Adversarial examples are not mysterious, generalization is

Angus Galloway
University of Guelph
gallowaa@uoguelph.ca

THE ADVERSARIAL EXAMPLES PHENOMENON

Machine learning models generalize well to an unseen test set, yet every input of a particular class is extremely close to an input of another class.

"Accepted" informal definition:
Any input designed to fool a machine learning system.

FORMAL DEFINITIONS

A "misclassification" adversarial candidate $\hat{x}$ on a neural network F with input x via some perturbation of x by δ:

$$\hat{x} = x + \delta$$

where δ is usually derived from the gradient of the loss $\nabla L(\theta, y, x)$ w.r.t x, and for some small scalar ϵ,

$$\|\delta\|_p \leq \epsilon, p \in \{1, 2, \infty\}$$

such that

$$F(x) \neq F(\hat{x})$$

GOODFELLOW ET AL. 2015

For input $x \in \mathbb{R}^n$, there is an adversarial example $\tilde{x} = x + \eta$ subject to the constraint $\|\eta\|_\infty < \epsilon$. The dot product between a weight vector w and an adversarial example $\tilde{x}$ is then:

$$w^T \cdot \tilde{x} = w^T \cdot x + w^T \cdot \eta$$

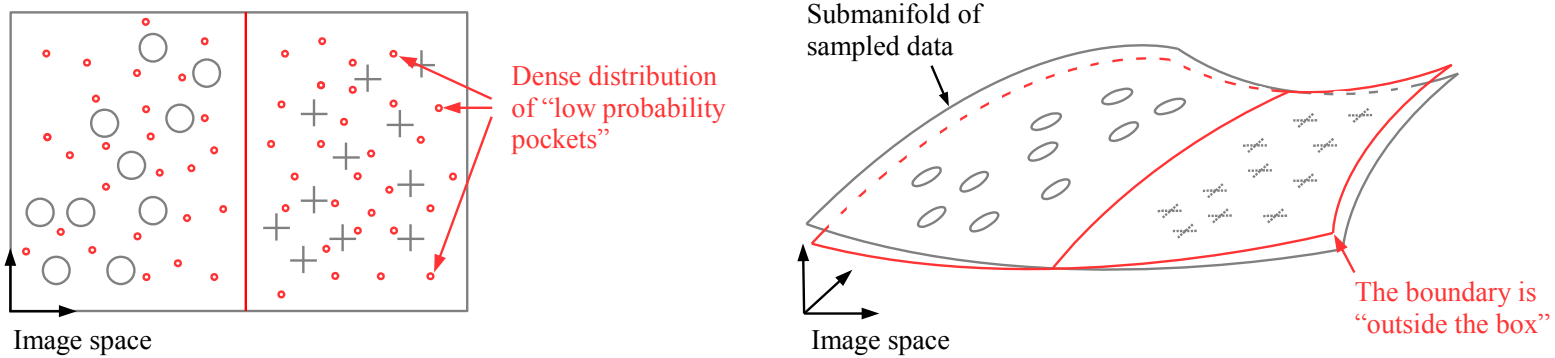
If w has mean m, activation grows linearly with $\epsilon\, m\, n \ldots$

TANAY & GRIFFIN 2016

But both $w^T \cdot x$ **and** $w^T \cdot \eta$ grow linearly with dimension n,

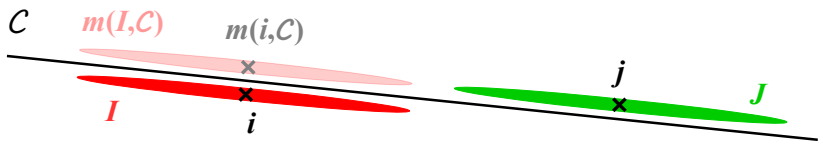
provided that the distribution of w and x do not change.

THE BOUNDARY TILTING PERSPECTIVE



(a) (b)

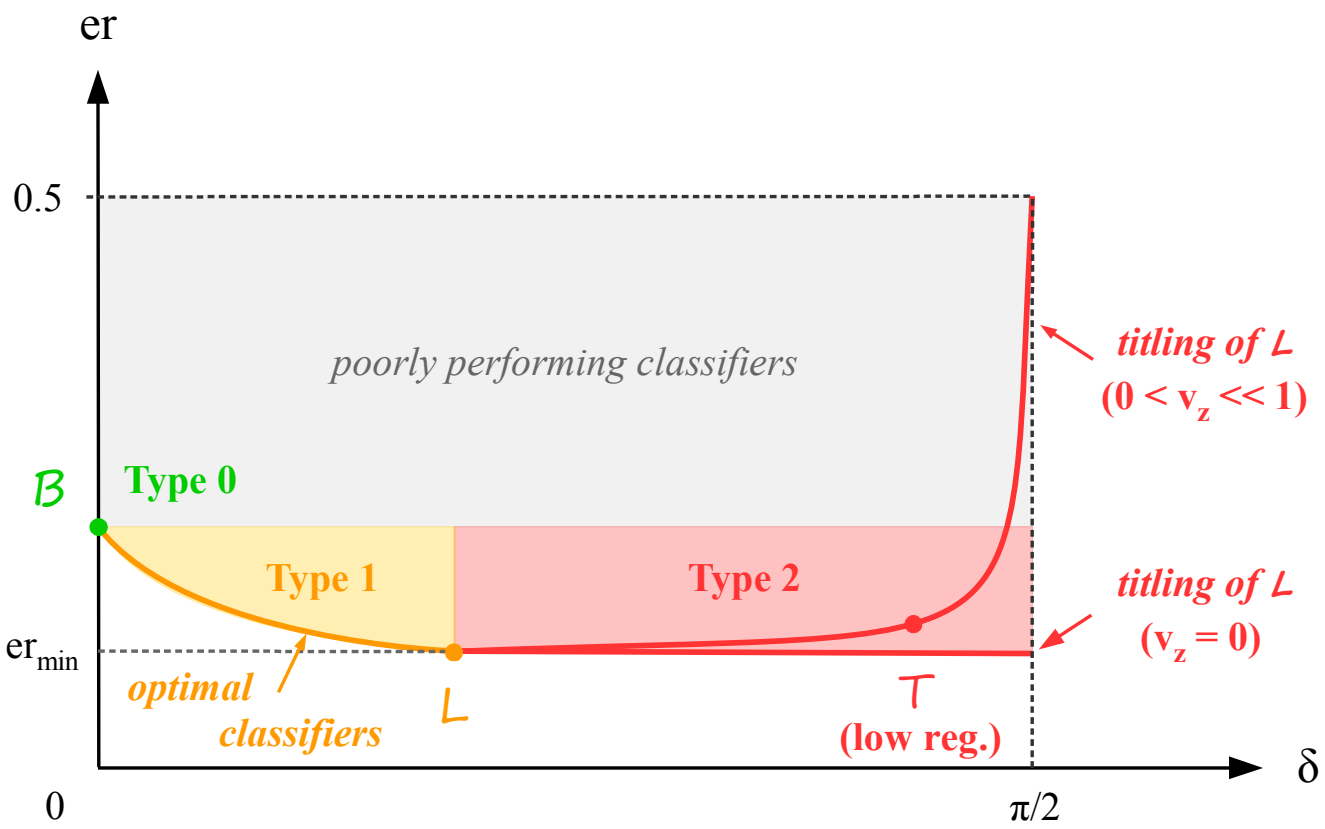
Recall **manifold learning hypothesis**: training data sub-manifold exists with finite topological dimension $f \langle\langle n$.



