Thermometer Encoding: One Hot Way to Resist Adversarial Examples

Stanford, 2017-11-16

Jacob Buckman*

Aurko Roy*

Colin Raffel

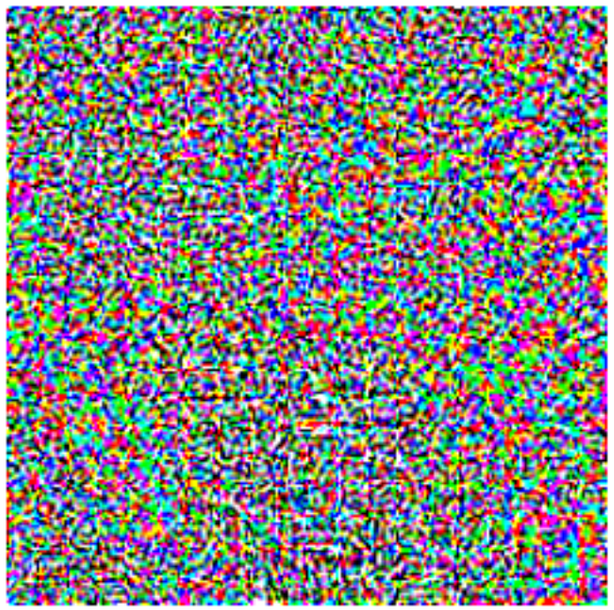
Ian Goodfellow

*joint first author

Adversarial Examples

Probably panda

$+ .007 \times$



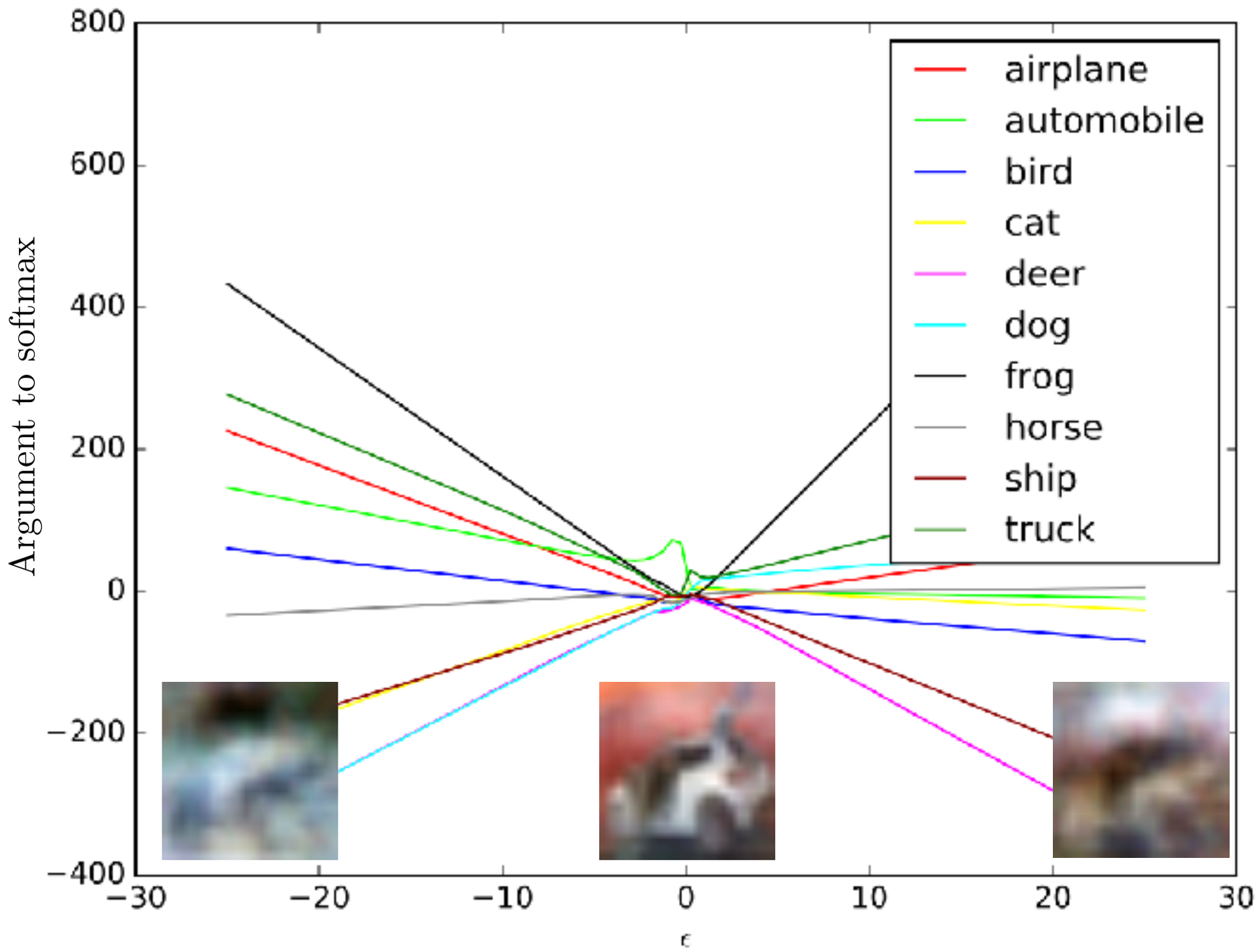
