A Baseline for Detecting Misclassified and Out-of-Distribution Examples In Neural Networks

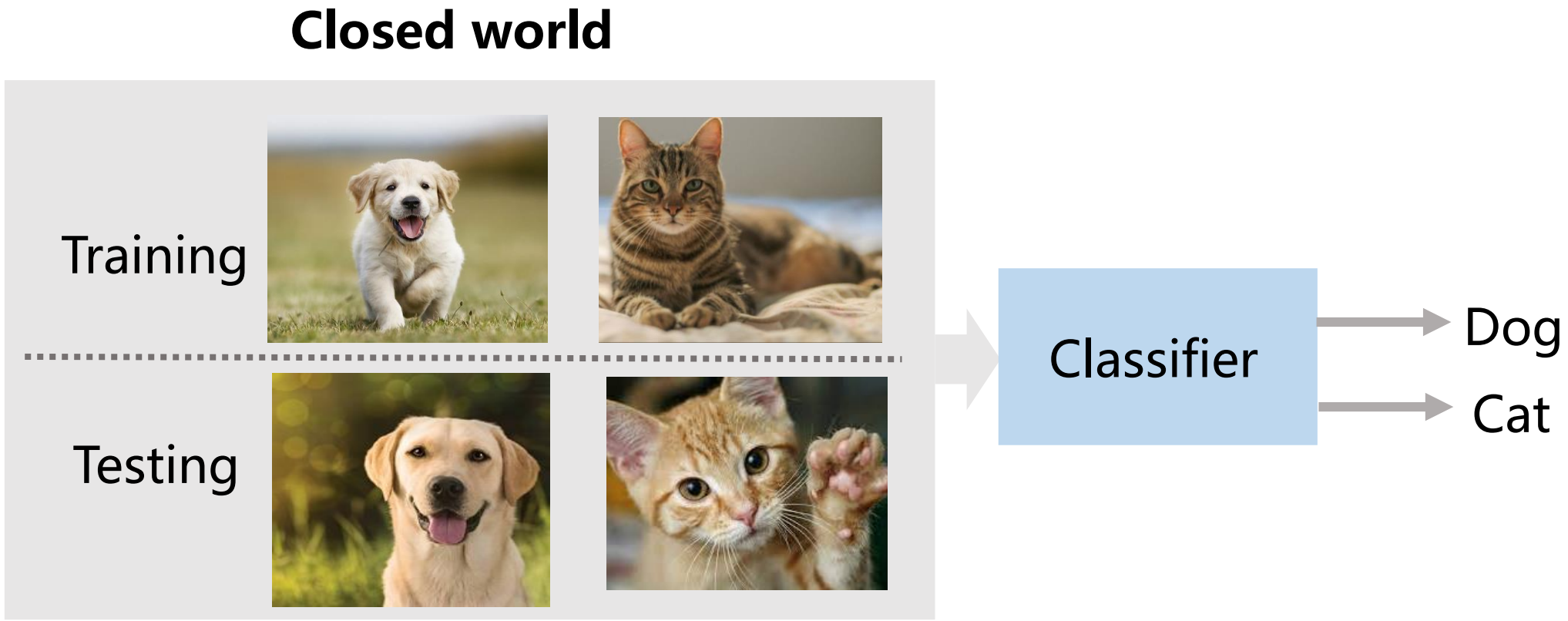
Dan Hendrycks*
University of California, Berkeley
hendrycks@berkeley.edu