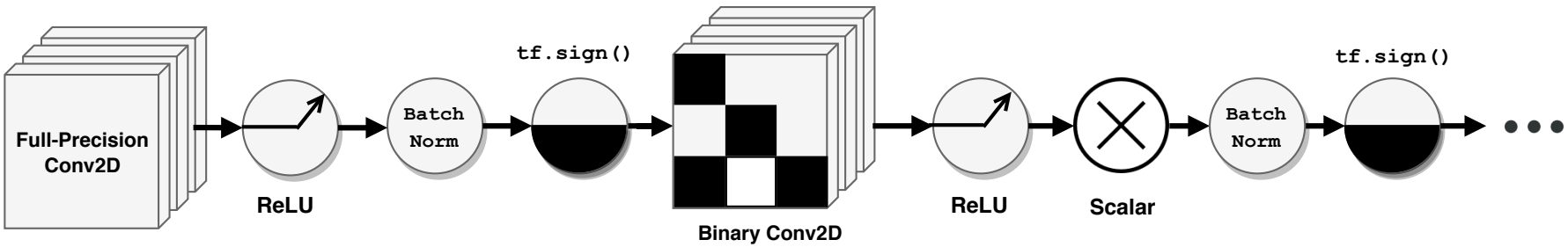
(c)

(d)

TAXONOMY


er
0.5
poorly performing classifiers
titling of L
(0 < v_z << 1)
B
Type 0
Type 1
Type 2
titling of L
(v_z = 0)
er_min
optimal
classifiers
L
T
(low reg.)
δ
0
π/2

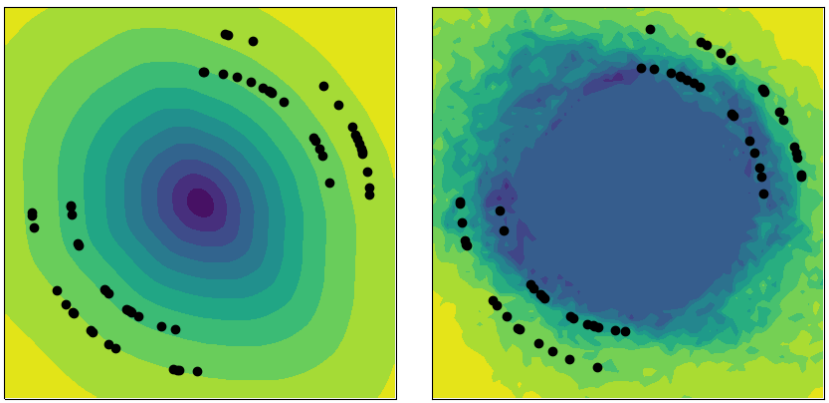
ATTACKING BINARIZED NEURAL NETWORKS



The empirical observation that BNNs with low-precision weights and activations are at least as robust as their full-precision counter parts.

ATTACKING BINARIZED NEURAL NETWORKS (2)

1. Regularizing effect due to decoupling between continuous and quantized parameters used in forward pass, biased gradient estimator (STE?)
2. Strikes better trade-off on IB curve in over-parameterized regime by discarding irrelevant information.