Adversarial perturbation

$=$

Definitely gibbon

Image from "Explaining and Harnessing Adversarial Examples", Goodfellow et al, 2014

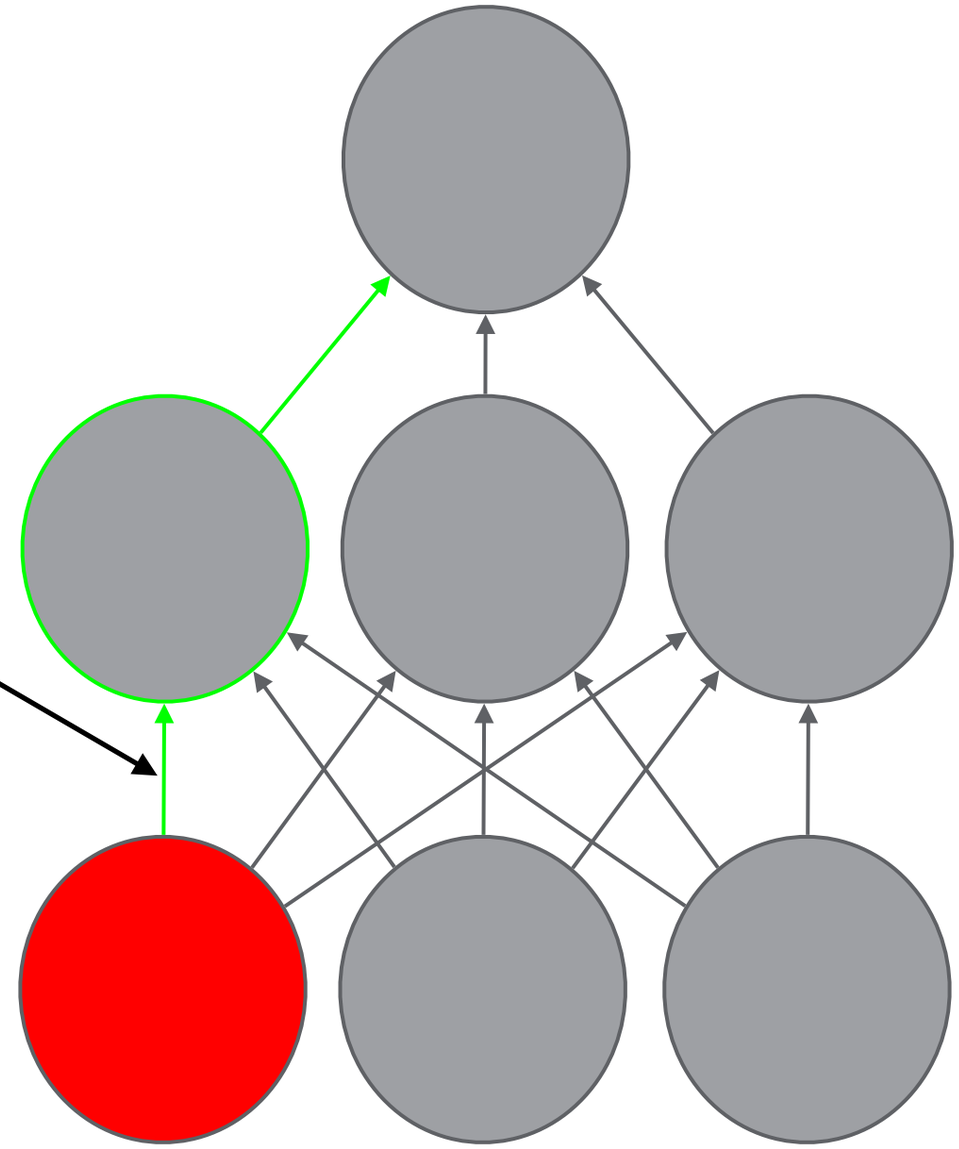
Unreasonable Linear Extrapolation



Plot from "Explaining and Harnessing Adversarial Examples", Goodfellow et al, 2014

