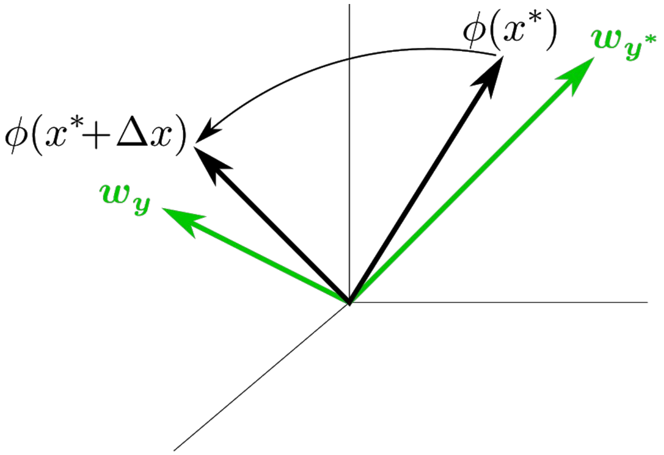
$$f_y(x) = \langle w_y, \phi(x) \rangle$$



Log-Odds & Adversarial Examples



$$f_y(x) = \langle w_y, \phi(x) \rangle$$

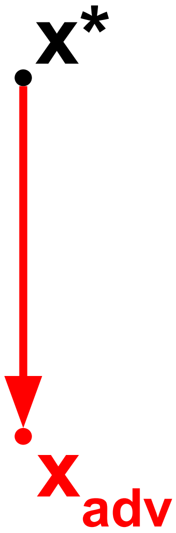


Adversarial examples cause atypically large feature space **perturbations along** the **weight-difference** direction