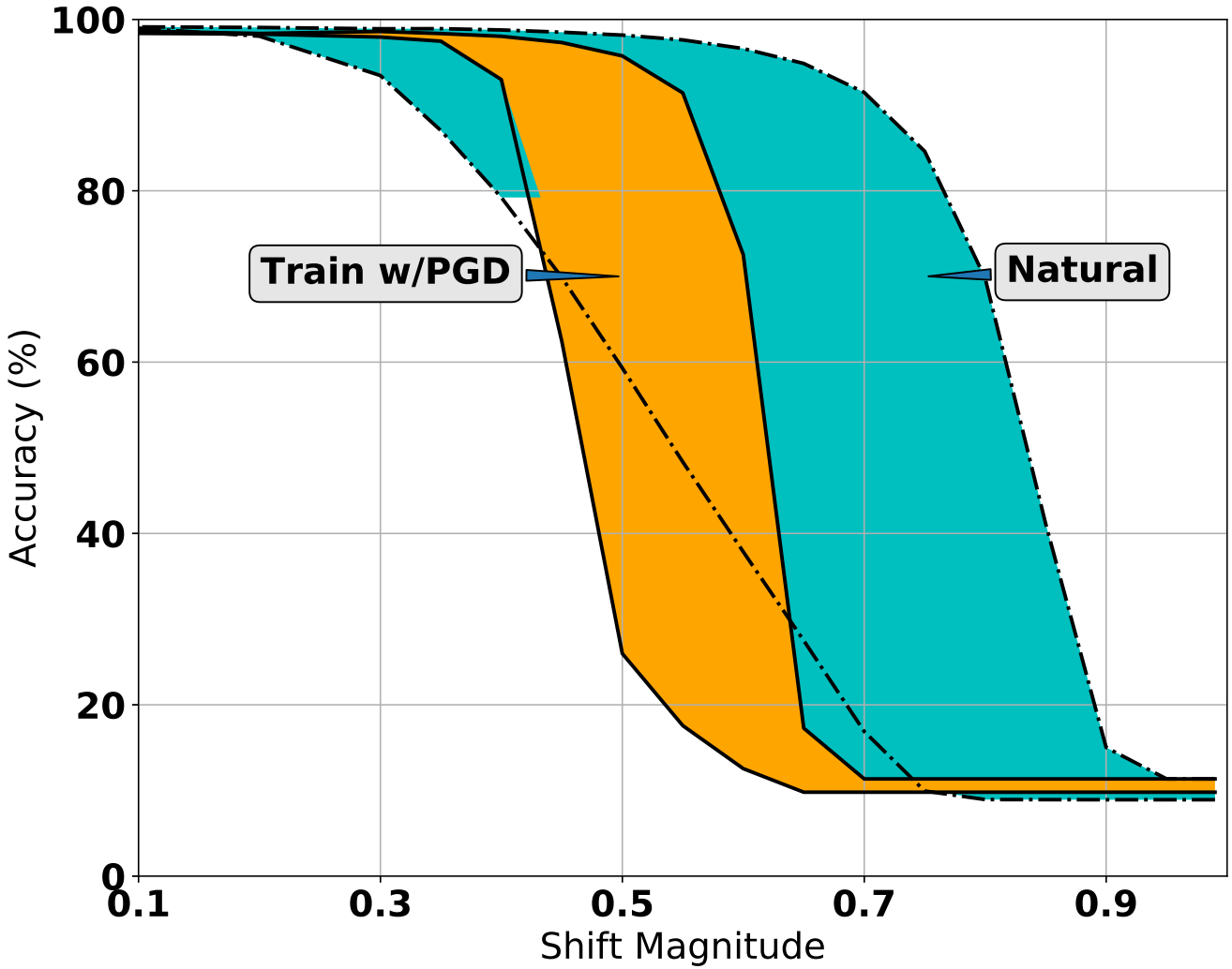


(e) (f)

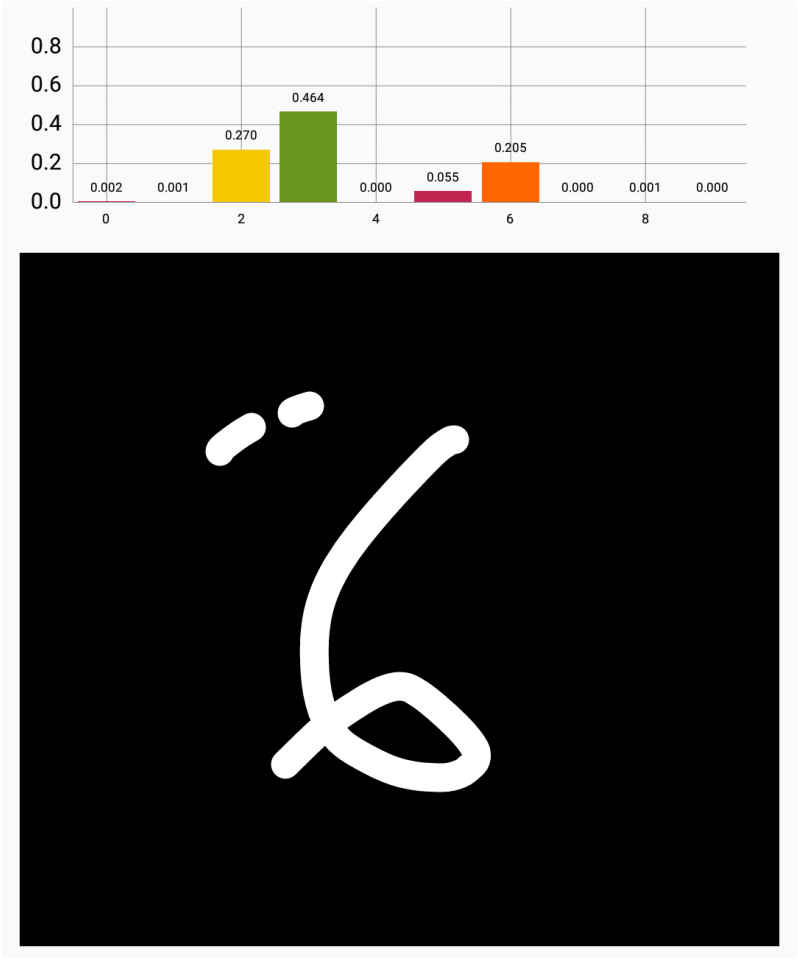
WHY ONLY CONSIDER SMALL PERTURBATIONS?

Fault tolerant engineering design:

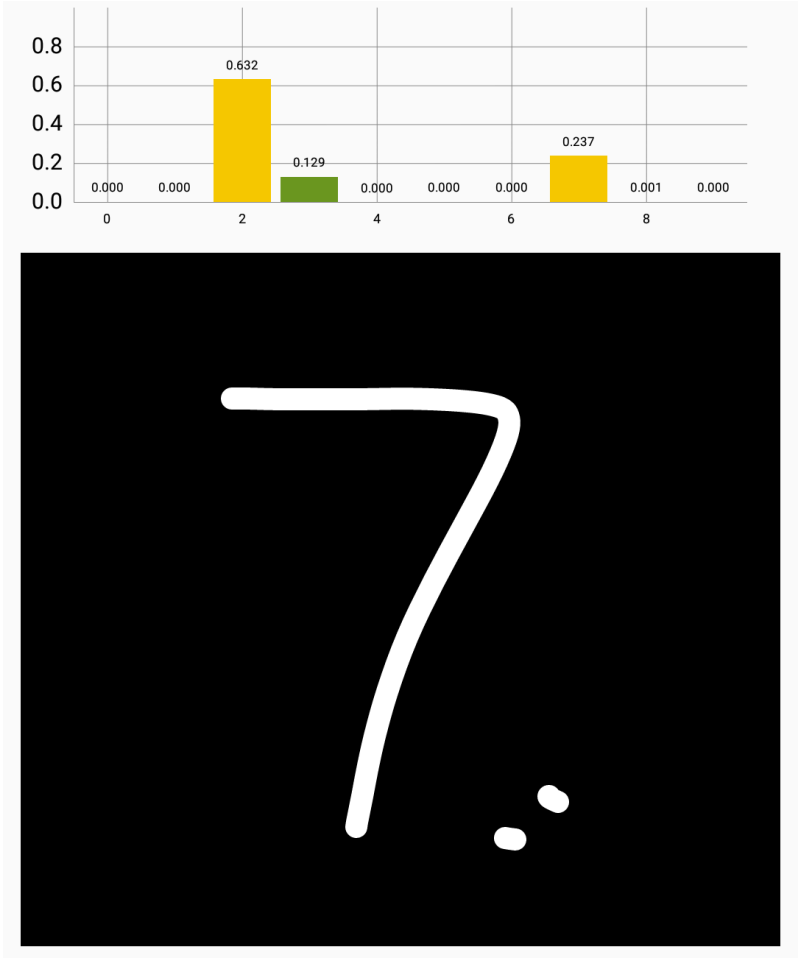
Want performance degradation to be proportional to perturbation magnitude, regardless of an attacker's strategy.


