Sapienza Università di Roma

On the Sensitivity and Uncertainty of Convolution Neural Networks to Adversarial Perturbations

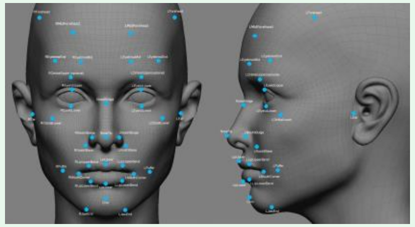
Senad Beadini

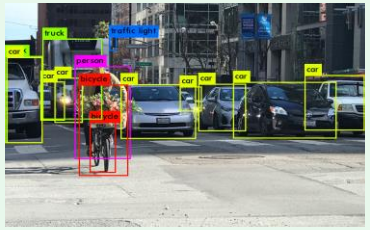
Iacopo Masi
Supervisor

Gabriele Tolomei
Co-supervisor

Deep Convolutional Networks for Classification

State-of-the-art models get **remarkable results in complex cognitive tasks**.

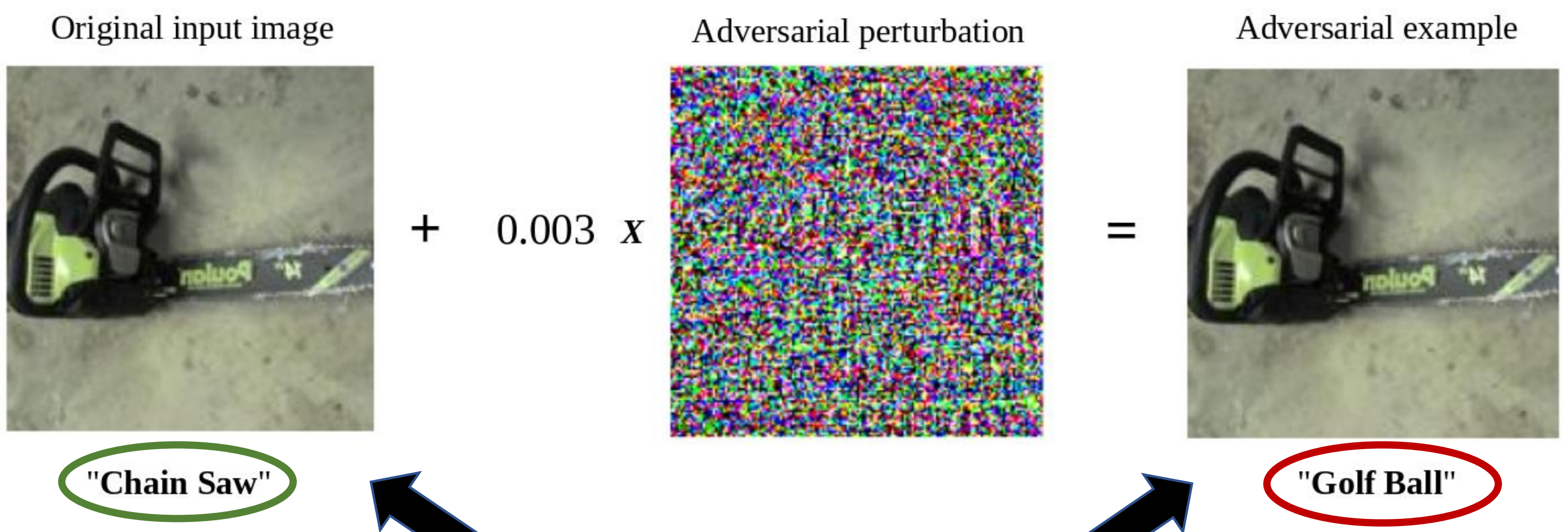




On ImageNet, convolutional neural networks (CNNs) **achieve accuracy on par with humans.**

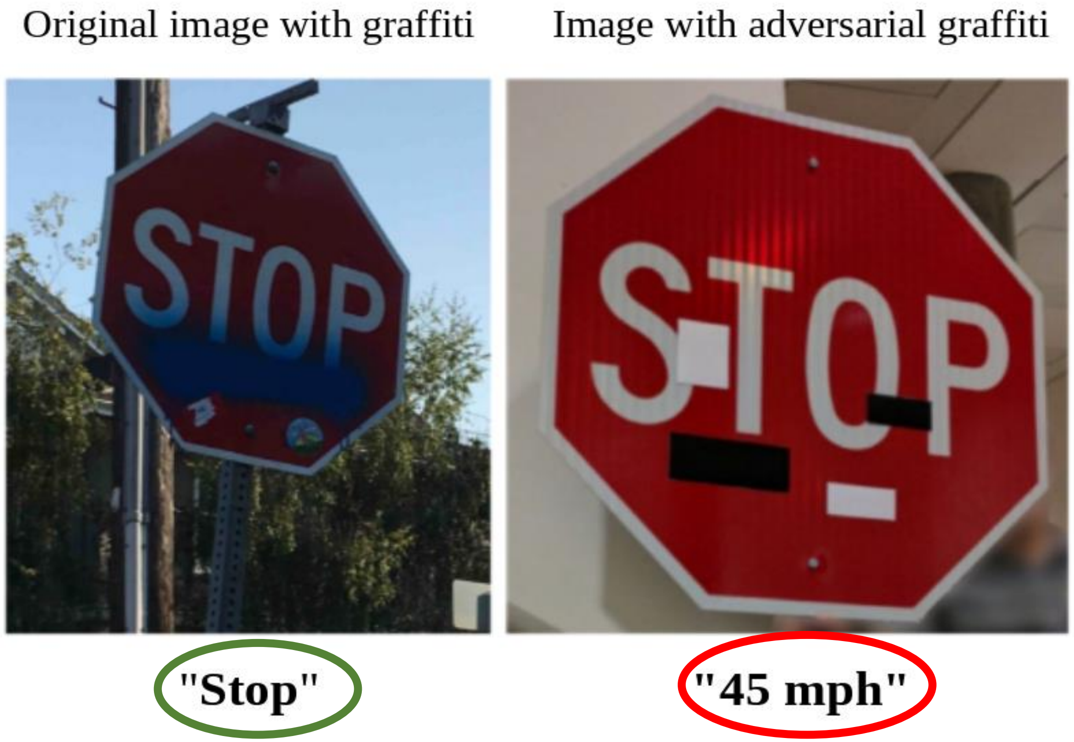
Adversarial Perturbations Make CNNs non-Robust

An imperceptible perturbation could break the performance of any model.



For human perception no difference.

Why Is Important to Study Adversarial Perturbations?

