e-Lab

Robust Convolutional Neural Networks under Adversarial Noise

Jonghoon Jin, Aysegul Dundar and Eugenio Culurciello

PURDUE UNIVERSITY

Abstract

We propose a new feedforward CNN that is **robust to adversarial noise**. With uncertainty noise added to input, all operators in CNNs are modified to benefit from the noise. The model is parameterized by mean and variance per pixel and **successfully applied to deep architecture** like ResNet-101.

Adversarial example

- generated to fool CNNs on purpose

Bee

(original)

+

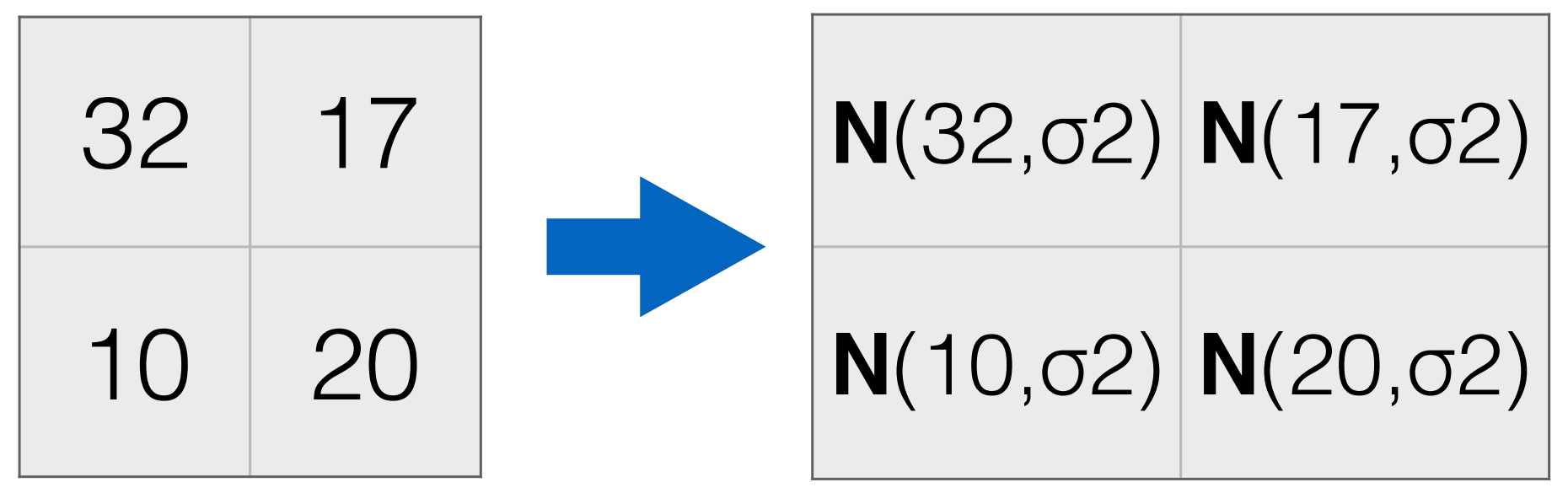
(adversarial noise)

=

Leaf beetle

(adversarial example)

Input with uncertainty



32	17
10	20

N(32,σ2)	**N**(17,σ2)
N(10,σ2)	**N**(20,σ2)

(2x2 example image)

Convolution

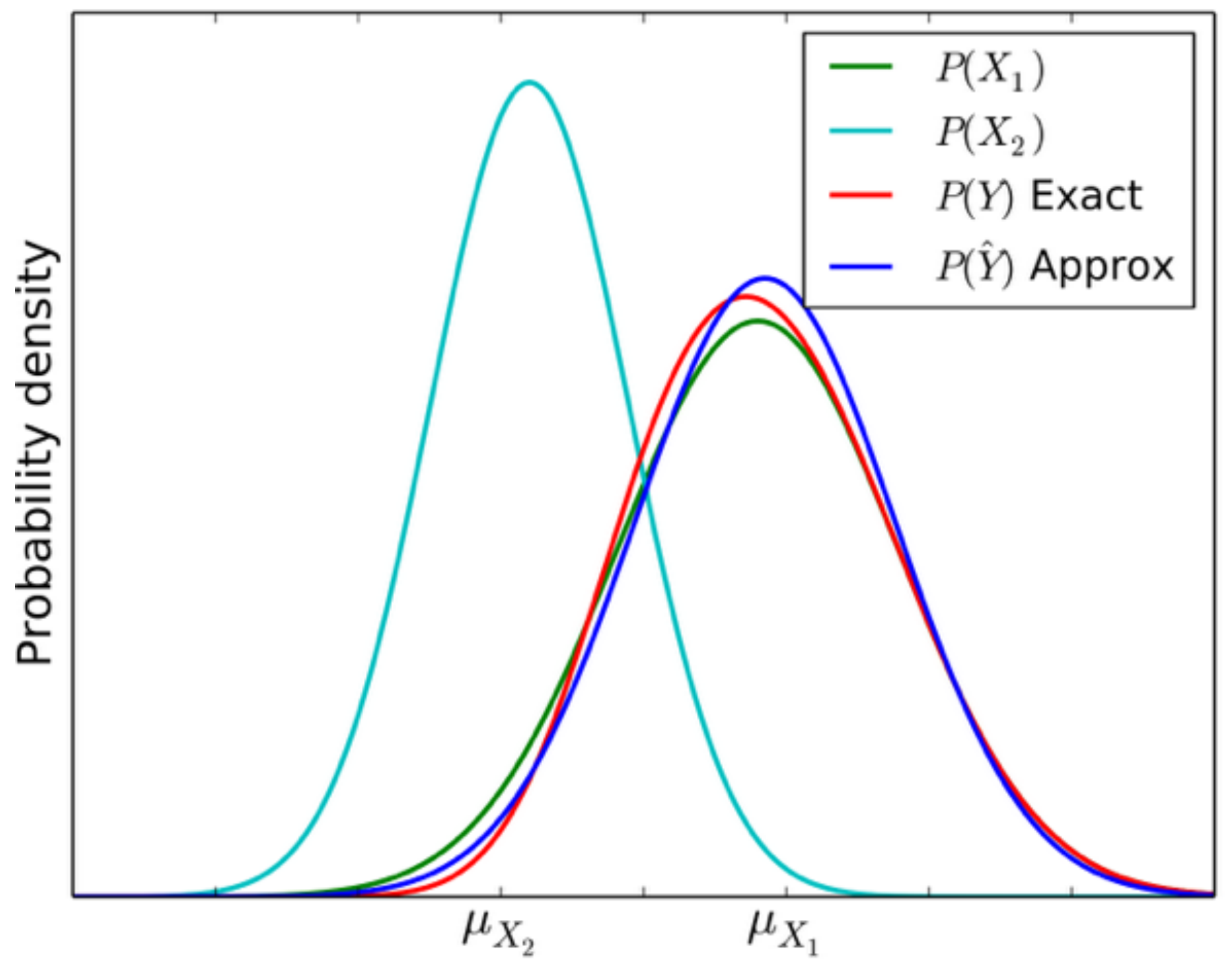
$$E\left[Y\right] = \sum \omega E\left[X\right] + b$$