100
80
60
40
20
0
Accuracy (%)
Train w/PGD
Natural
0.1
0.3
0.5
0.7
0.9
Shift Magnitude

HUMAN-DRIVEN ATTACKS

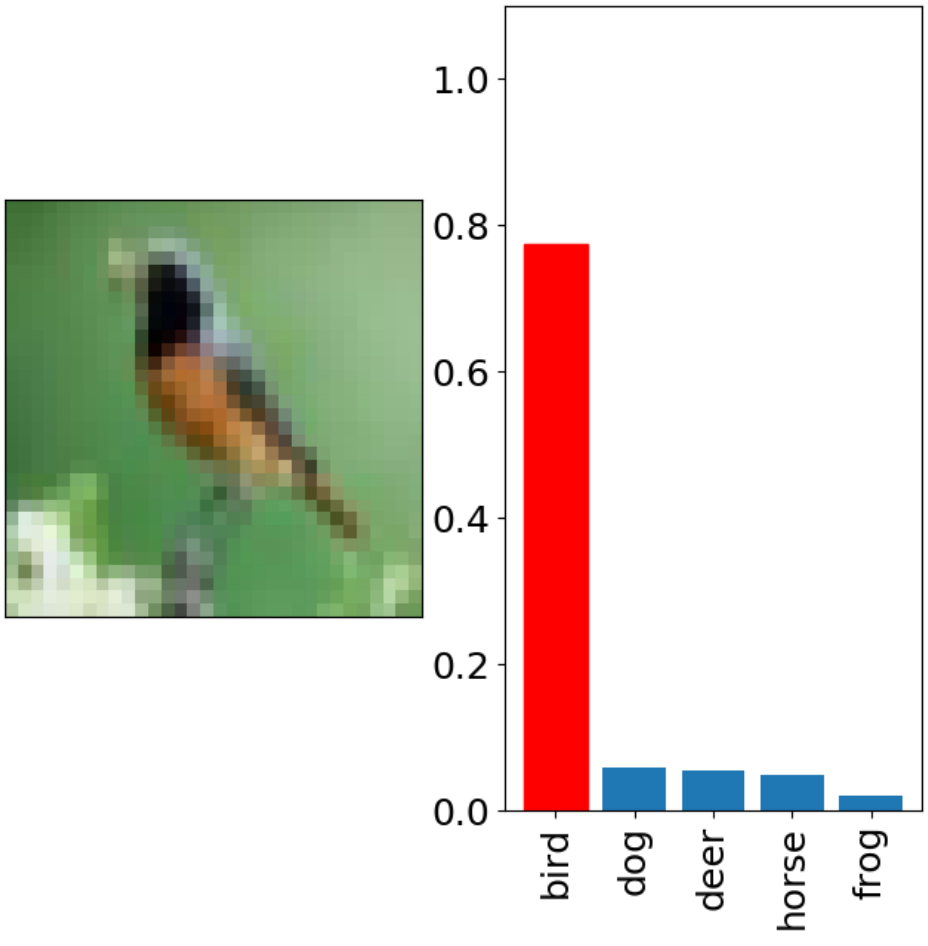


(g)

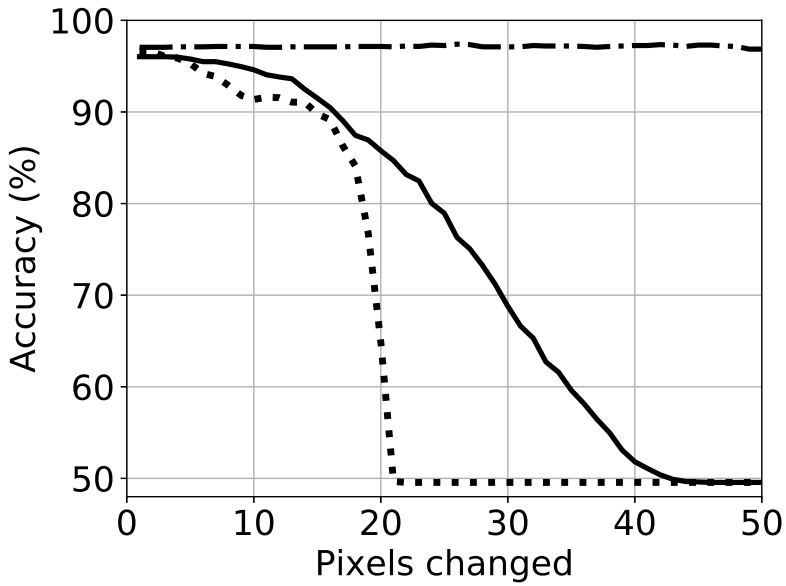


(h)

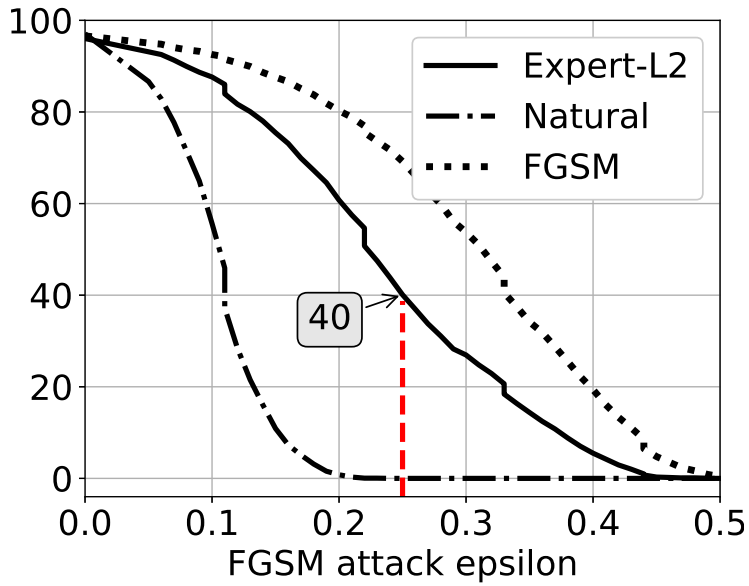
A PRACTICAL BLACK-BOX ATTACK



TRADE-OFFS



(i)

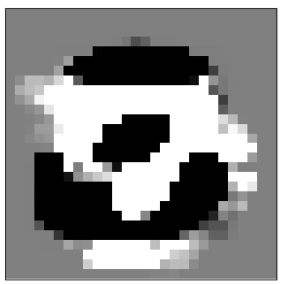


(j)

INTERPRETABILITY OF LOGISTIC REGRESSION

(k) (l) (m) (n)

Candidate Examples



(o)

(p)

(q)

(r)

(s)