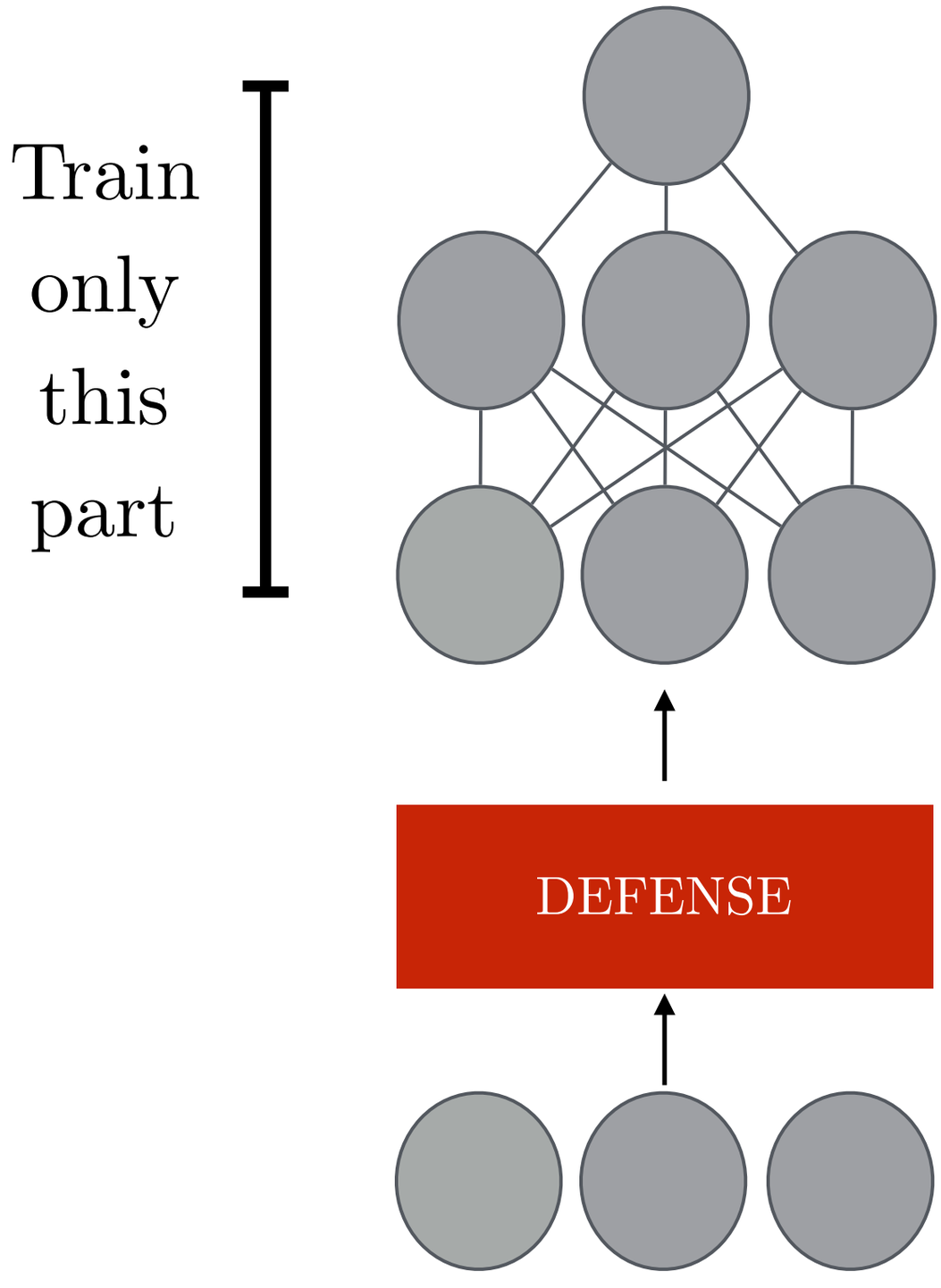
Difficult to train extremely nonlinear hidden layers