Kevin Gimpel
Toyota Technological Institute at Chicago
kgimpel@ttic.edu

ICLR 2017

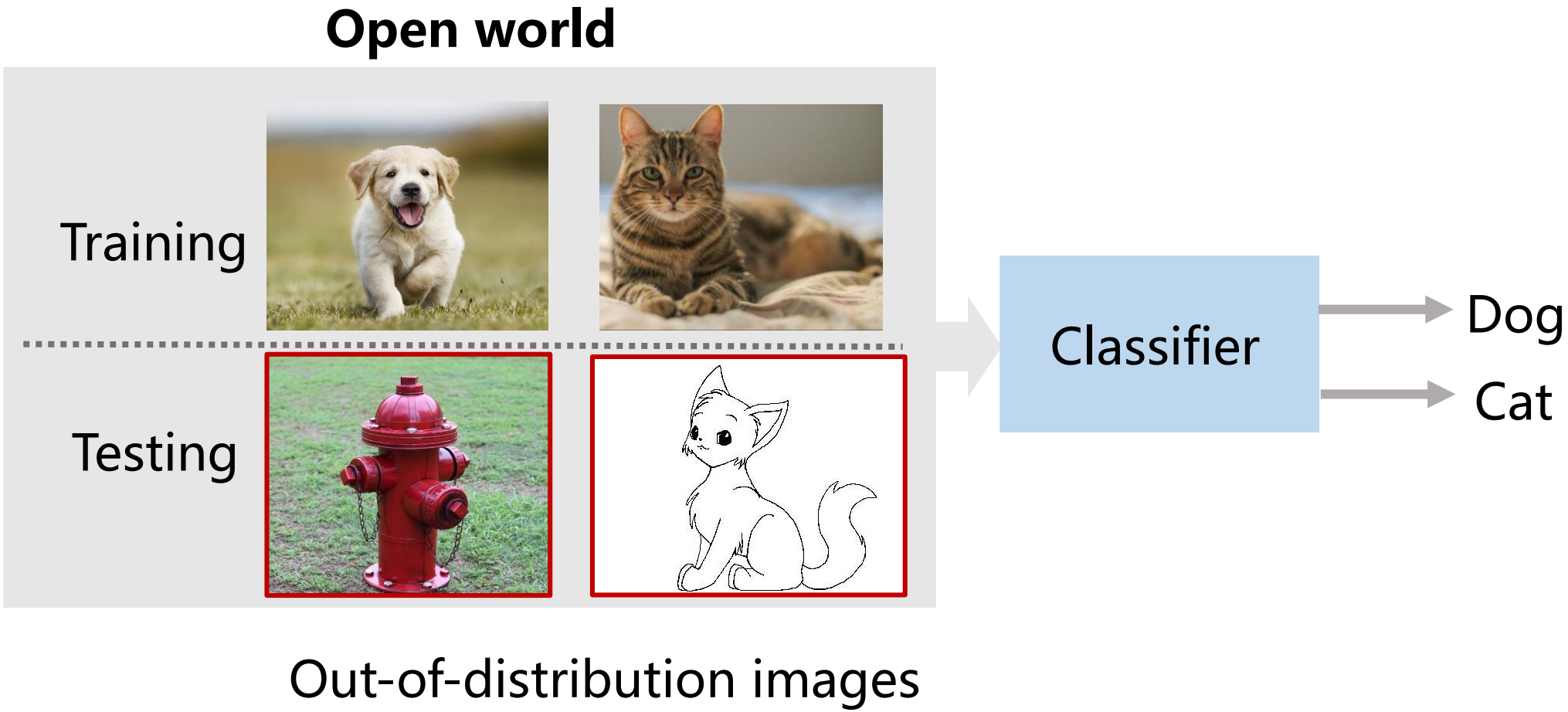
Background

- A standard classifier assumes a closed world, i.e., classes and domains are the same between training and testing



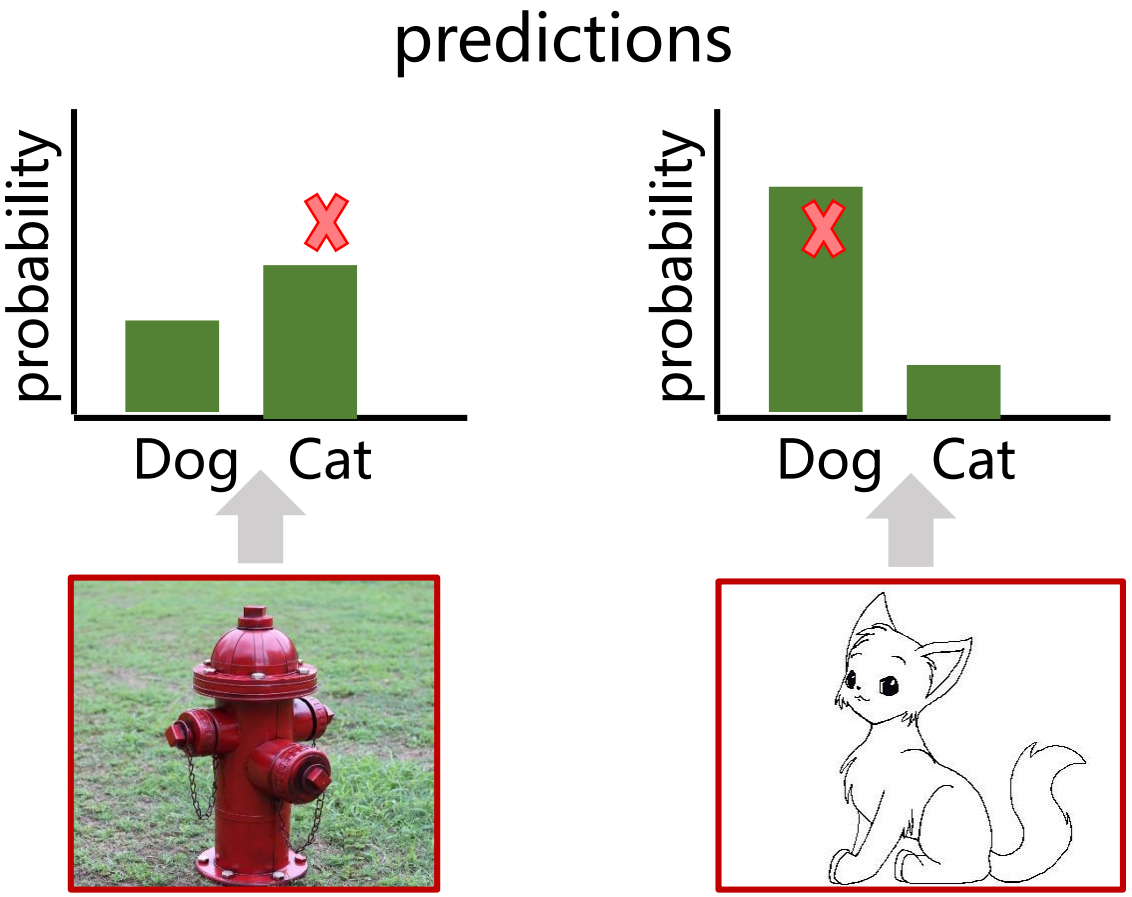
Background

- However, in an open world, the testing images may come from **unknown classes** or **unknown domains**.



Background

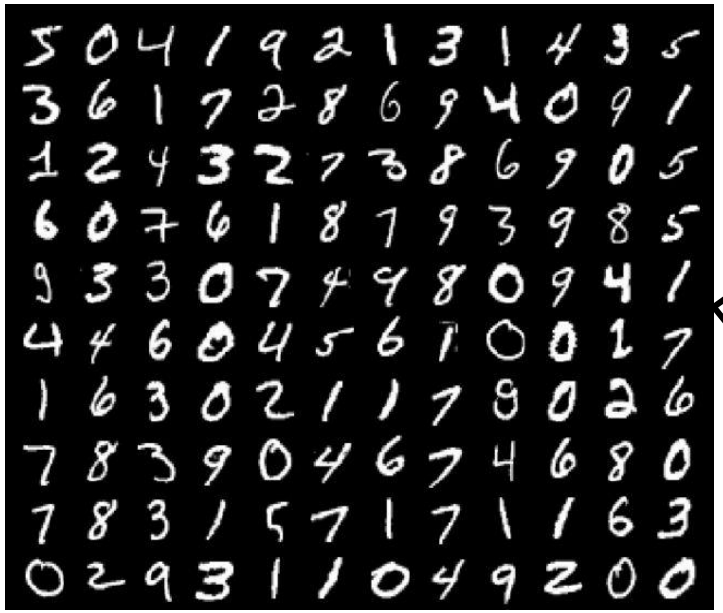
- The classifier still makes high confident predictions for out-of-distribution images while being wrong.