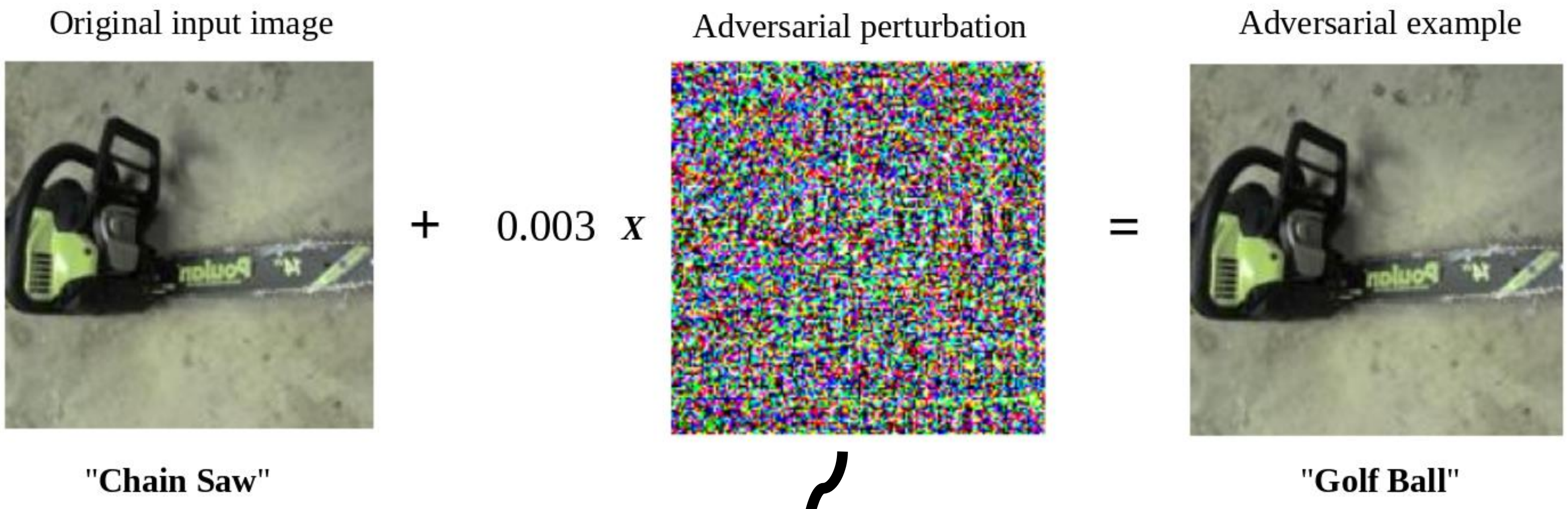


[Eykholt et al. CVPR 2018]

What kinds of perturbations exist?

Pixel-wise Perturbations :

The perturbation in the pixel space is limited by a threshold ε under a norm p.



The CNN achieves 85% accuracy without any perturbation.

The accuracy of the CNN under attack is 0%.

Transformation-based Adversarial Example :

We apply a transformation to the input image to get a new image. [Engstrom et al. ICML 2018]

Original input image

$\boldsymbol{x}$

$$\boldsymbol{x}' = T(\boldsymbol{x}, \gamma)$$

T

T parametrized by γ

Transformed image

$\boldsymbol{x}'$

The perturbation is created in the parameter space of T

Contribution and Related Work

1) **Kamath et al. NeurIPS 2021** proved that exists a trade-off between pixel-wise robustness and Rotation/Translation robustness.

- **We replicate the results of this work by showing the trade-off for rotation and extend the analysis to the perspective and other transformations.**

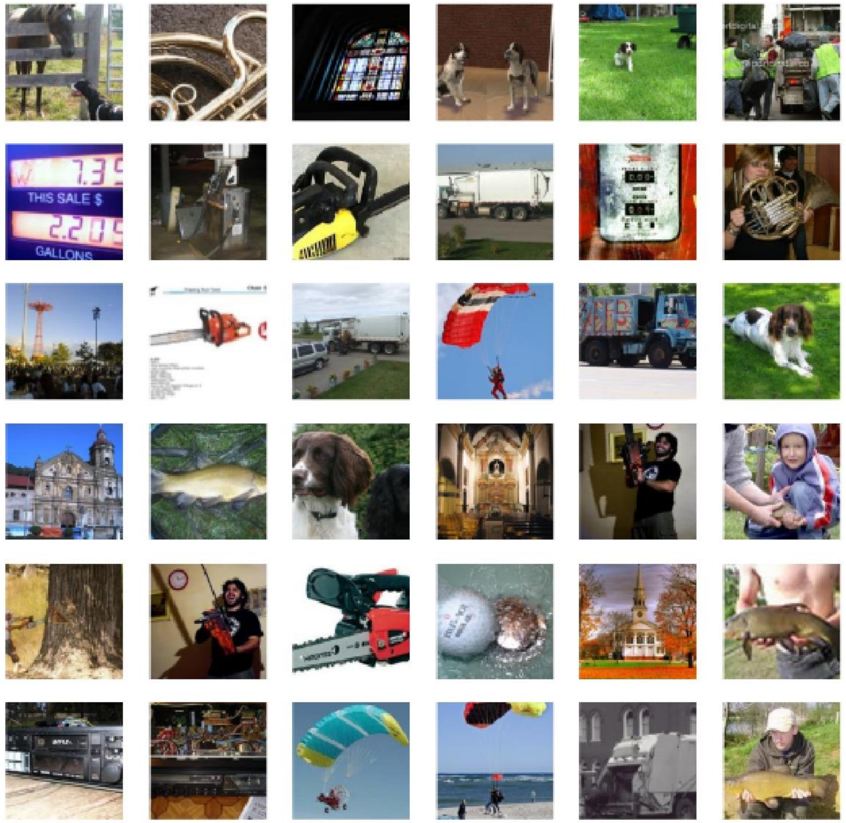
2) **Grathwohl et al. ICLR 2020** and **Zhu et al. ICCV 2021** find that CNNs can be reinterpreted as an energy-based model.

- **Exploiting these results, we show that there’s a correlation between the energy of a discriminative model and the strength of a pixel-wise adversarial attack.**
- **We provide a detector based on the energy function to detect adversarial attacks.**
- **We propose an algorithm for generating low-energy adversarial data.**