To train:
changing this weight needs to have a large, predictable effect



To defend:
changing this input needs to have a small or unpredictable effect

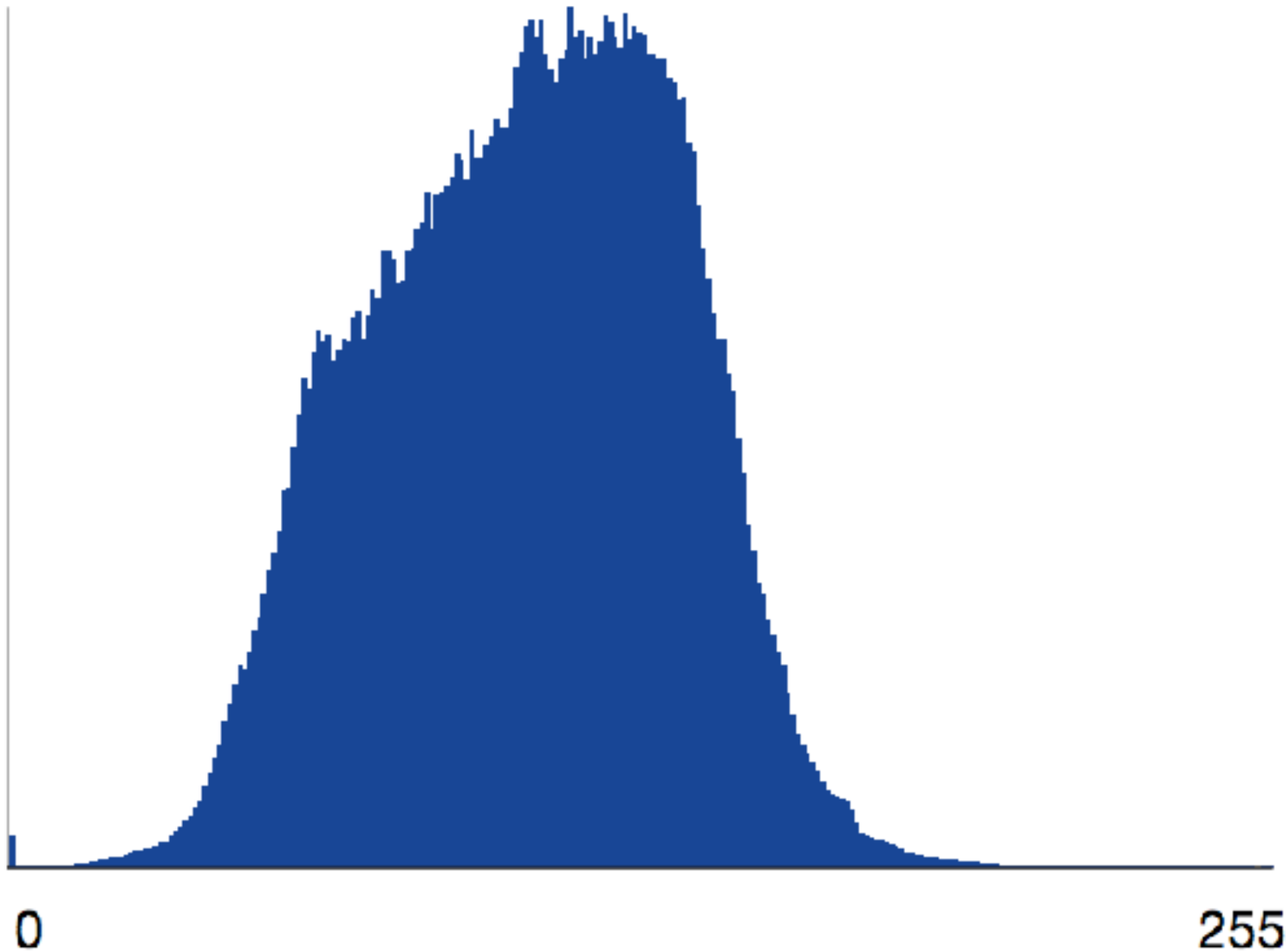
Idea: edit only the input layer



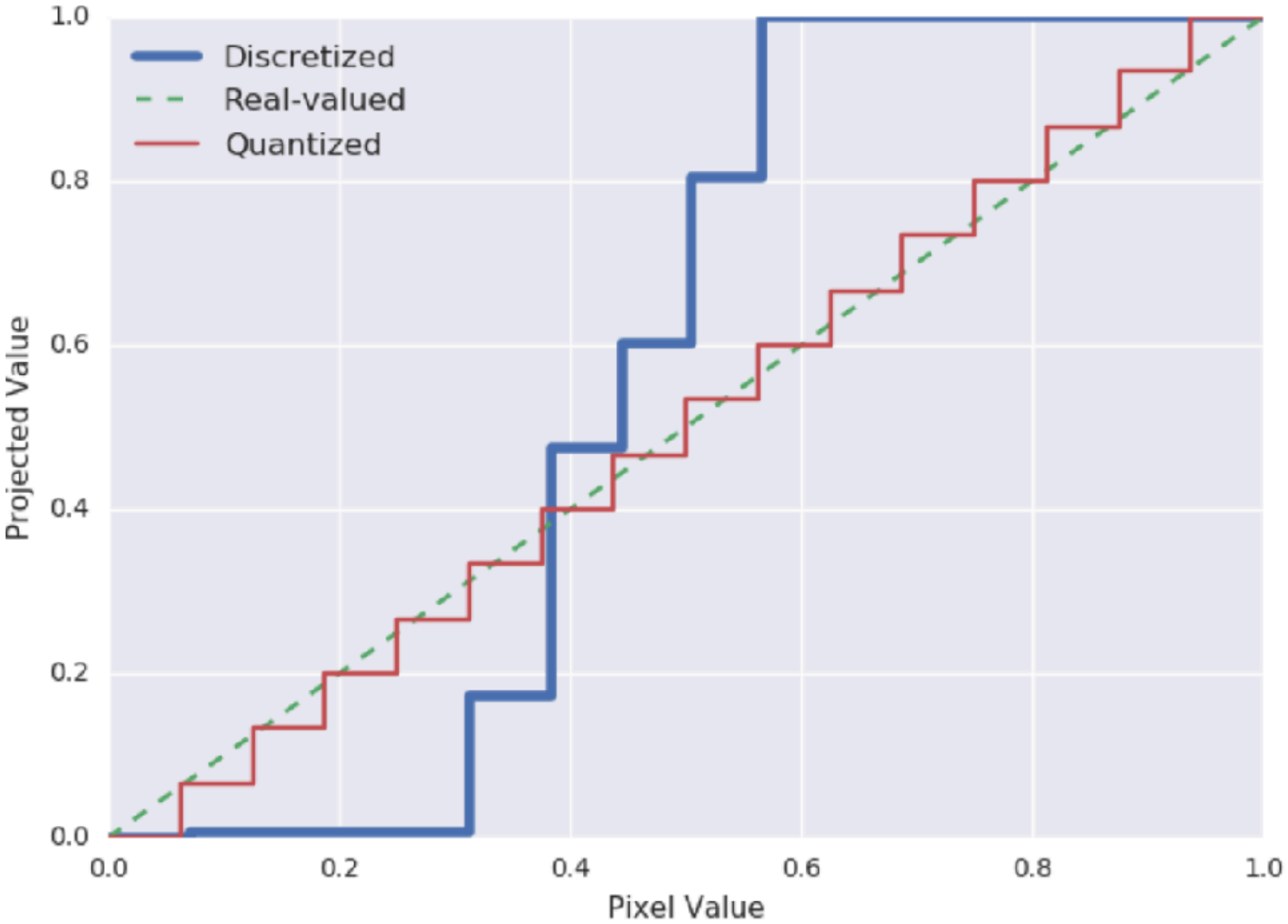