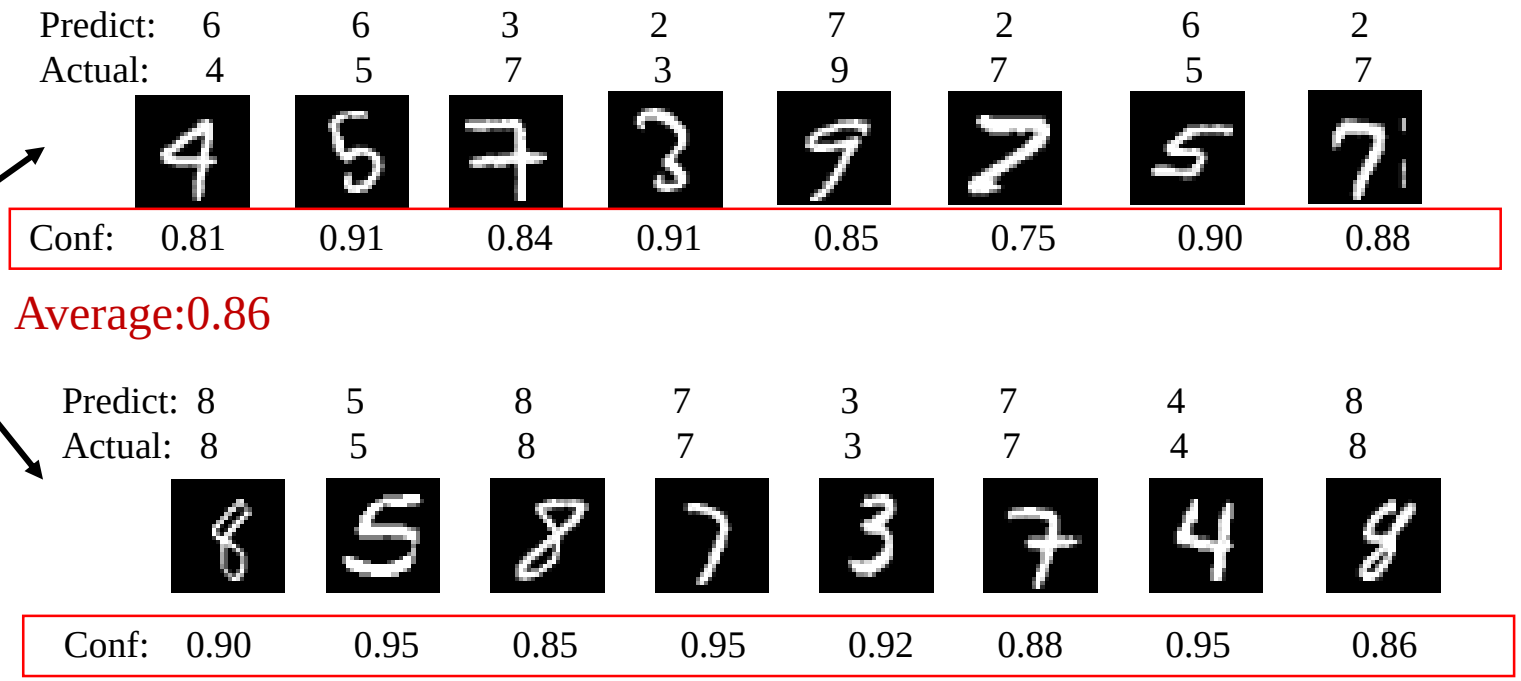


Contributions

- They show the prediction probability of incorrect and **out-of–distribution** examples tends to be **lower** than the prediction probability for correct examples.
- These prediction probability form detection baseline, and demonstrate **its efficacy** through **various experiments**.
- They propose one method which **outperforms the baseline** on some tasks.

Experiments(Misclassified?)





Dataset	AUROC /Base	AUPR Succ/Base	AUPR Err/Base	Pred. Prob Wrong(mean)	Test Set Error
MNIST	97/50	100/98	48/1.7	86	1.69
CIFAR-10	93/50	100/95	43/5	80	4.96
CIFAR-100	87/50	96/79	62/21	66	20.7

- correctly classified and incorrectly classified examples are sufficiently **distinct** and thus allow **reliable discrimination**

Experiments(out of distribution)

Select Maximum Softmax Probability and use it as out-of-distribution score

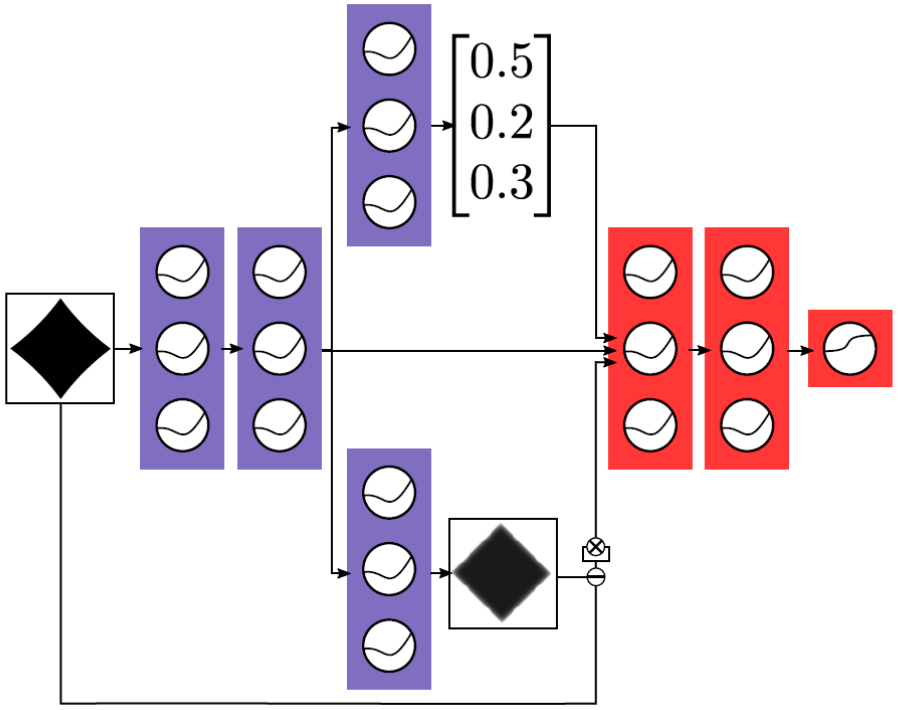
In-Distribution / Out-of-Distribution	AUROC /Base	AUPR In /Base	AUPR Out/Base	Pred. Prob (mean)
CIFAR-10/SUN	95/50	89/33	97/67	72
CIFAR-10/Gaussian	97/50	98/49	95/51	77
CIFAR-10/All	96/50	88/24	98/76	74
CIFAR-100/SUN	91/50	83/27	96/73	56
CIFAR-100/Gaussian	88/50	92/43	80/57	77
CIFAR-100/All	90/50	81/21	96/79	63
MNIST/Omniglot	96/50	97/52	96/48	86
MNIST/notMNIST	85/50	86/50	88/50	92
MNIST/CIFAR-10bw	95/50	95/50	95/50	87
MNIST/Gaussian	90/50	90/50	91/50	91
MNIST/Uniform	99/50	99/50	98/50	83
MNIST/All	91/50	76/20	98/80	89