Dataset and Models

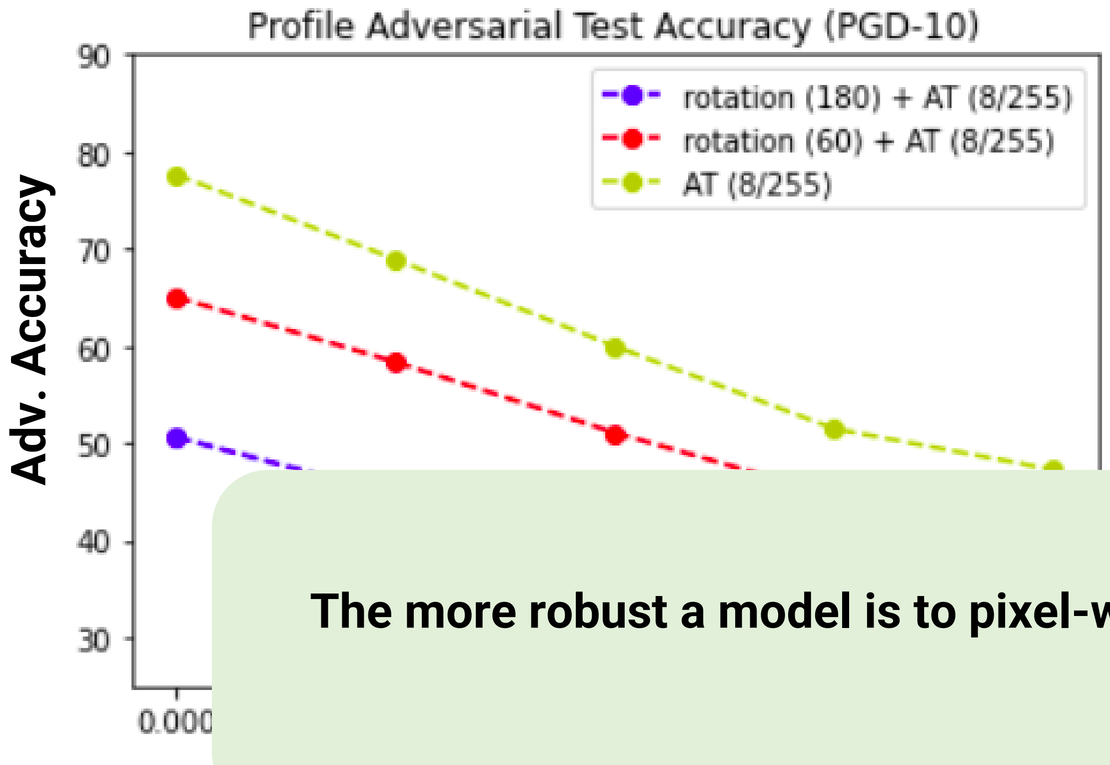
The datasets used are CIFAR-10 and Imagenette, a subset of ImageNet.

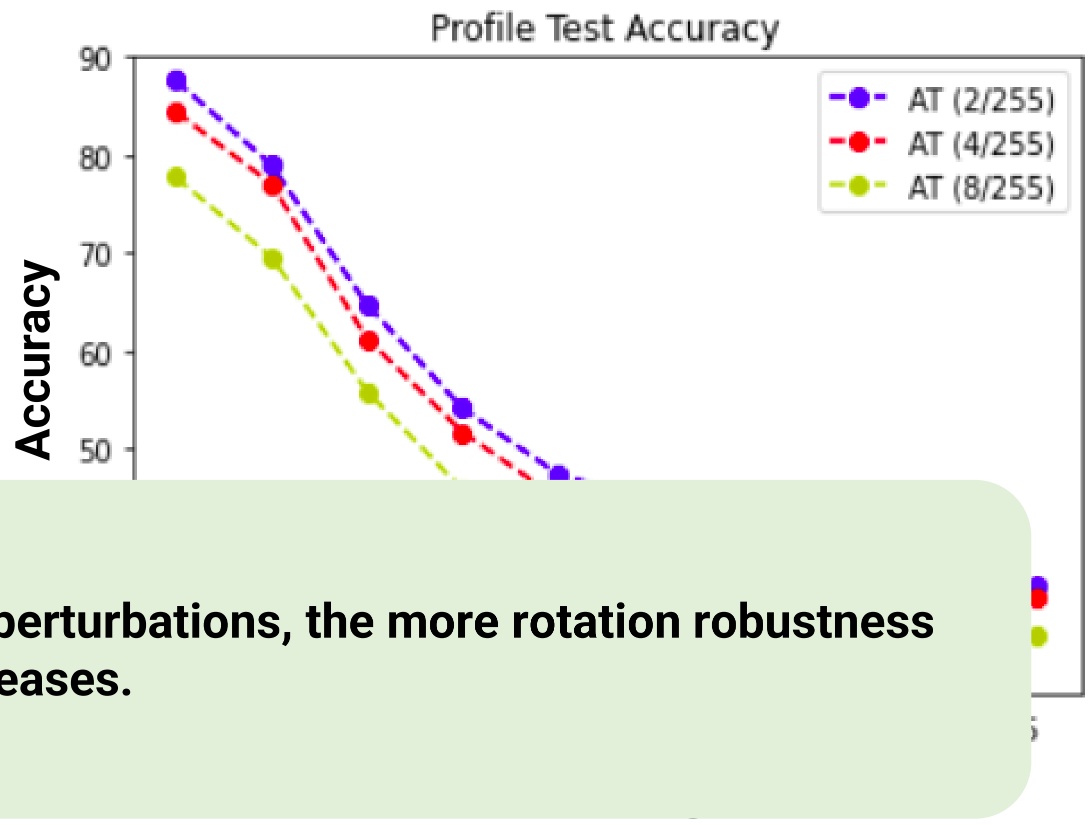
All the following experiments have been performed using a CNN with residual connections (i.e., ResNet10) .



Trade-off Analysis for Different Transformations

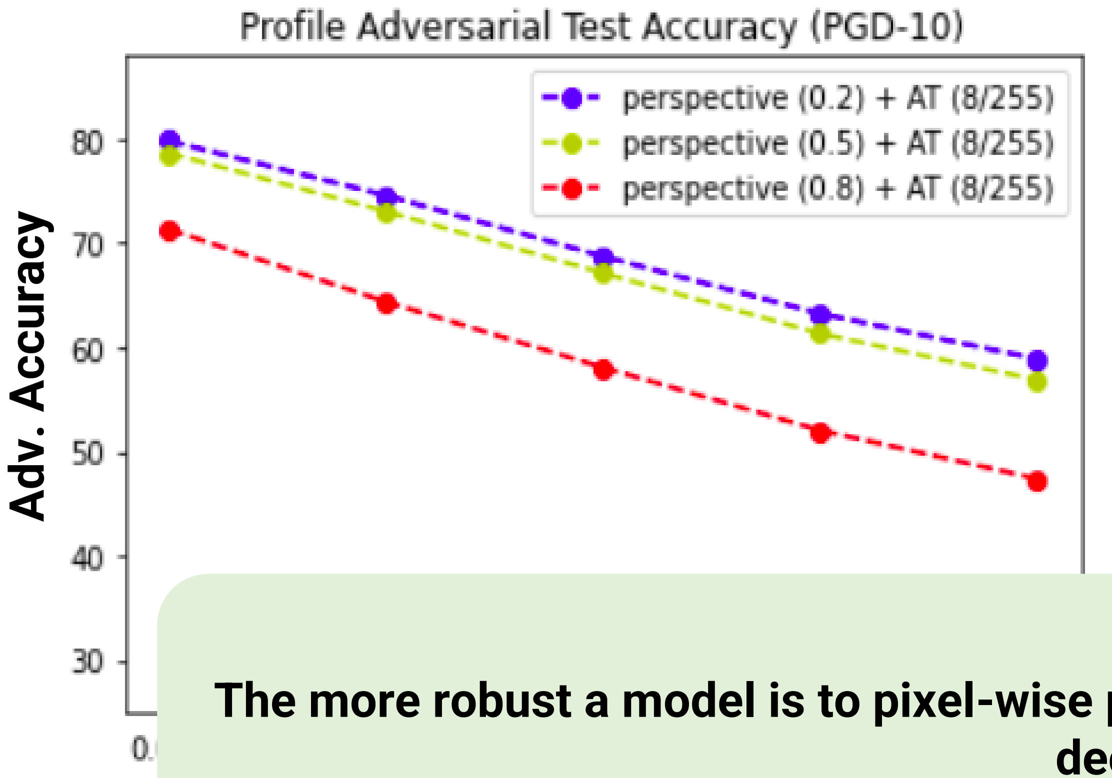
The basic defense for robustness consists of generating at training time perturbations and perform training on perturbed data.