Real-valued	Quantized
0.13	0.15
0.66	0.65
0.92	0.95

Discretized (one-hot)	Discretized (thermometer)
[01000000000]	[01111111111]
[00000010000]	[00000011111]
[00000000001]	[00000000001]

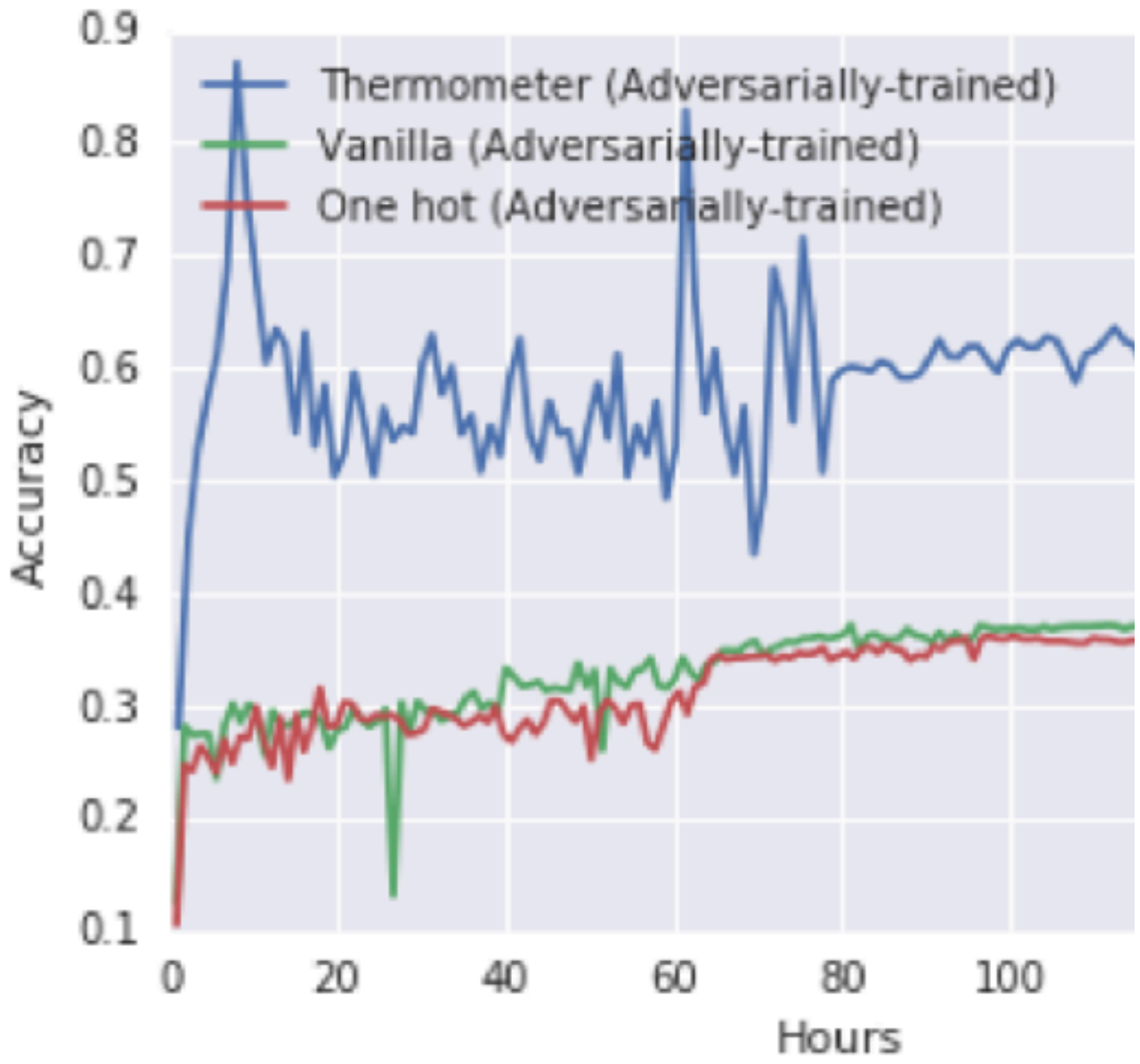
Observation: PixelRNN shows one-hot codes work