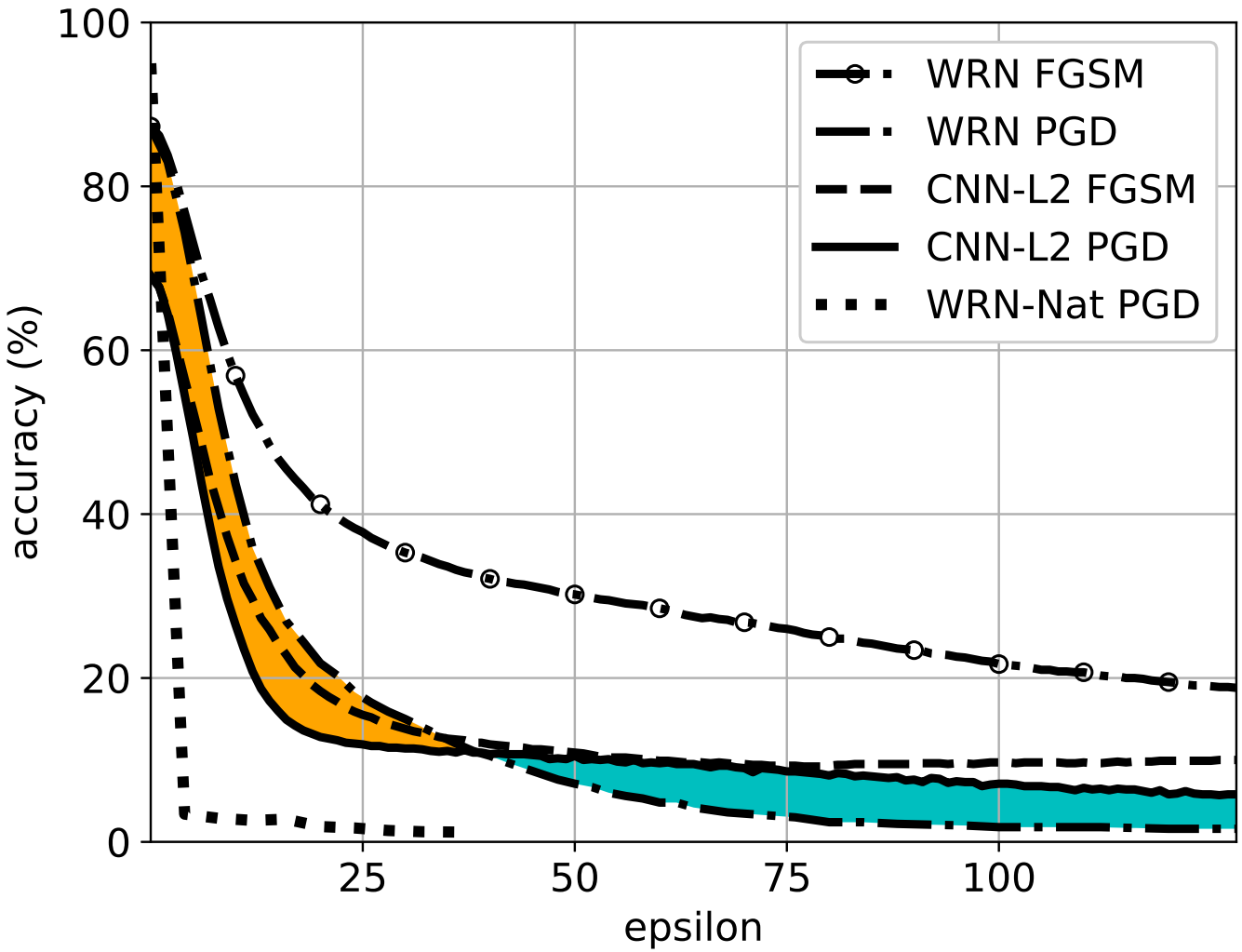
CIFAR-10 architecture

Table: Simple fully-convolutional architecture adapted from the CleverHans library. Model uses ReLU activations, and does not use batch normalization or pooling.

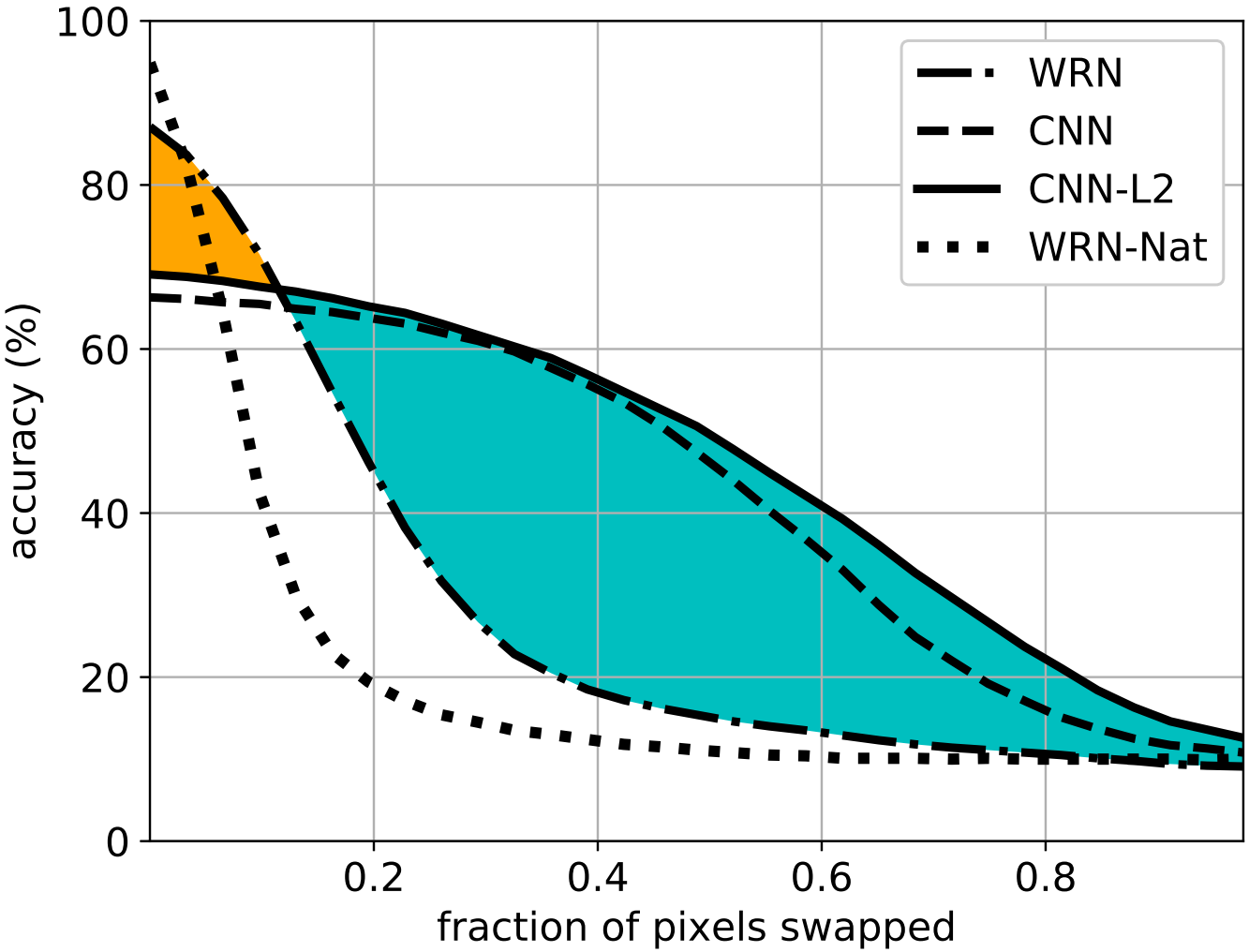
Layer	h	w	c_{in}	c_{out}	s	params
`Conv1`	8	8	3	32	2	6.1k
`Conv2`	6	6	32	64	2	73.7k
`Conv3`	5	5	64	64	1	102.4k
`Fc1`	1	1	256	10	1	2.6k
Total	–	–	–	–	–	**184.8k**

Model has 0.4% as many parameters as WideResNet.