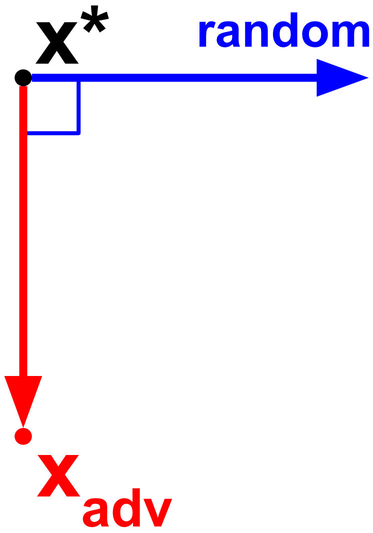
Adversarial Cone

$\bullet\, \mathbf{x}^*$

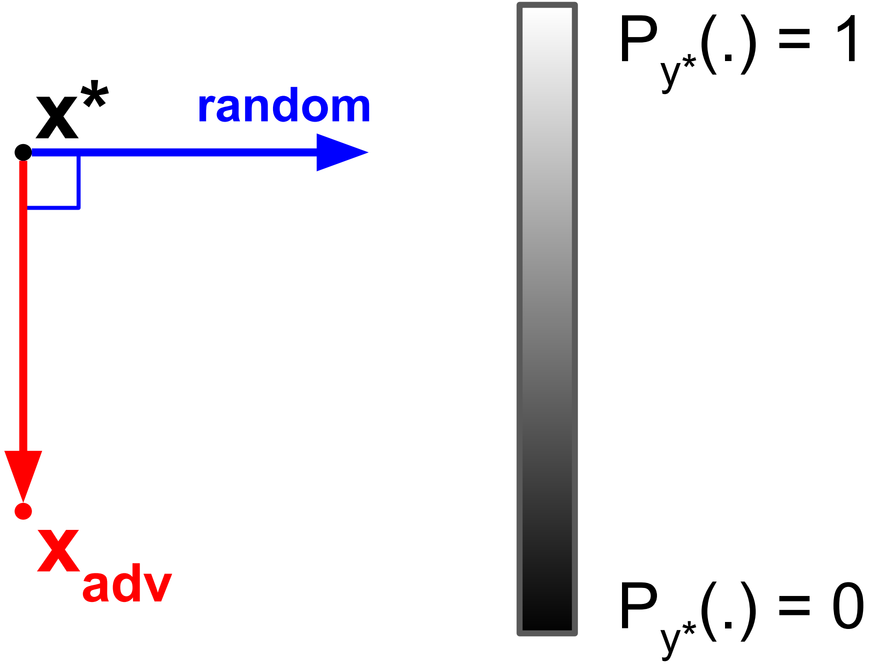
Adversarial Cone