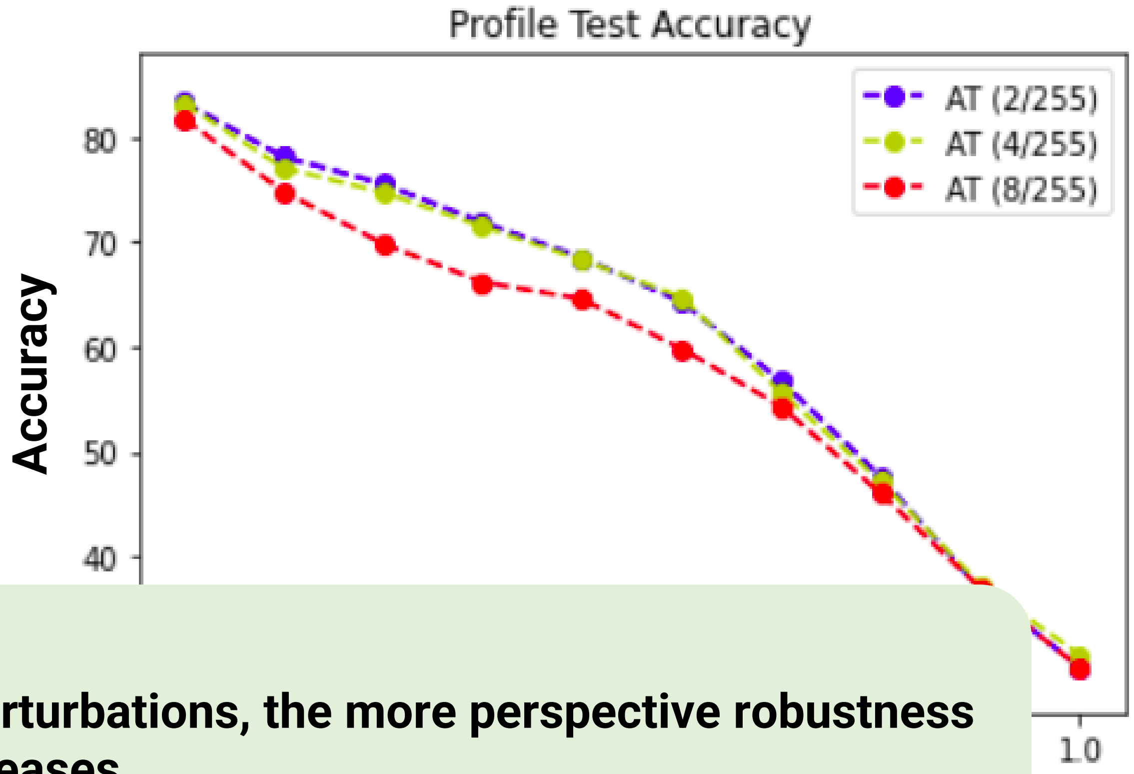




The more robust a model is to pixel-wise perturbations, the more rotation robustness decreases.

Trade-off Analysis for Perspective Transformation and Others

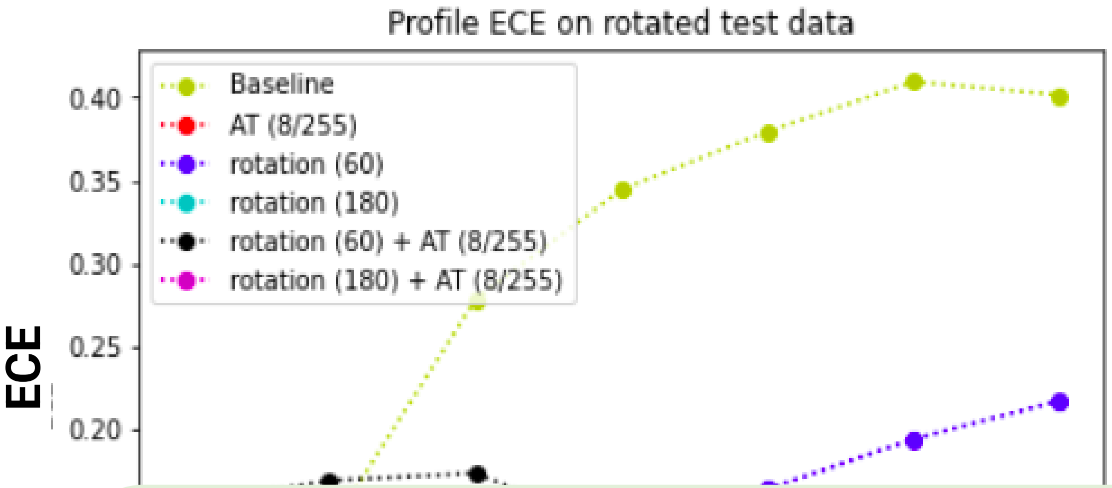


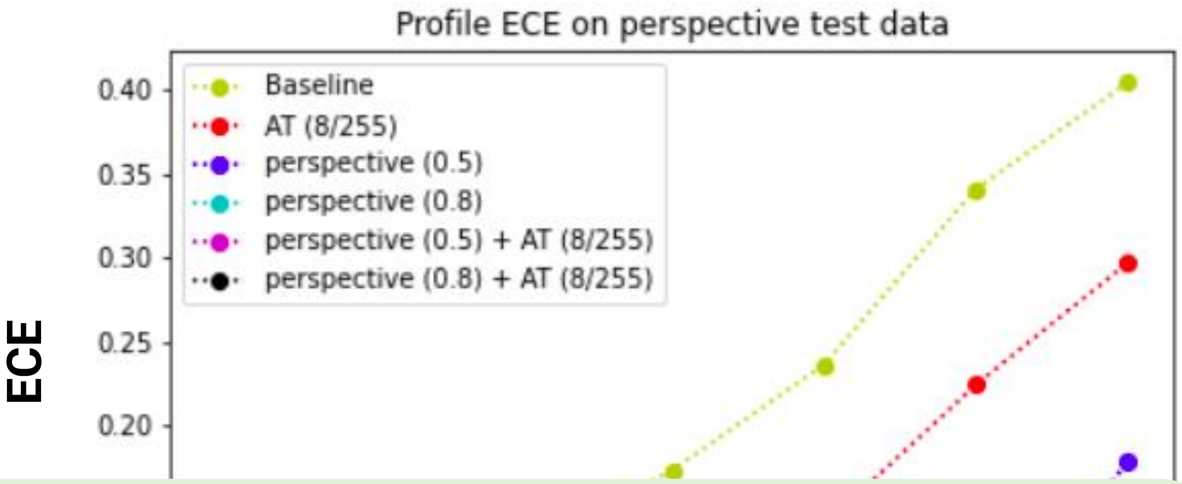


The more robust a model is to pixel-wise perturbations, the more perspective robustness decreases.