Table 2: Distinguishing in- and out-of-distribution test set data for image classification. CIFAR-10/All is the same as CIFAR-10/(SUN, Gaussian). All values are percentages.

- softmax probabilities should not be viewed as a direct representation of confidence.
- out-of-distribution examples sufficiently differ in the prediction probabilities from in-distribution examples

Improved Method

Abnormally Module



1. Train a normal classifier and append an auxiliary decoder which reconstructs the input with in-distribution dataset.
2. Froze the blue
3. Train red layers on clean and noised training examples
4. Finally, the sigmoid output of the red layers scores how normal the input is

Experiments

In-Distribution / Out-of-Distribution	AUROC /Base Softmax	AUROC /Base AbMod	AUPR In/Base Softmax	AUPR In/Base AbMod	AUPR Out/Base Softmax	AUPR Out/Base AbMod
MNIST/Omniglot	95/50	100/50	95/52	100/52	95/48	100/48
MNIST/notMNIST	87/50	100/50	88/50	100/50	90/50	100/50
MNIST/CIFAR-10bw	98/50	100/50	98/50	100/50	98/50	100/50
MNIST/Gaussian	88/50	100/50	88/50	100/50	90/50	100/50
MNIST/Uniform	99/50	100/50	99/50	100/50	99/50	100/50
Average	93	100	94	100	94	100

Table 11: Improved detection using the abnormality module. All values are percentages.

Abnormality Module is useful to detect out-of-distribution samples

Enhancing the Reliability of Out-of-Distribution Image Detection In Neural Networks

Shiyu Liang
Coordinated Science Lab, Department of ECE
University of Illinois at Urbana-Champaign
sliang26@illinois.edu

Yixuan Li
University of Wisconsin-Madison*
sharonli@cs.wisc.edu

R. Srikant
Coordinated Science Lab, Department of ECE
University of Illinois at Urbana-Champaign
rsrikant@illinois.edu

ICLR 2018

Method

- **Temperature Scaling**

$$S_i(\boldsymbol{x};T) = \frac{\exp\left(f_i(\boldsymbol{x})/T\right)}{\sum_{j=1}^{N}\exp\left(f_j(\boldsymbol{x})/T\right)},$$

- **Input preprocessing**

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} - \varepsilon \mathrm{sign}(-\nabla_{\boldsymbol{x}} \log S_{\hat{y}}(\boldsymbol{x};T)),$$