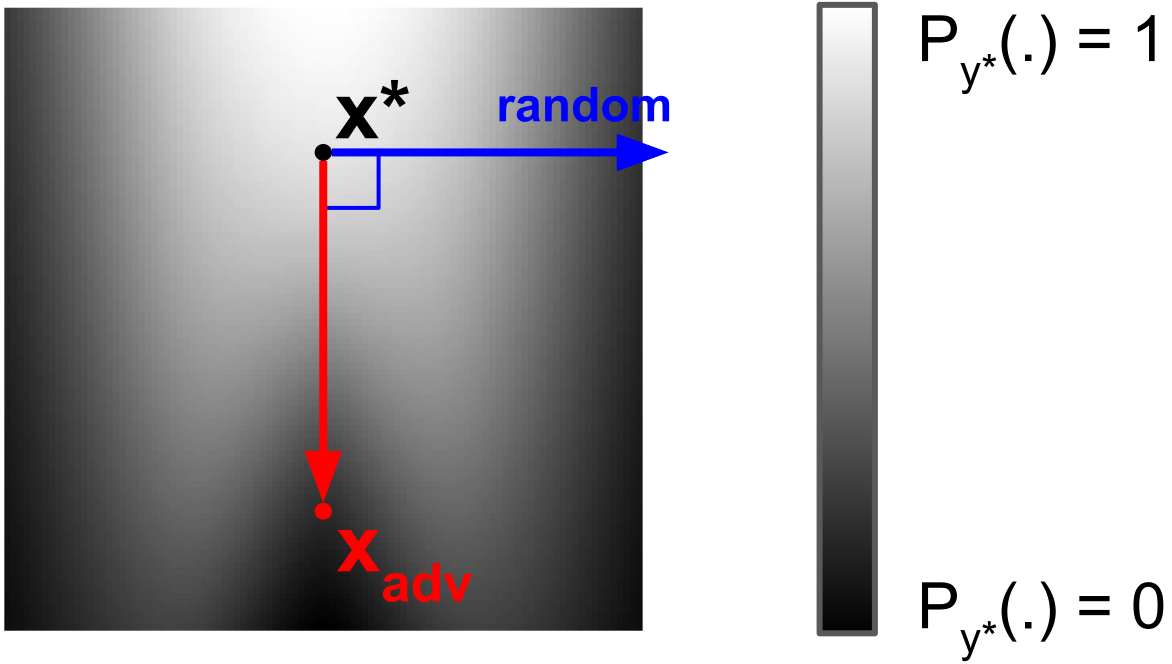
Adversarial Cone



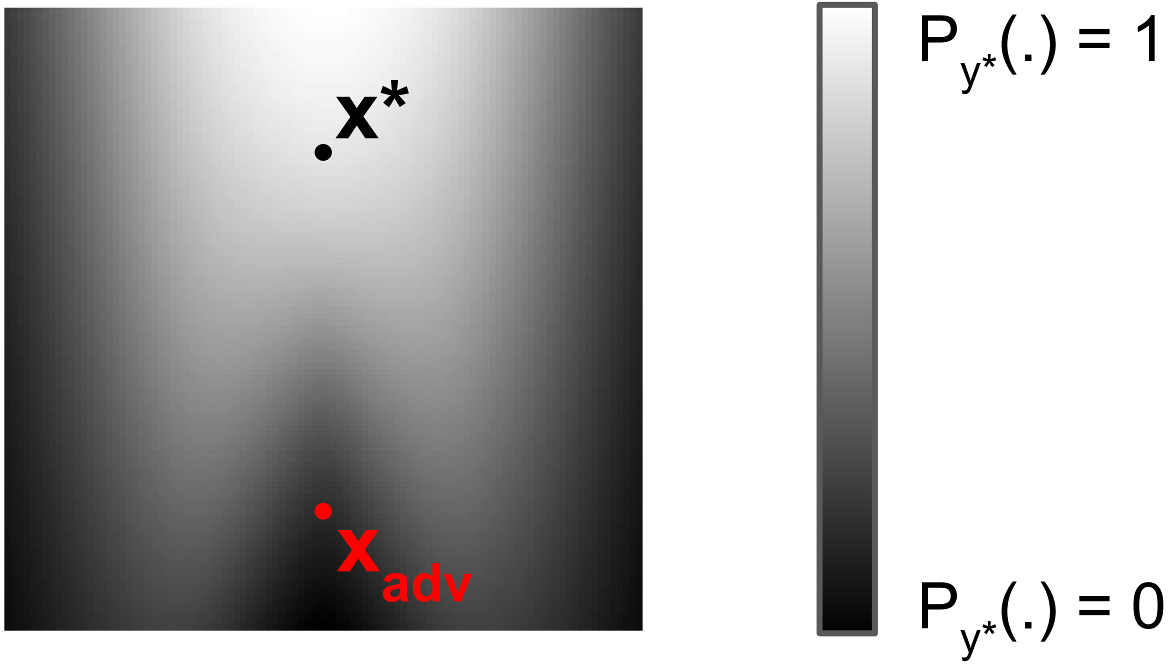
Adversarial Cone



Adversarial Cone