Calibration Analysis for Different Transformation

We measure calibration using the Expected Calibration Error (ECE).





Pixel-wise robustness provides more calibrated CNNs.

Combining both robustnesses still provides more calibrated CNNs.
This holds for other transformations like Gaussian Noise and Motion Blur.

Hypothesis :

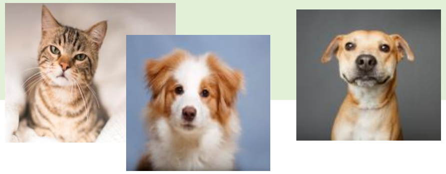
Pixel-wise adversarial examples are out-of-distribution (OOD) data.

Energy-based Model (EBM)

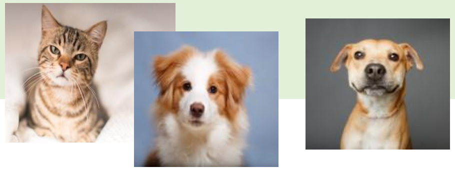
While classifiers are discriminative, EBMs are motivated from a different perspective: density estimation.

Density estimation : given a set of images that represents dogs and cats we want to learn a probability denstity function $p_\theta(x)$ over all possible images such that :

If $x =$  $\Rightarrow$ x **has high likelihood** $p_\theta(x)$



If $x =$  $\Rightarrow$ x **has low likelihood** $p_\theta(x)$



Energy-based Model (EBM)

While classifiers are discriminative, EBMs are motivated from a different perspective: density estimation.

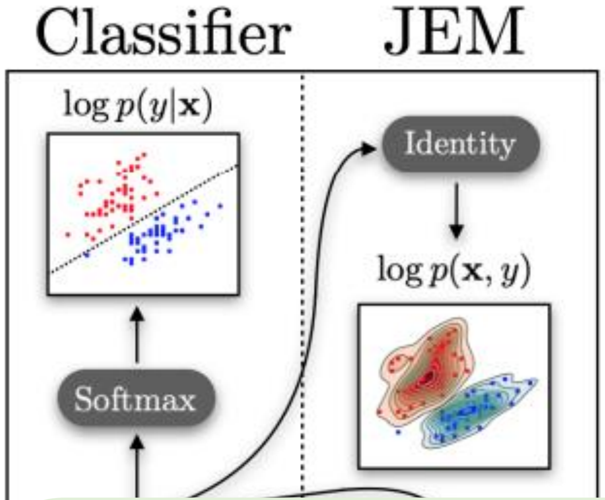
The fundamental idea of EBM is the energy function such that we approximate $p(x)$ via :

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\exp\left(-E_{\boldsymbol{\theta}}(\boldsymbol{x})\right)}{Z_{\boldsymbol{\theta}}}$$

Since $Z_{\boldsymbol{\theta}}$ is intractable for high dimensional data, EBM focuses directly on energy $E_{\boldsymbol{\theta}}(\boldsymbol{x})$ and not $p_{\boldsymbol{\theta}}(\boldsymbol{x})$.

In this framework, data points with high likelihood have a low energy, while data points with low likelihood have a high energy.

A CNN is Actually an EBM



A CNN F with softmax is a discriminative model :

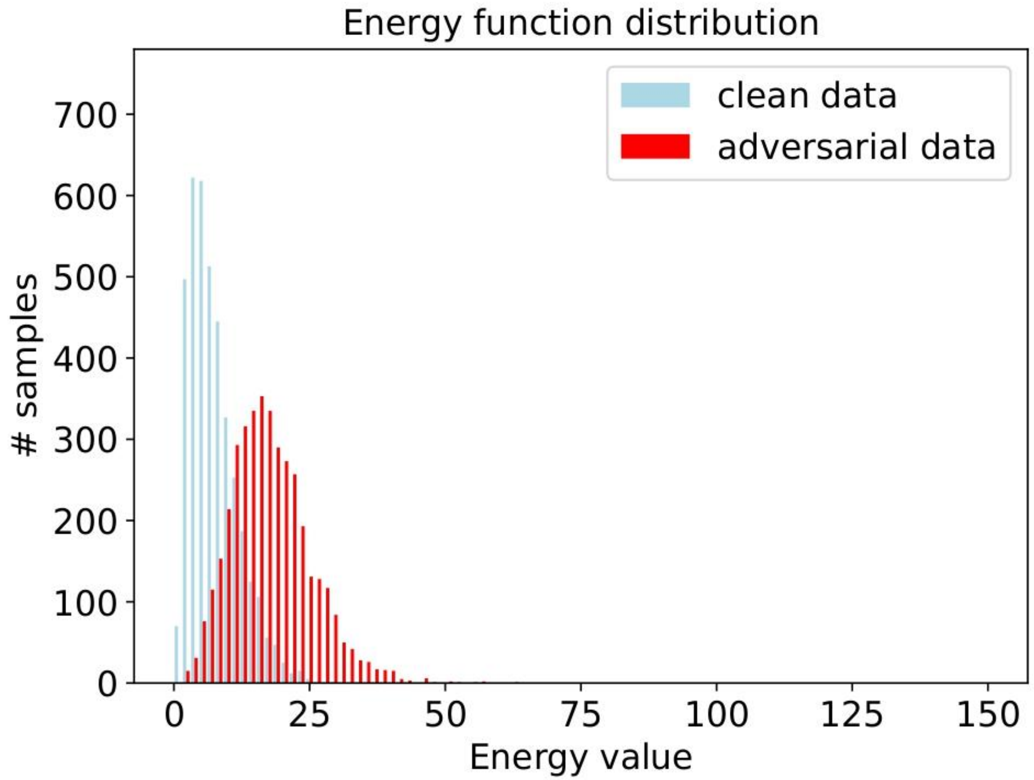
$$p(y = i|\boldsymbol{x}) = \frac{\exp F_{\boldsymbol{\theta}}(\boldsymbol{x})[i]}{\sum_{j=1}^{K} \exp F_{\boldsymbol{\theta}}(\boldsymbol{x})[j]}$$

Grathwohl et al. ICLR 2020 observed that we can reinterpret without any change F as an EBM :