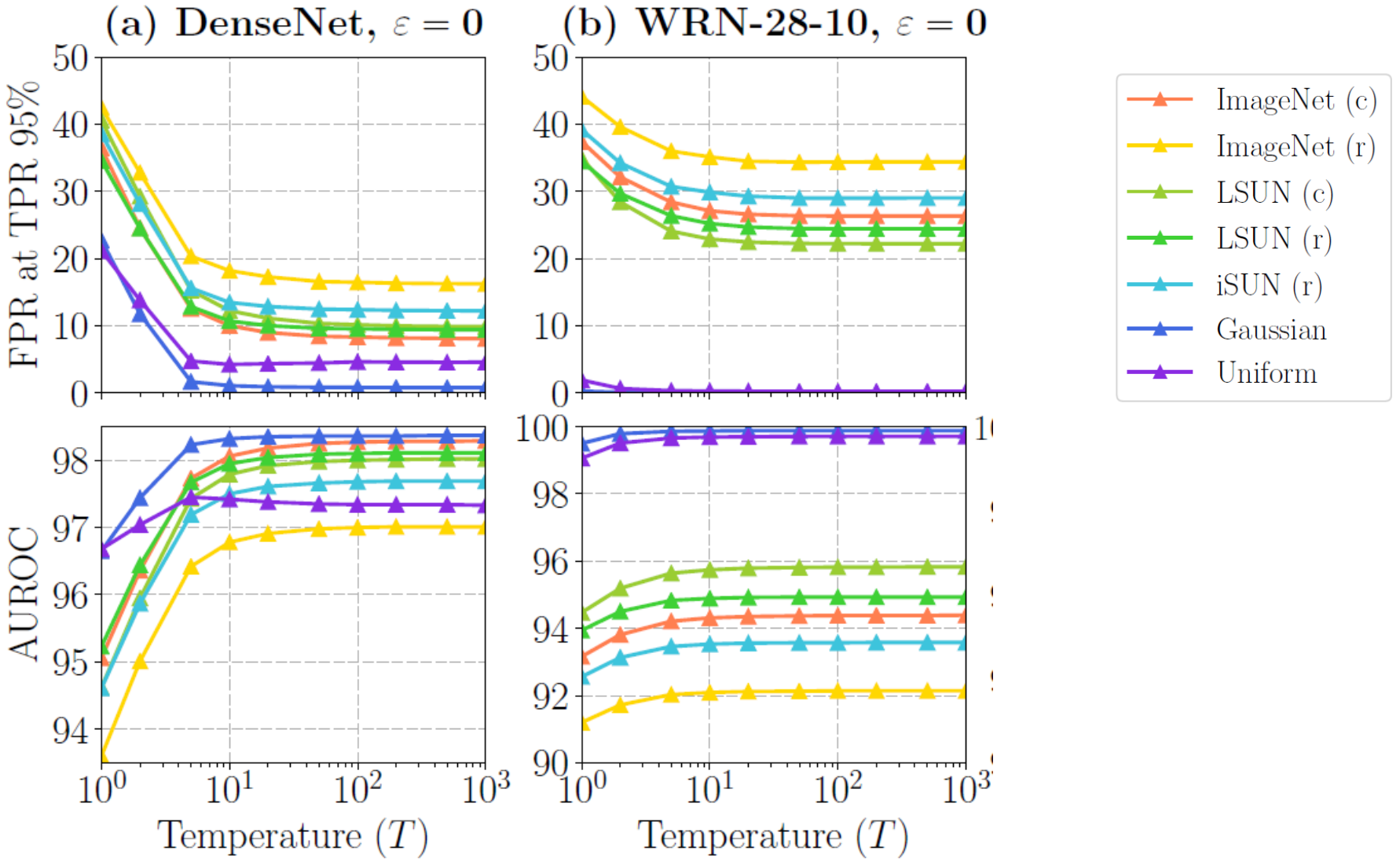
Experiment

	Out-of-distribution dataset	FPR (95% TPR) ↓	Detection Error ↓	AUROC ↑	AUPR In ↑	AUPR Out ↑
		Baseline (Hendrycks & Gimpel, 2017) / ODIN				
Dense-BC CIFAR-10	TinyImageNet (crop)	34.7/**4.3**	10.0/**4.7**	95.3/**99.1**	96.4/**99.1**	93.8/**99.1**
	TinyImageNet (resize)	40.8/**7.5**	11.5/**6.1**	94.1/**98.5**	95.1/**98.6**	92.4/**98.5**
	LSUN (crop)	39.3/**11.4**	10.2/**7.2**	94.8/**97.9**	96.0/**98.0**	93.1/**97.9**
	LSUN (resize)	33.6/**3.8**	9.8/**4.4**	95.4/**99.2**	96.4/**99.3**	94.0/**99.2**
	Uniform	23.5/**0.0**	5.3/**0.5**	96.5/**99.0**	97.8/**100.0**	93.0/**99.0**
	Gaussian	12.3/**0.0**	4.7/**0.2**	97.5/**100.0**	98.3/**100.0**	95.9/**100.0**
Dense-BC CIFAR-100	TinyImageNet (crop)	67.8/**26.9**	36.4/**12.9**	83.0/**94.5**	85.3/**94.7**	80.8/**94.5**
	TinyImageNet (resize)	82.2/**57.0**	43.6/**22.7**	70.4/**85.5**	71.4/**86.0**	68.6/**84.8**
	LSUN (crop)	69.4/**18.6**	37.2/**9.7**	83.7/**96.6**	86.2/**96.8**	80.9/**96.5**
	LSUN (resize)	83.3/**58.0**	44.1/**22.3**	70.6/**86.0**	72.5/**87.1**	68.0/**84.8**
	Uniform	100.0/**100.0**	35.86/**17.9**	43.1/**99.5**	63.2/**87.5**	41.9/**65.1**
	Gaussian	100.0/**100.0**	41.2/**38.0**	30.6/**40.5**	53.4/**60.5**	37.6/**40.9**

Table 2: Distinguishing in- and out-of-distribution test set data for image classification. All values are percentages. ↑ indicates larger value is better, and ↓ indicates lower value is better. We use $T = 1000$ for all experiments. The noise magnitude ε was selected on a **separate validation dataset**, which is different from the out-of-distribution test sets. On CIFAR-10 pretrained model, we use $\varepsilon = 0.0014$ for all OOD test datasets; and $\varepsilon = 0.002$ for CIFAR-100 pretrained model.