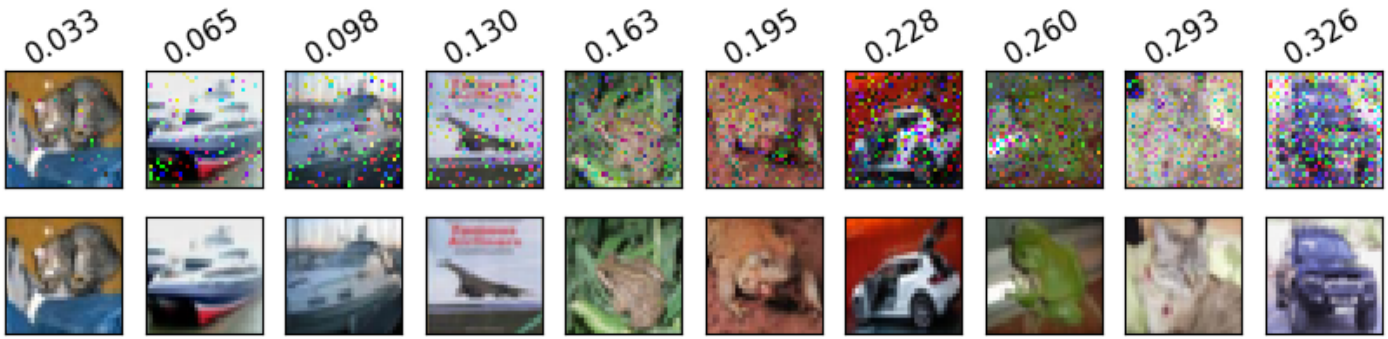
L_∞ Adversarial Examples



Robustness


100
80
60
40
20
0
accuracy (%)
0.2
0.4
0.6
0.8
fraction of pixels swapped
WRN
CNN
CNN-L2
WRN-Nat

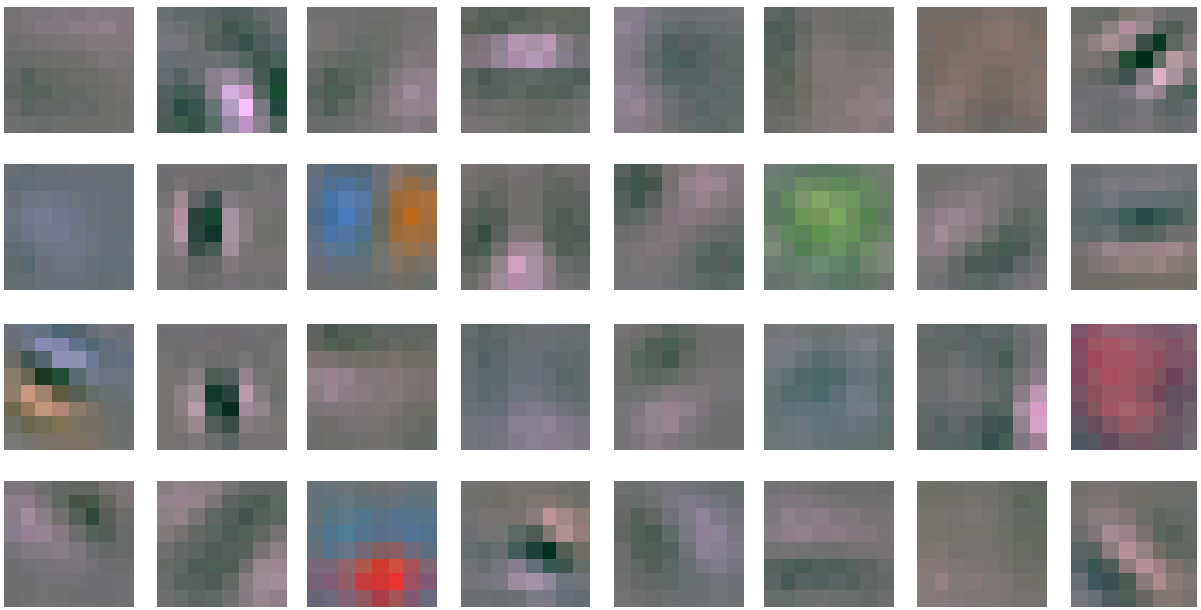
NOISY EXAMPLES



WITH L_2 WEIGHT DECAY

The "independent components" of natural scenes are edge filters (Bell & Sejnowski 1997).