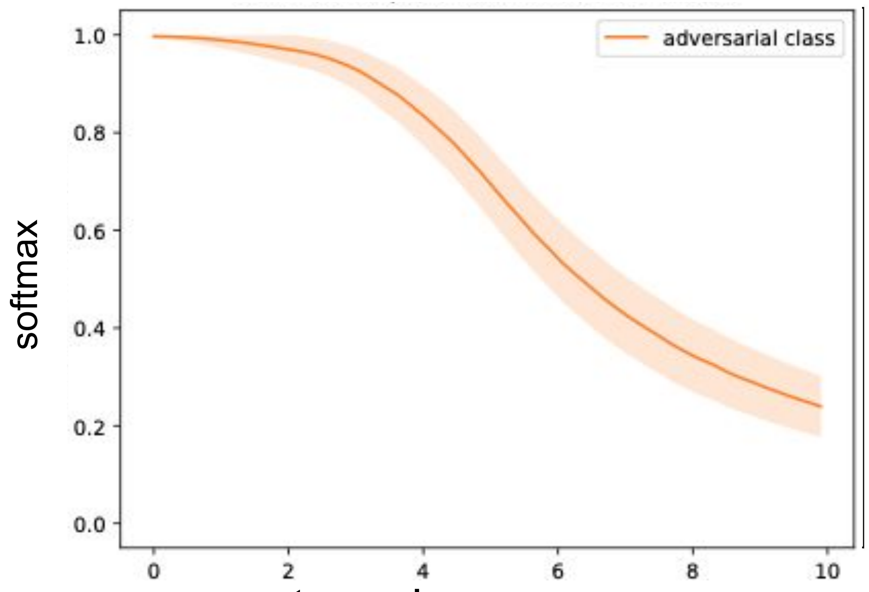


Adversarial Cone



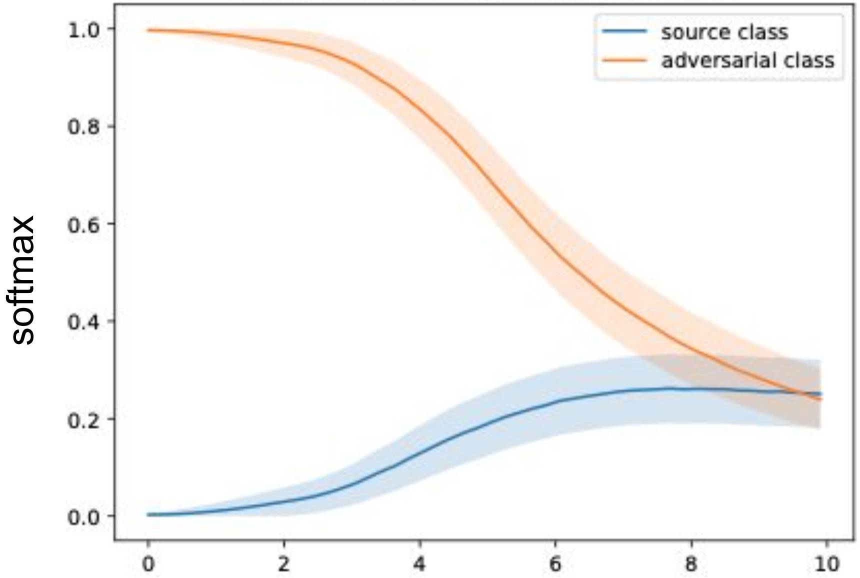
Adversarial examples are embedded in a cone-like structure