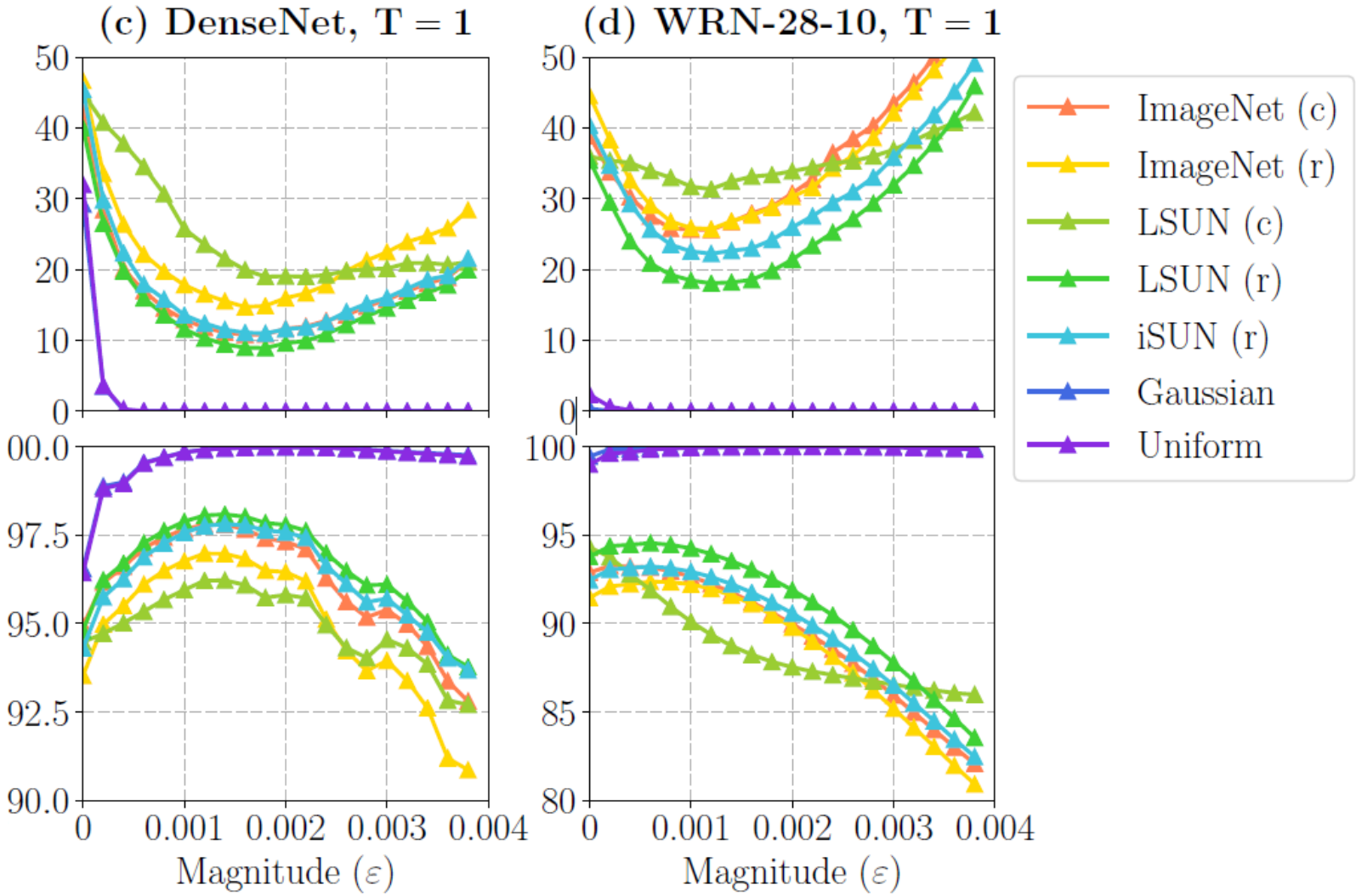
Experiment

- **Effects of temperatures T when ε=0**



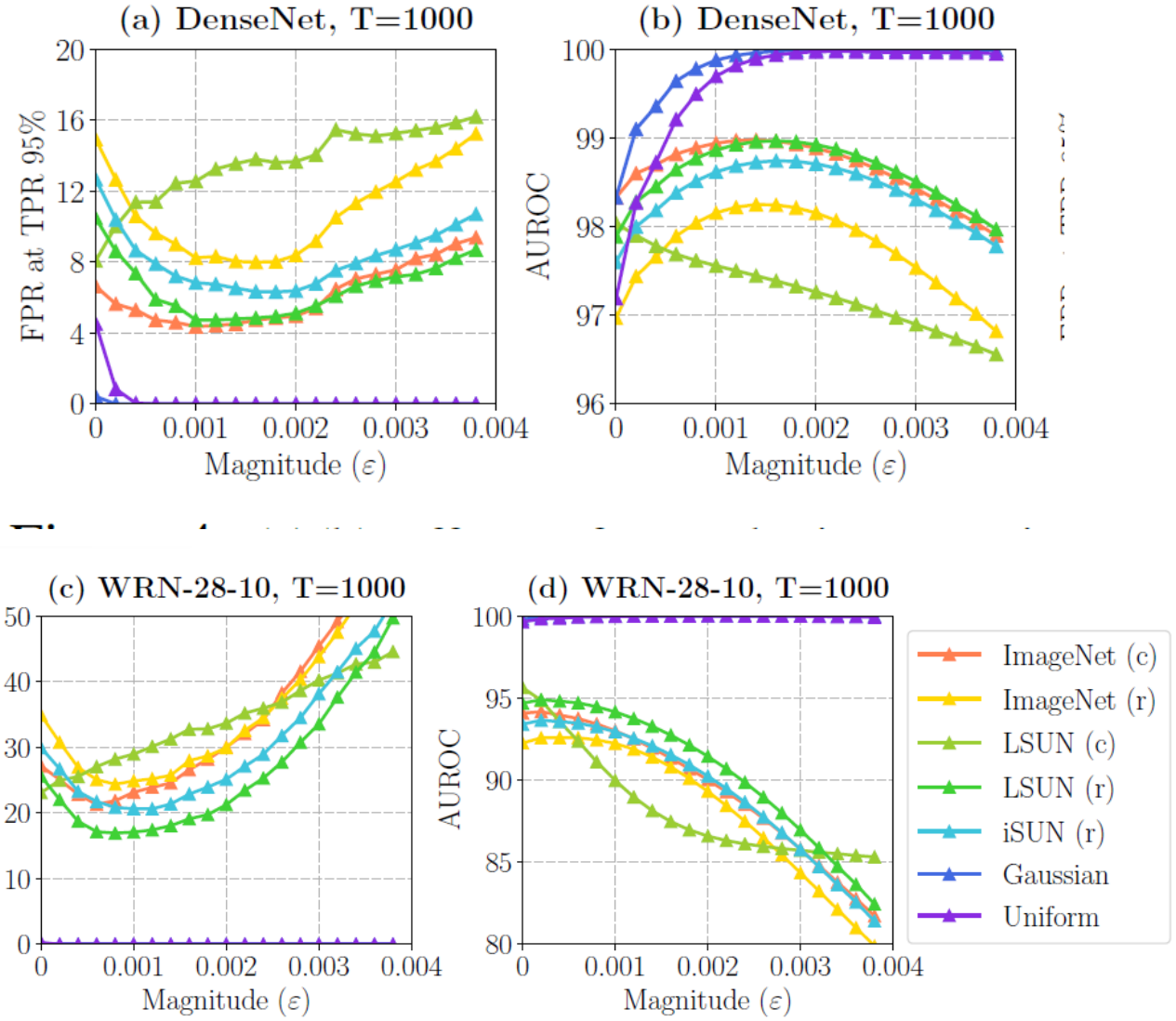
Experiment

➢ **Effects of perturbation magnitude ε when T = 1**



Experiment

- **Effects of perturbation magnitude ε when T is large**



Discussion

- **Analysis on temperature scaling**

$$
\begin{aligned}
S_{\hat{y}}(\boldsymbol{x};T) &= \frac{\exp\left(f_{\hat{y}}(\boldsymbol{x})/T\right)}{\sum_{i=1}^{N}\exp(f_i(\boldsymbol{x})/T)} \\
&= \frac{1}{\sum_{i=1}^{N}\exp\left(\frac{f_i(\boldsymbol{x})-f_{\hat{y}}(\boldsymbol{x})}{T}\right)} \\
&= \frac{1}{\sum_{i=1}^{N}\left[1+\frac{f_i(\boldsymbol{x})-f_{\hat{y}}(\boldsymbol{x})}{T}+\frac{1}{2!}\frac{(f_i(\boldsymbol{x})-f_{\hat{y}}(\boldsymbol{x}))^2}{T^2}+o\left(\frac{1}{T^2}\right)\right]} \\
&\approx \frac{1}{N-\frac{1}{T}\sum_{i=1}^{N}[f_{\hat{y}}(\boldsymbol{x})-f_i(\boldsymbol{x})]+\frac{1}{2T^2}\sum_{i=1}^{N}[f_i(\boldsymbol{x})-f_{\hat{y}}(\boldsymbol{x})]^2}
\end{aligned}
$$