WITHOUT WEIGHT DECAY

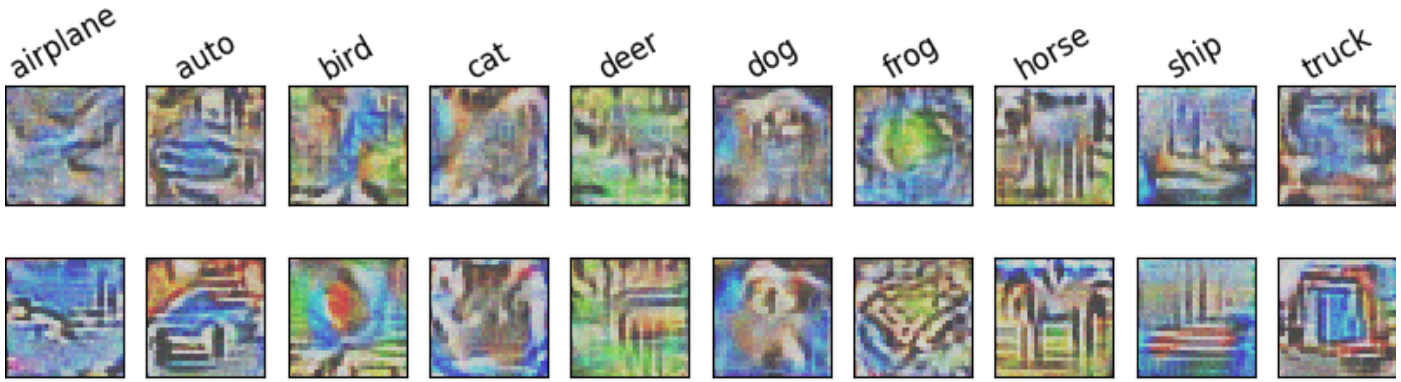


Fooling Images

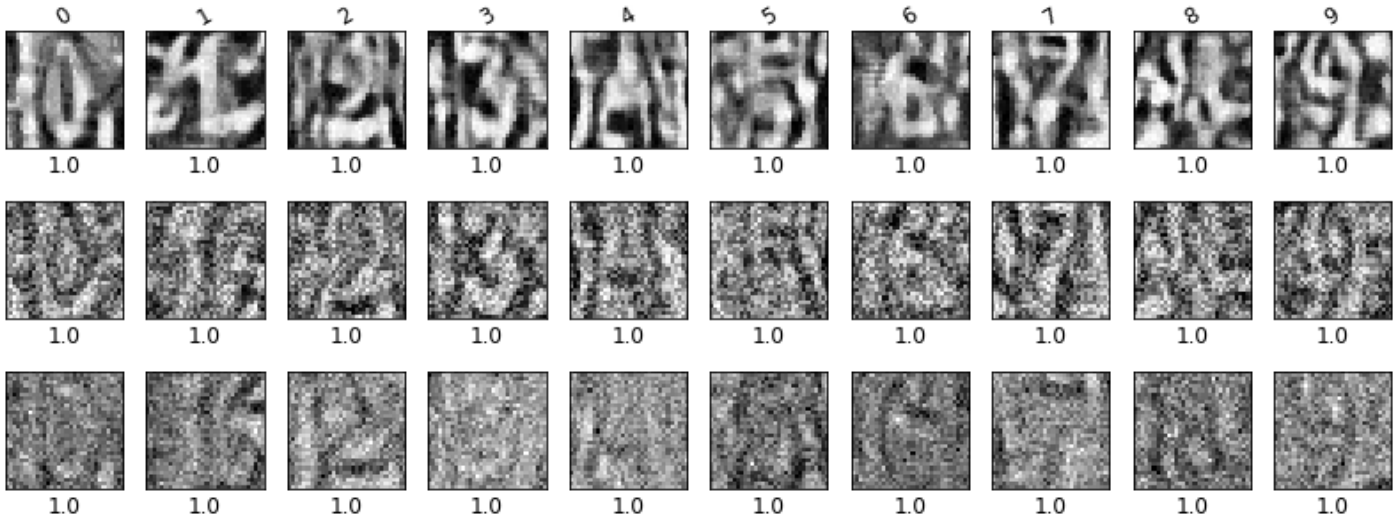
4 years ago I didn't think small-perturbation adversarial examples were going to be so hard to solve. I thought after another n months of working on those, I'd be basically done with them and would move on to ***fooling attacks****.*

Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images (CVPR 2015)

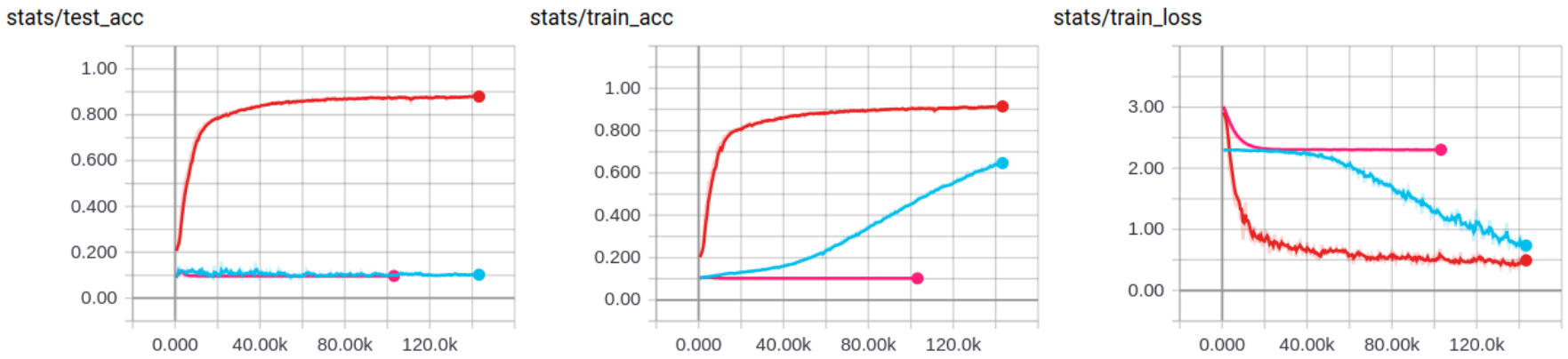
FOOLING IMAGES (CIFAR-10)



FOOLING IMAGES (SVHN)