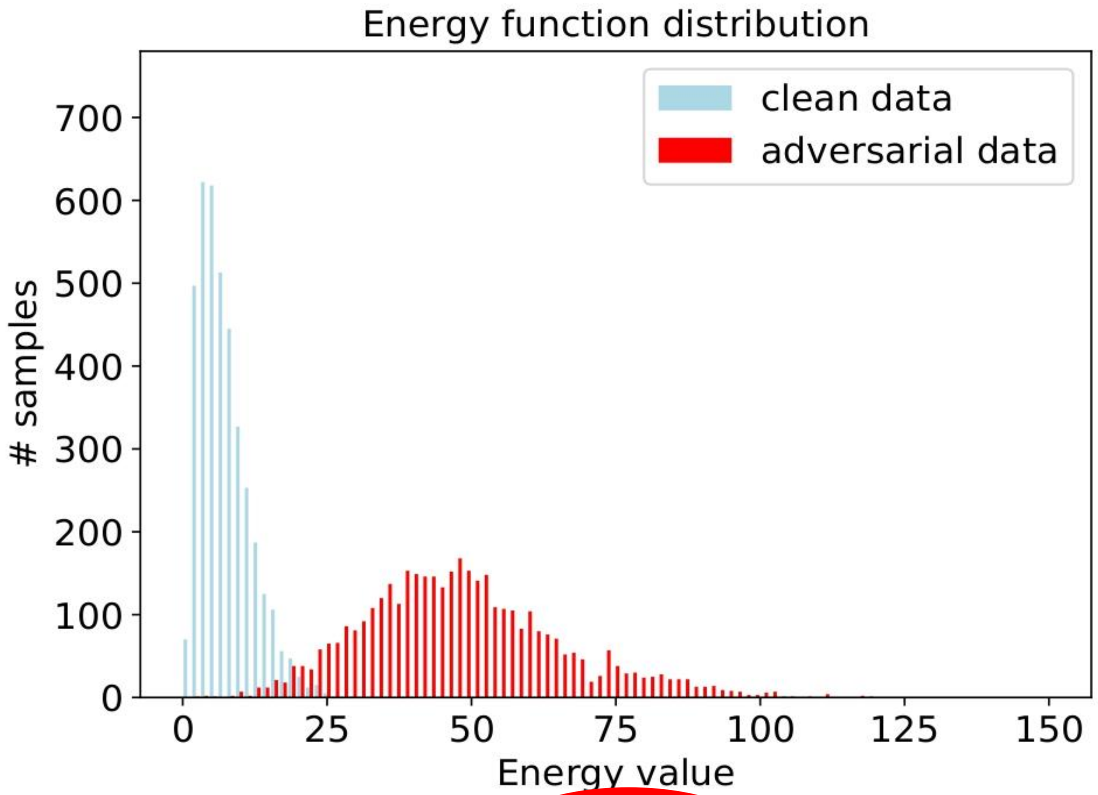
The energy function is the denominator of the softmax operation.

[taken from ICLR2020]

Energy Function vs Attack Strength



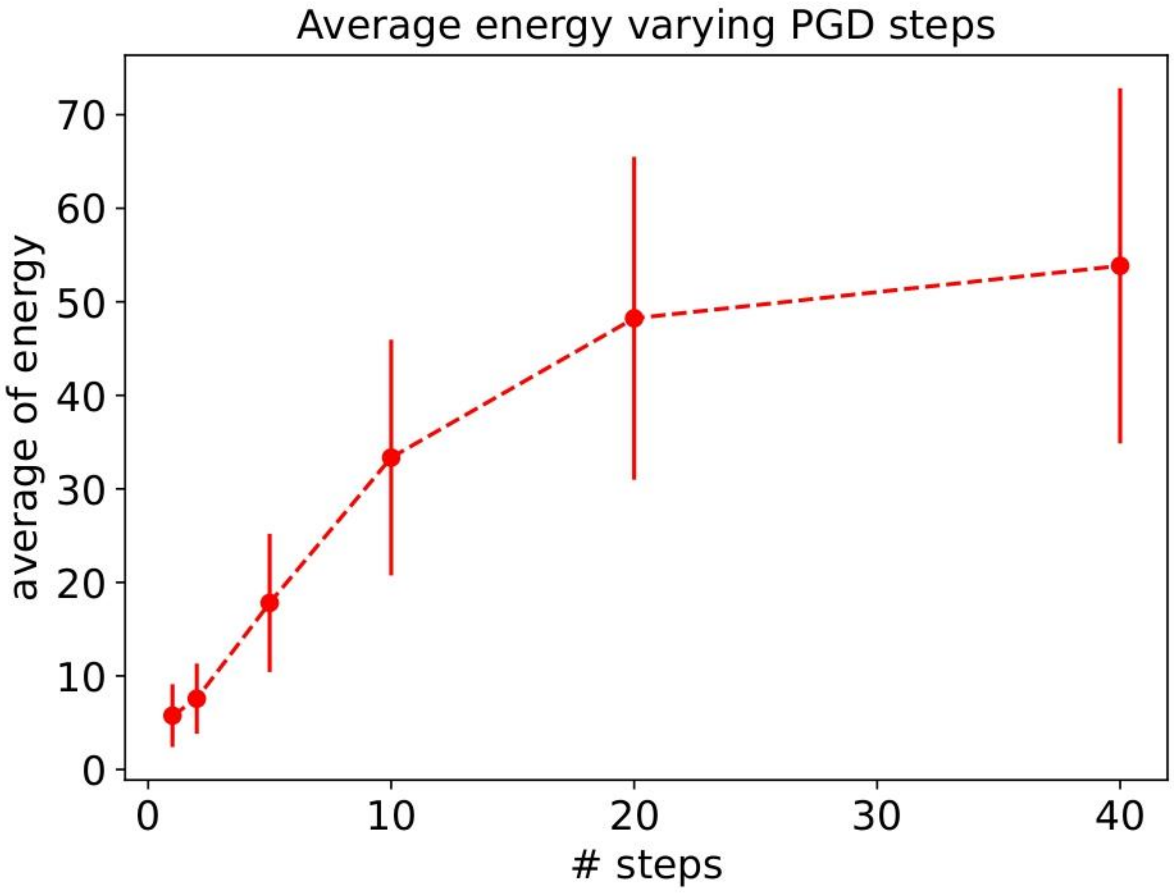
PGD-5



PGD-20

Projected Gradient Descent (PGD) Madry et al. ICLR2018

Energy Function vs Attack Strength



The attack strength is correlated with the energy value.

Notice that the norm of the perturbation is bounded, we only increase the iterations of the attack.

The increase in energy is a "clue" that we are generating OOD data.