Plot from “Pixel Recurrent Neural Networks”, van den Oord et al, 2016

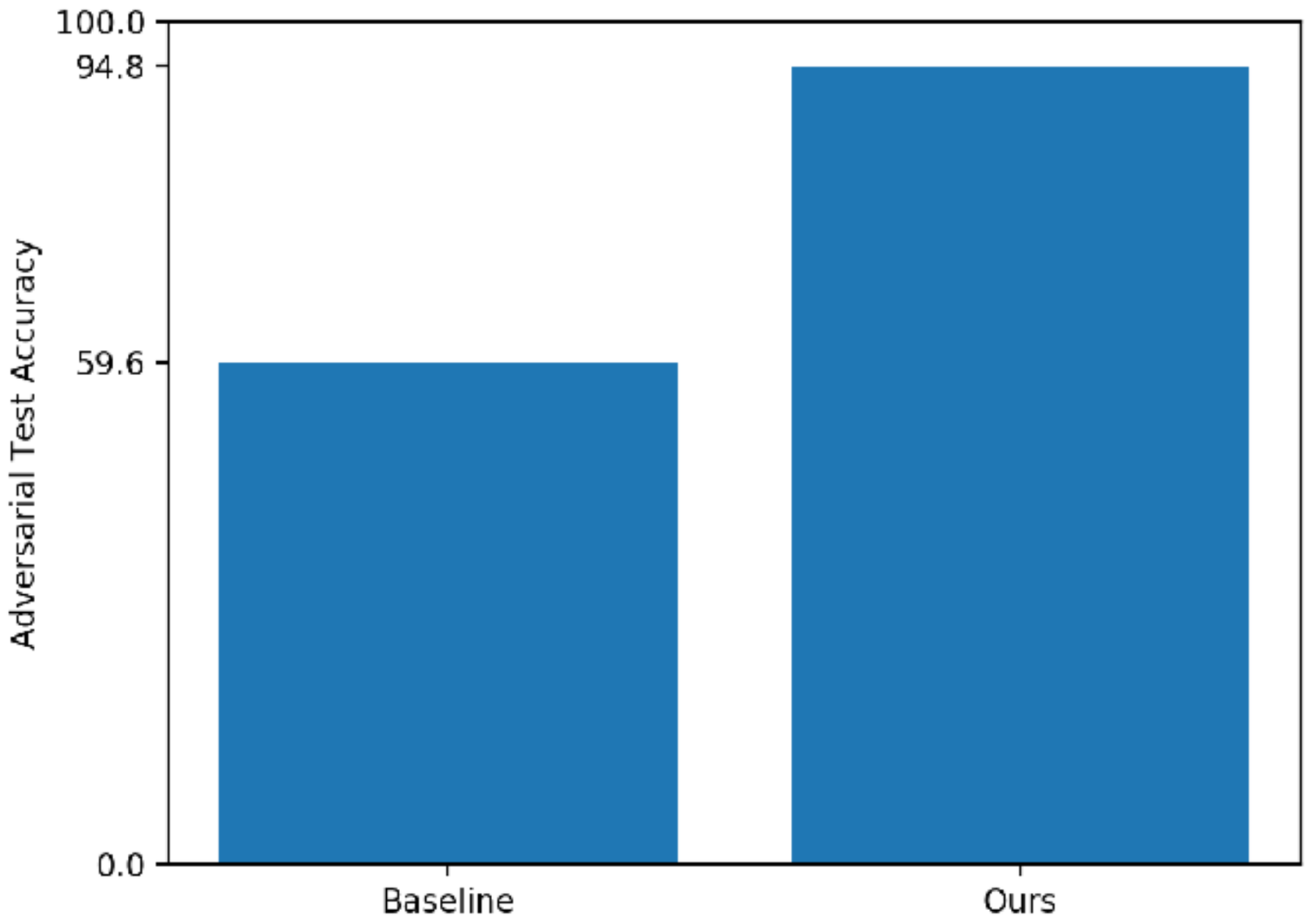

Discretized
Real-valued
Quantized
1.0
0.8
0.6
0.4
0.2
0.0
Projected Value