by Taylor expansion

Define: $U_1(\boldsymbol{x}) = \frac{1}{N-1}\sum_{i\neq\hat{y}}[f_{\hat{y}}(\boldsymbol{x})-f_i(\boldsymbol{x})]$ $U_2(\boldsymbol{x}) = \frac{1}{N-1}\sum_{i\neq\hat{y}}[f_{\hat{y}}(\boldsymbol{x})-f_i(\boldsymbol{x})]^2.$

$$S \propto (U_1 - U_2/2T)/T.$$

Discussion

➢ Interpretations of U1 and U2

For simplicity of the notations, let $\Delta_i = f_{\hat{y}} - f_i$ and thus $\Delta = \{\Delta_i\}_{i \neq \hat{y}}$. Besides, let $\bar{\Delta}$ denote the mean of the set Δ. Therefore,

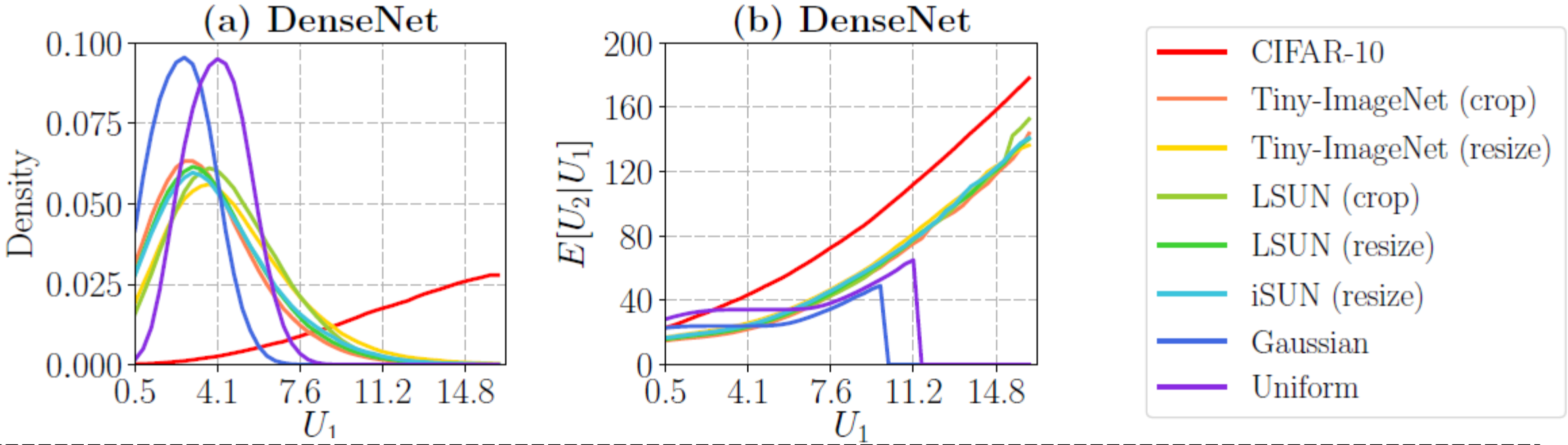
$$\bar{\Delta} = \frac{1}{N-1} \sum_{i \neq \hat{y}} \Delta_i = \frac{1}{N-1} \sum_{i \neq \hat{y}} [f_{\hat{y}} - f_i] = U_1.$$

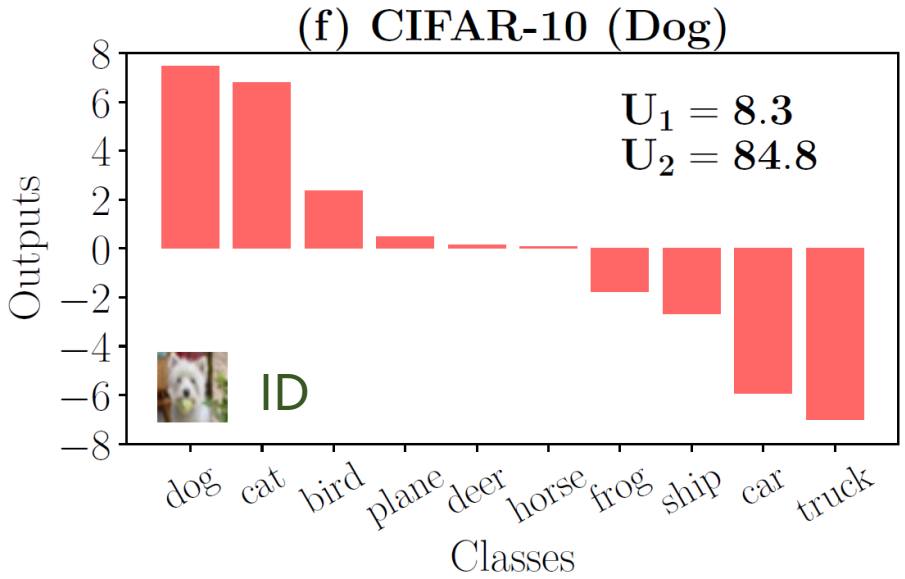
$\Rightarrow U_1 = \text{Mean}(\Delta).$