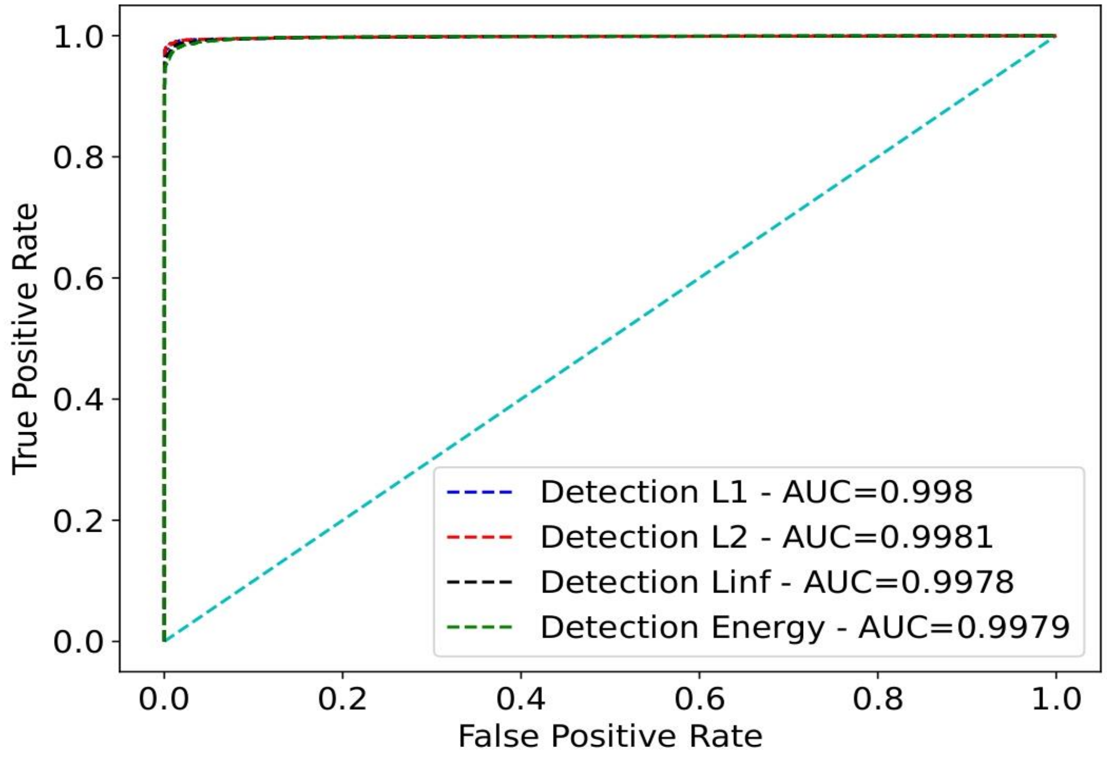
Projected Gradient Descent (PGD) Madry et al. ICLR2018

Energy-based Adversarial Examples Detection

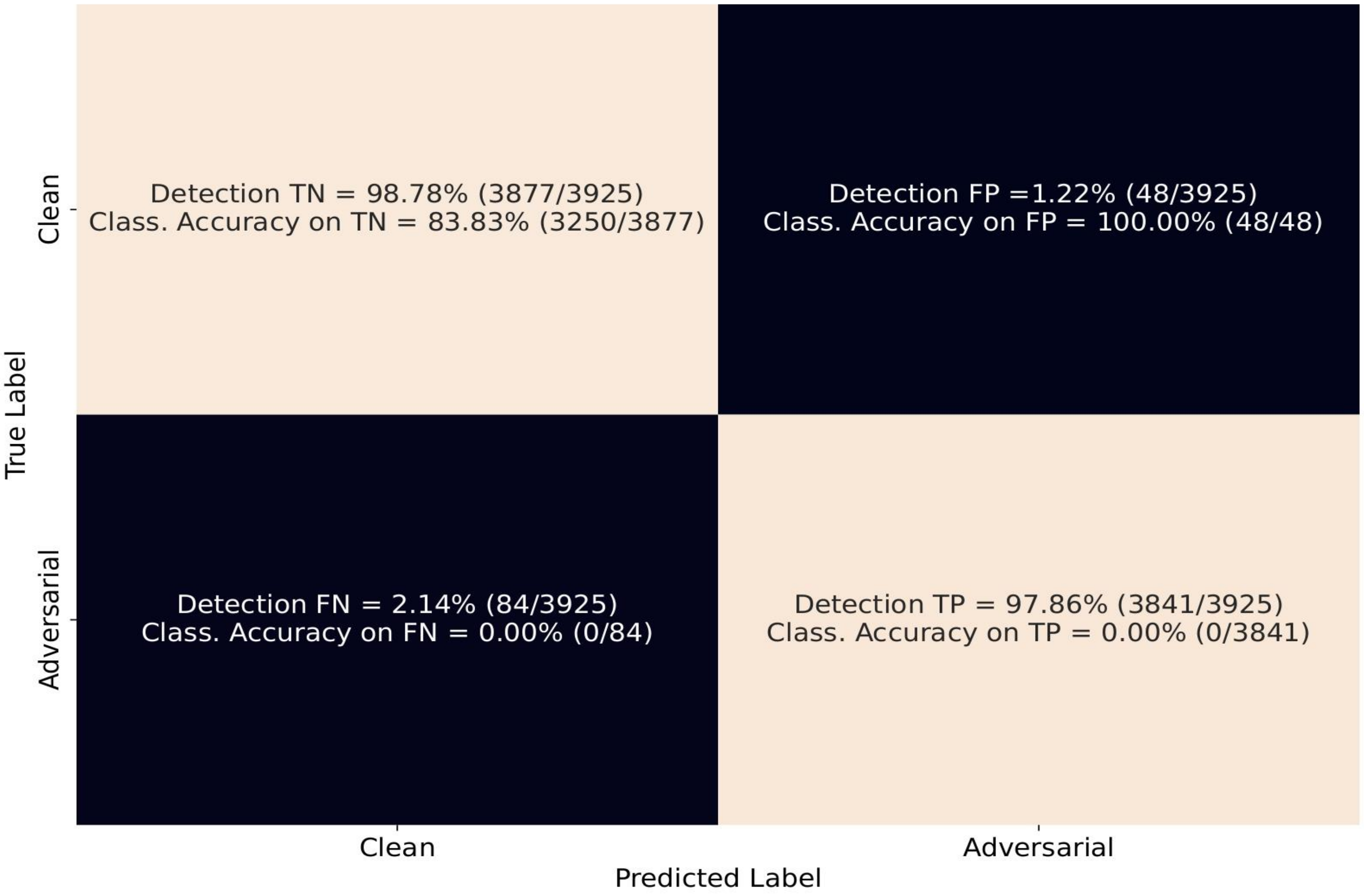
Exploiting the previous result is possible to build a detector for adversarial data on the top of energy.

The threshold can be optimized on a given metric in the validation set.

ROC under PGD-30 attack



Energy-based Adversarial Examples Detection



Under PGD-30

TPR = 98%

FPR = 1%

Can we bypass the energy-based detection algorithm?

That is, can we craft an adversarial attack that fools the classifier yet has low energy?

Low Energy Projected Gradient Descent (LE-PGD)

PGD is a constrained optimization procedure.

PGD builds a perturbation through an iterative procedure in which at each step the direction of the gradient is followed to maximize the loss. The result is clipped to bound the Lp norm.