0.0
0.2
0.4
0.6
0.8
1.0
Pixel Value

Fast Improvement Early in Learning



(Goodfellow 2017)

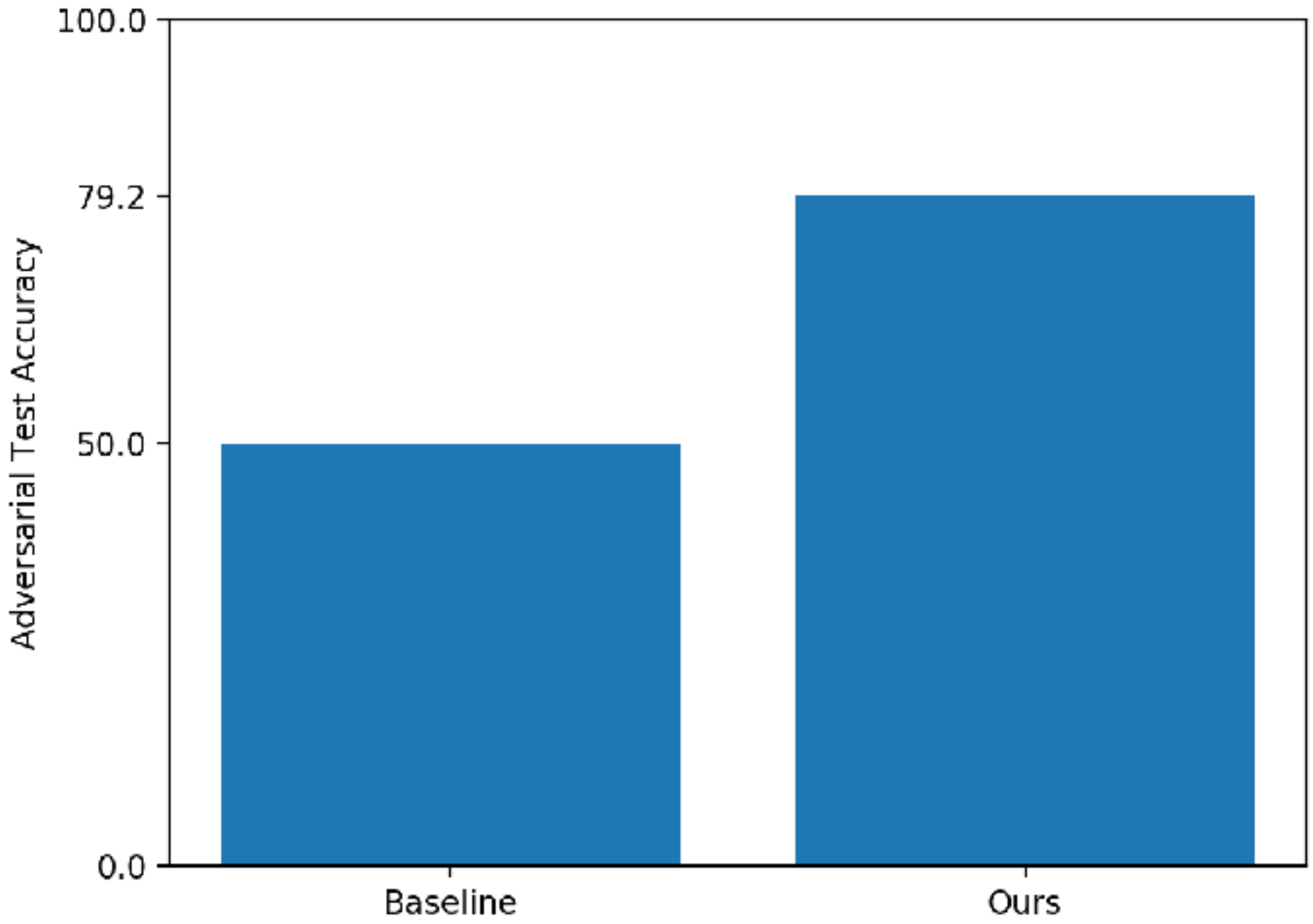
Large improvements on SVHN white box attacks