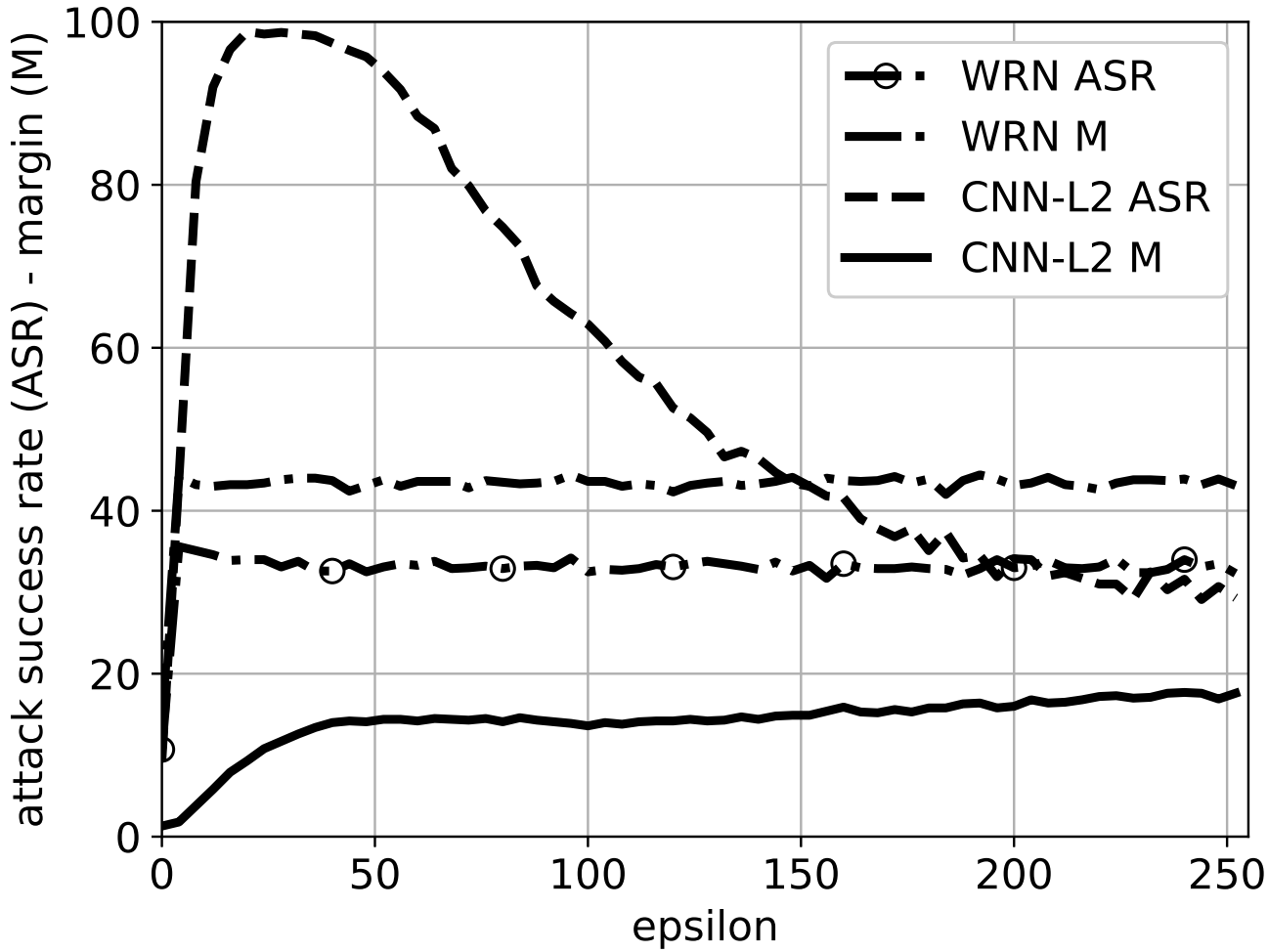


FOOLING IMAGES (SVHN)



Robust training procedure does not learn random labels (lower Rademacher complexity).

Fooling Images



DIVIDE AND CONQUER?

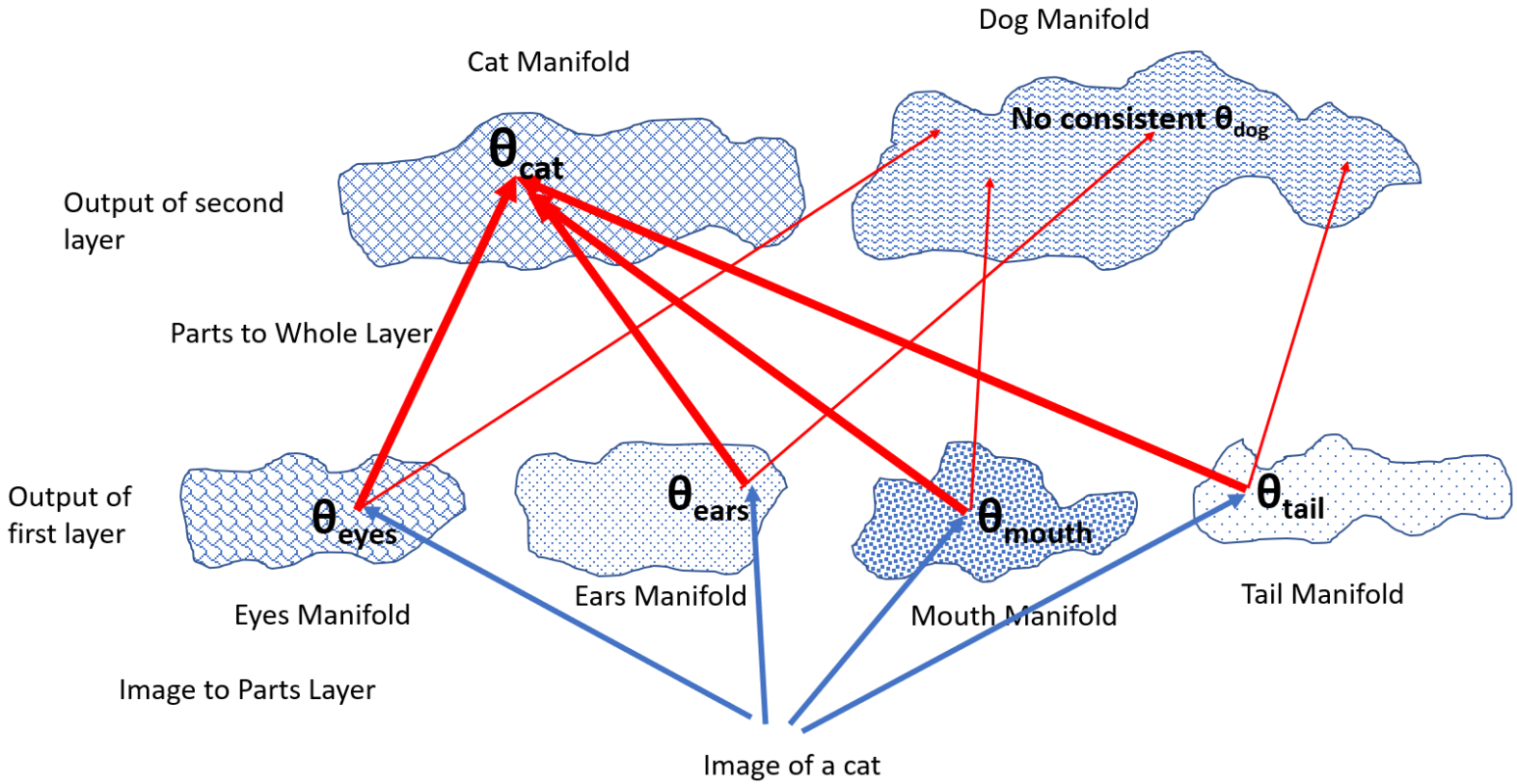


Image from Dube (2018).

REMARKS

- Test accuracy on popular ML benchmarks a weak measure of generalization.
- Plethora of band-aid fixes to std DNNs do not yield compelling results (e.g. provably robust framework).
- Incorporate expert knowledge, e.g. by excplicitly modeling part-whole relationships, other priors that relate to the known causal features such as edges in natural scenes.
- Good generalization implies some level of privacy, and more "fair" models assuming original intent is fair.

FUTURE WORK

Information bottleneck (IB) theory seems essential for efficiently learning robust models from finite data. But why do models with no bottleneck generalize well on common machine learning datasets?

i-RevNet retains all information until final layer and achieves high accuracy, but is extremely sensitive to adversarial examples.