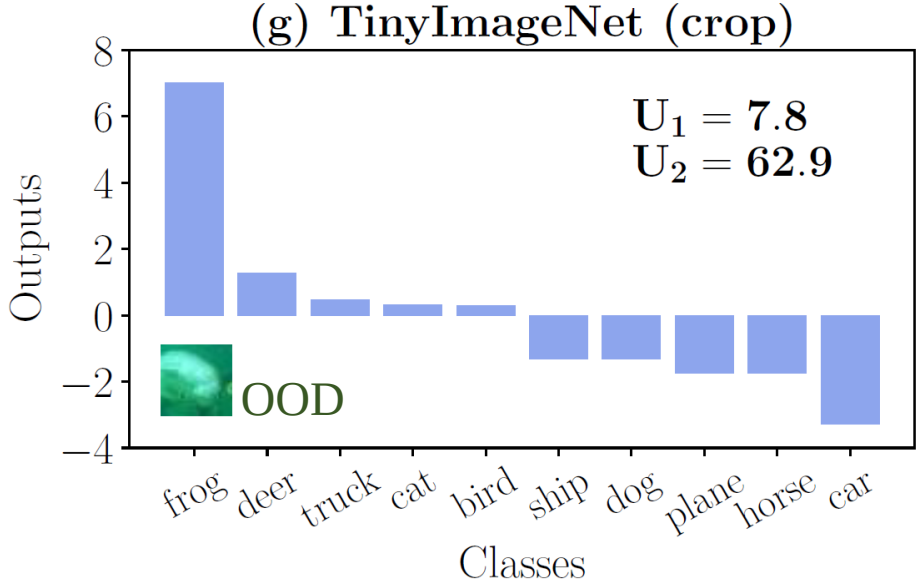
$$\Rightarrow U_2 = \frac{1}{N-1} \sum_{i \neq \hat{y}} [f_{\hat{y}} - f_i]^2 = \overbrace{\frac{1}{N-1} \sum_{i \neq \hat{y}} [\Delta_i - \bar{\Delta}]^2}^{\text{Variance}^2(\Delta)} + \overbrace{\bar{\Delta}^2}^{\text{Mean}^2(\Delta)}.$$

- U1 measures the extent to which the largest unnormalized output of the neural network deviates from the remaining outputs.
- U2 measures the extent to which the remaining smaller outputs deviate from each other.

Discussion


(a) DenseNet
Density
0.100
0.075
0.050
0.025
0.000
0.5
4.1
7.6
11.2
14.8
U_1
(b) DenseNet
$E[U_2|U_1]$
200
160
120
80
40
0
0.5
4.1
7.6
11.2
14.8
U_1
CIFAR-10
Tiny-ImageNet (crop)
Tiny-ImageNet (resize)
LSUN (crop)
LSUN (resize)
iSUN (resize)
Gaussian
Uniform


(f) CIFAR-10 (Dog)
Outputs
8
6
4
2
0
−2
−4
−6
−8
$U_1 = 8.3$
$U_2 = 84.8$
ID
dog
cat
bird
plane
deer
horse
frog
ship
car
truck
Classes


(g) TinyImageNet (crop)
Outputs
8
6
4
2
0
−2
−4
$U_1 = 7.8$
$U_2 = 62.9$
OOD
frog
deer
truck
cat
bird
ship
dog
plane
horse
car
Classes

Discussion

- **The effects of T**

$$S \propto (U1 - U2/T)/T$$

$S \propto U1$: in-distribution images produce larger softmax scores than out-of distribution images

$S \propto -\frac{U2}{T}$: Opposite effect

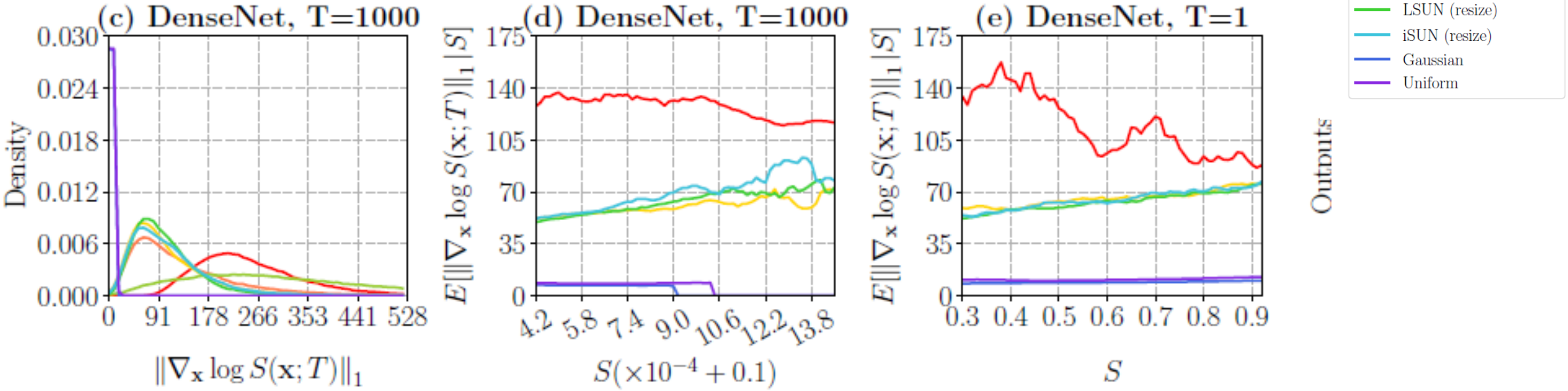
$T \uparrow \longrightarrow U2/T \downarrow$: compensate the negative impacts of U2/T on the detection performance

Discussion

➢ Analysis on input preprocessing

- first order Taylor expansion of the log-softmax function for the perturbed image $\hat{x}$,

$$\log S_{\hat{y}}(\tilde{\boldsymbol{x}}; T) = \log S_{\hat{y}}(\boldsymbol{x}; T) + \varepsilon \left\| \nabla_{\boldsymbol{x}} \log S_{\hat{y}}(\boldsymbol{x}; T) \right\|_1 + o(\varepsilon)$$



- in-distribution results in a much larger value on the norm of softmax gradient than that of out-of-distribution, in- and out-of-distribution images are more separable from each other after input preprocessing

Thanks