PGD

$$\boldsymbol{x}^* = \text{clip}_\epsilon \left(\boldsymbol{x}^* + \alpha \nabla_{x^*} \mathcal{L}_\theta(\boldsymbol{x}^*, y) \right)$$

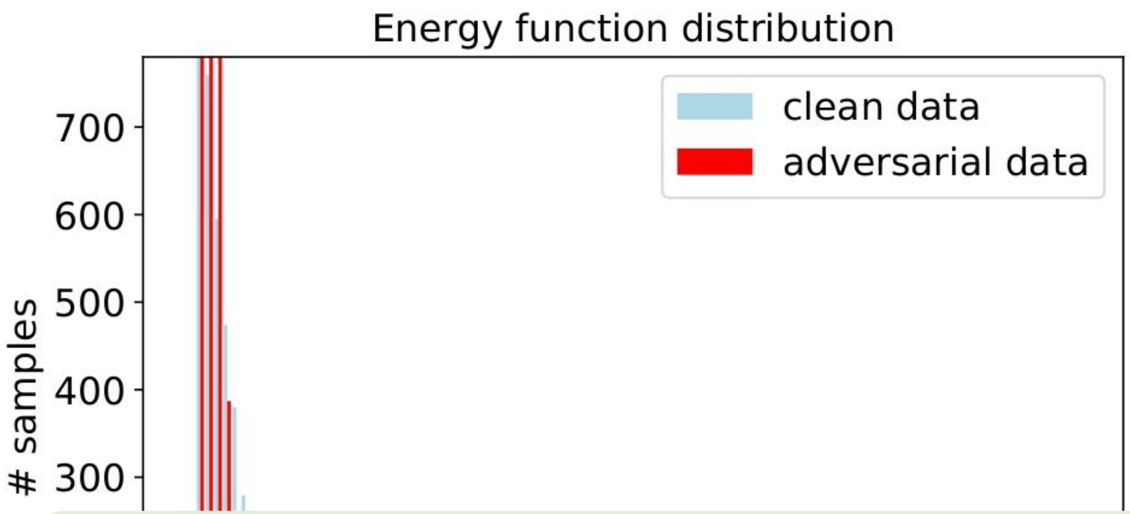
LE-PGD

$$\boldsymbol{x}^* = \text{clip}_\epsilon \left[\boldsymbol{x}^* + \alpha \nabla_{x^*} \left(\mathcal{L}_\theta(\boldsymbol{x}^*, y) - \lambda\, E_\theta(\boldsymbol{x}^*) \right) \right]$$

We add an energy regularizer in the optimization procedure.

Projected Gradient Descent (PGD) Madry et al. ICLR2018

Low Energy Projected Gradient Descent (LE-PGD)



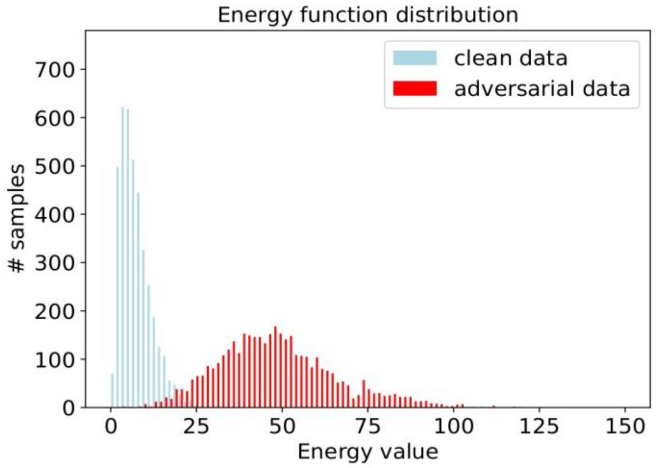
The two distributions completely overlap.

Under LE-PGD the CNN accuracy still drops to 0%.

LE-PGD produces adversarial data that are «in-distribution» according to the model but the accuracy is still 0%.

Conclusion and Take Home Message

- **We illustrated that it is not always possible to have both pixel-wise and transformation robustness for CNNs.**



- **We have shown that there is a correlation between energy and attack strength. Under this validation, we provide a detector for pixel-wise adversarial examples.**

- **We proposed a modification of the PGD attack to produce low energy adversarial examples to break the energy-based detector.**

$$\boldsymbol{x}^* = \mathrm{clip}_\epsilon \left[\boldsymbol{x}^* + \alpha \nabla_{x^*} \Big(\mathcal{L}_\theta(\boldsymbol{x}^*, y) - \lambda\, E_\theta(\boldsymbol{x}^*) \Big) \right]$$

Thank you!

Detector Performances for Different Attacks

Attack	ϵ	DR	FPR	FNR
PGD-5	8/255	83%	19%	17%
PGD-10	8/255	93%	4%	7%
PGD-30	8/255	98%	1%	2%
PGD-50	8/255	98%	0.7%	2%
PGD-50	16/255	99%	0.02%	0.1%

Energy-based Model (EBM)

While most of the previous models had the goal of classification, EBMs are motivated from a different perspective: density estimation.

The fundamental idea of EBM is the energy function such that we approximate $p(x)$ via :

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\exp\left(-E_{\theta}(\boldsymbol{x})\right)}{\int_x \exp\left(-E_{\boldsymbol{\theta}}(\boldsymbol{x})\right) d\boldsymbol{x}}$$

Notice that : $$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = -E_{\boldsymbol{\theta}}(\boldsymbol{x}) - Z_{\boldsymbol{\theta}}$$ Constant for all *x*

Data points with high likelihood have a low energy, while data points with low likelihood have a high energy.

Beyond Accuracy Metric :

We would like to have probabilistic models that are highly confident in their correct predictions and low in their incorrect ones.

Given 100 prediction, each with confidence 0.8, we expect that 80 should be correctly classified.

Despite their high accuracy modern CNNs are not well-calibrated.

A CNN overestimates/underestimates its predictions.

[Guo et al. ICML 2017]

ECE Metric :

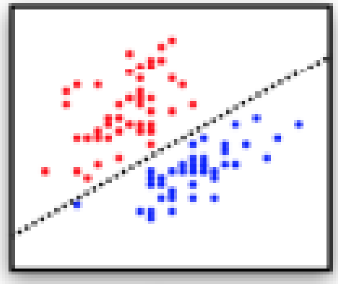
Divide the interval [0,1] in m bins. Then, sort the predictions according the winning confidence class.

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \mathrm{Acc}(B_m) - \mathrm{Conf}(B_m) \right|$$

What is Classification?

Input data

High dimensional data

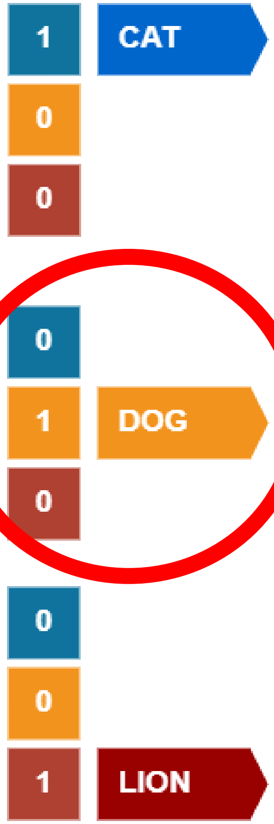


$$p(y|\boldsymbol{x})$$

Model

Mapping Procedure

Output



Low dimensional data