5 years ago, this would have been SOTA on *clean* data

(Goodfellow 2017)

Large Improvements against CIFAR-10 white box attacks



6 years ago,
this would have
been SOTA
on *clean* data

Other results

- Improvement on CIFAR-100
  - (Still very broken)
- Improvement on MNIST
  - Please quit caring about MNIST

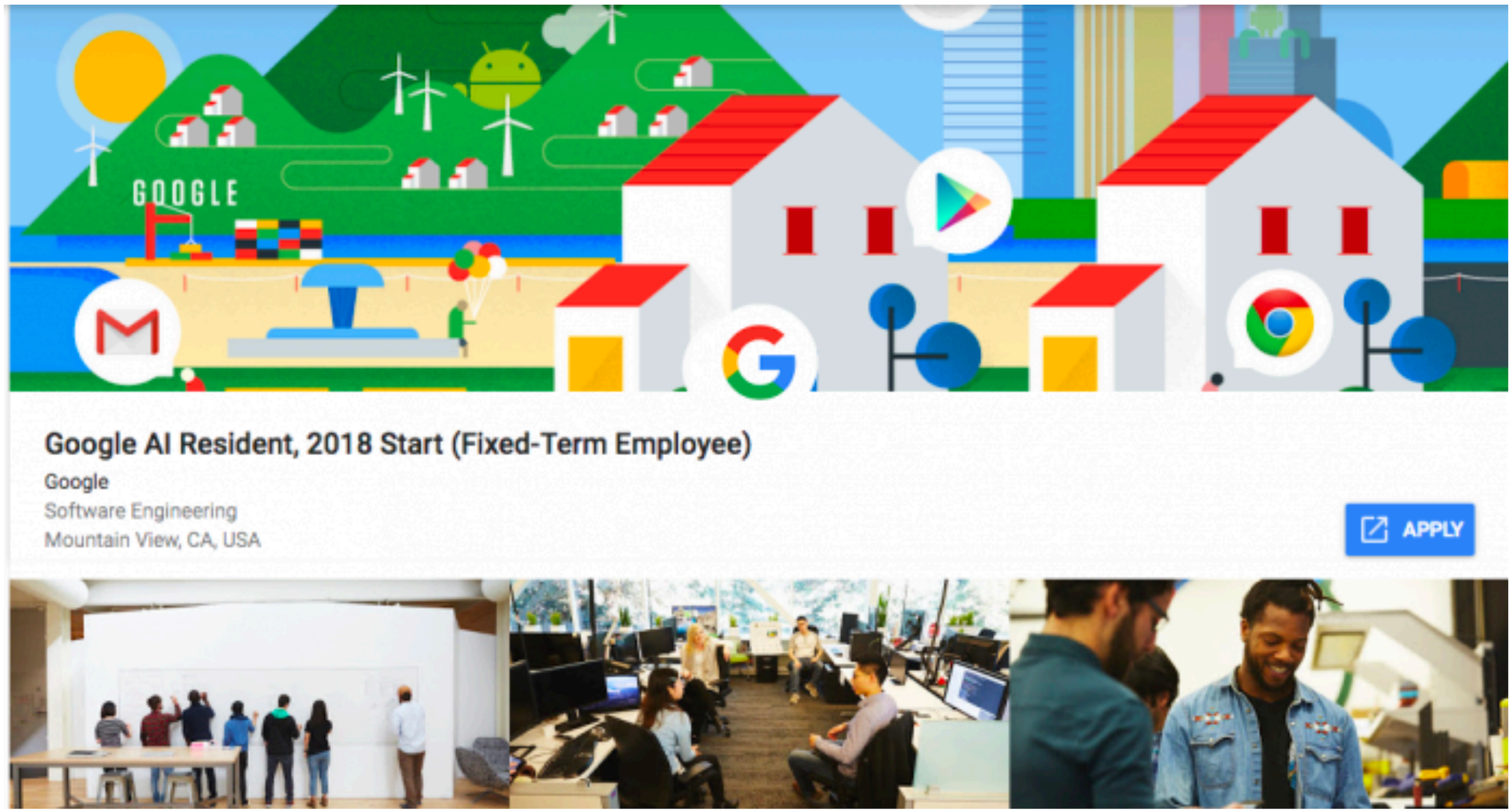
Caveats

- Slight drop in accuracy on clean examples
- Only small improvement on black-box adversarial examples

Get involved!

https://github.com/tensorflow/cleverhans

g.co/airesidency



(Goodfellow 2017)