Adversarial Cone



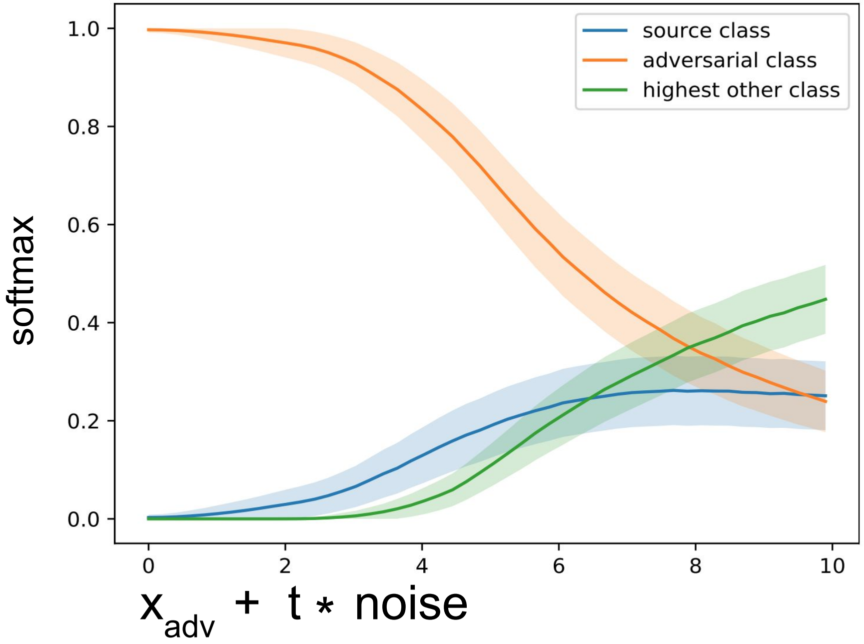
x_{adv} + t * noise

Adversarial Cone



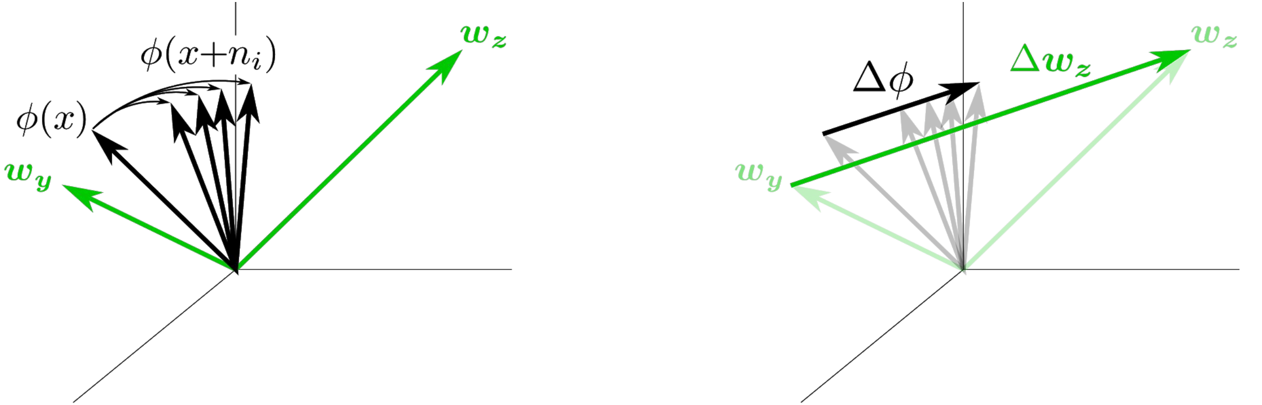
$x_{adv} + t * \text{noise}$

Adversarial Cone



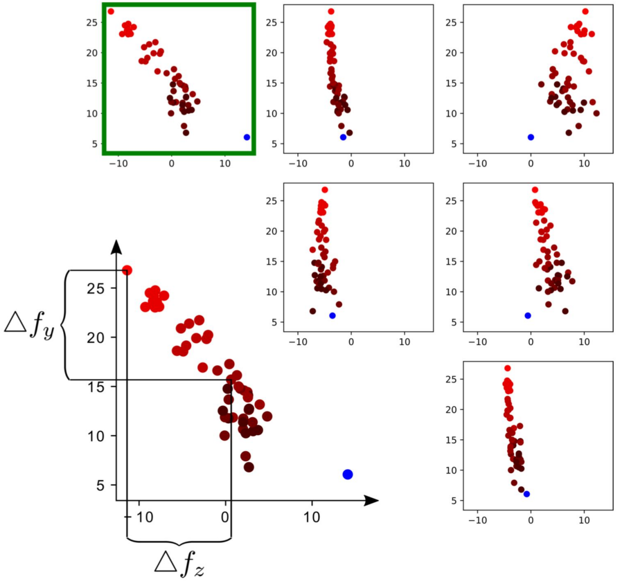
Noise as a *probing instrument*

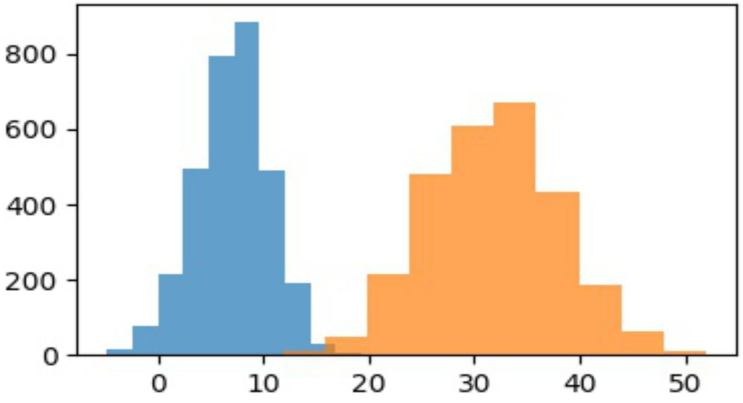
Main Idea: Log-Odds Robustness



- The robustness properties of $\phi(x+n_i)$ are different dependent on whether $x = x^*$ or $x = x^* + \Delta x$
- $\Delta\phi$ tends to have a **characteristic direction** if $x = x^* + \Delta x$ whereas it tends not to have a specific direction if $x = x^*$

Main Idea: Log-Odds Robustness





natural adversarial

Noise can partially undo effect of adversarial perturbation and directionally revert log-odds towards the true class y*

Statistical Test & Corrected Classification

- We propose to use **noise-perturbed pairwise log-odds**

$$g_{y,z}(x,\eta) = \langle w_z - w_y, \phi(x+\eta) - \phi(x) \rangle$$

to test whether x **classified as** y **should be thought of as a manipulated example of true class** z **:**

$$x \textbf{ adversarial if } \quad \max_{z \neq y} \left\{ \mathbf{E}_\eta \left[\bar{g}_{y,z}(x,\eta) \right] - \tau_{y,z} \right\} \geq 0$$

- **Corrected classification :** $G(x) = \arg\max_z \left\{ \bar{g}_{y,z}(x) - \tau_{y,z} \right\}$

Detection Rates & Corrected Classification

Table 1: CIFAR10

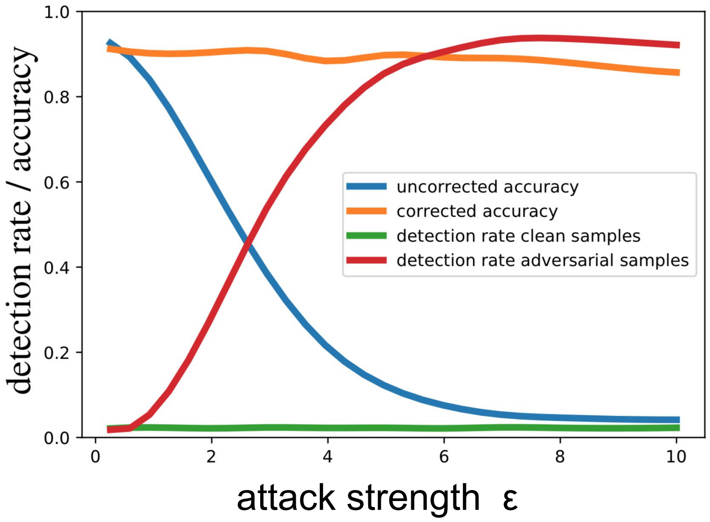
Model	Detection rate (clean / pgd)	Corrected Accuracy (clean / pgd)
WResNet	**0.2% / 99.1%**	**96.0% / 92.7%**
CNN7	**0.8% / 95.0%**	**93.6% / 89.5%**
CNN4	**1.4% / 93.8%**	**71.0% / 67.6%**

Table 2: ImageNet

Model	Detection rate (clean / pgd)
Inception V3	**1.9% / 99.6%**
ResNet 101	**0.8% / 99.8%**
VGG16(+BN)	**0.3% / 99.9%**

- Our statistical test **detects nearly all adversarial examples** with FPR ~1%
- Our correction method **reclassifies almost all adversarial examples** successfully
- **Drop in performance on clean samples is negligible**

Detection Rates & Corrected Classification



- **Detection rate increases with increasing attack strength**
- **Corrected classification manages to compensate for decay in uncorrected accuracy** due to increase in attack strength

Defending against Defense-Aware Attacks

Model	Detection rate (clean / attack)	Accuracy (clean / attack)
WResNet	**4.5% / 71.4%**	**91.7% / 56.0%**
CNN7	**2.8% / 75.5%**	**91.2% / 56.6%**
CNN4	**4.1% / 81.3%**	**69.0% / 56.5%**

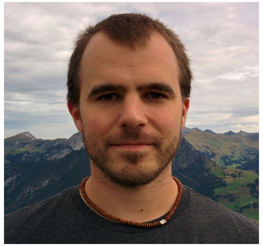
Attacker has **full knowledge of the defense** :

perturbations that work in expectation under noise source used for detection

Detection rates and corrected accuracies remain remarkably high

poster #62

Thank You



Kevin Roth **Yannic Kilcher** **Thomas Hofmann**

data analytics lab ETH zürich

Follow-Up Work: *Adversarial Training Generalizes Data-dependent Spectral Norm Regularization*

ICML Workshop on Generalization (June 14)

References

The approaches most related to our work are those that detect whether or not the input has been perturbed, either by detecting characteristic regularities in the adversarial perturbations themselves or in the network activations they induce.

- Grosse, Kathrin, et al. "On the (statistical) detection of adversarial examples." (2017).
- Metzen, Jan Hendrik, et al. "On detecting adversarial perturbations." (2017).
- Feinman, Reuben, et al. "Detecting adversarial samples from artifacts." (2017).
- Xu, Weilin, David Evans, and Yanjun Qi. "Feature squeezing: Detecting adversarial examples in deep neural networks." (2017).
- Song, Yang, et al. "Pixeldefend: Leveraging generative models to understand and defend against adversarial examples." (2017).
- Carlini, Nicholas, and David Wagner. "Adversarial examples are not easily detected: Bypassing ten detection methods." (2017).
- … and many more