$$Var\left[Y\right] = \sum \omega^2 Var\left[X\right]$$

(Gaussian output)

Max-pooling

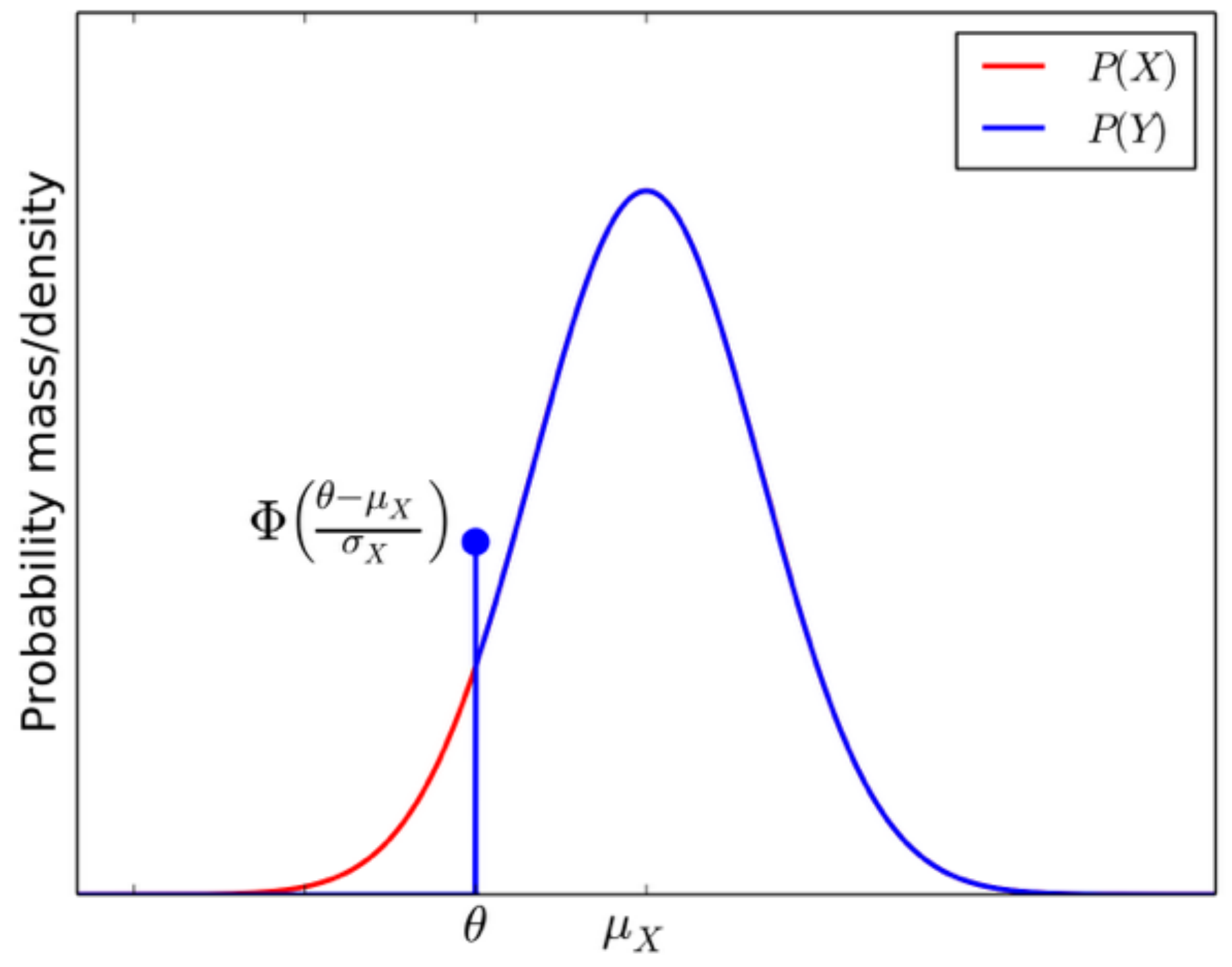
$$Y = max(X_1, X_2)$$



(non-Gaussian output)

ReLU

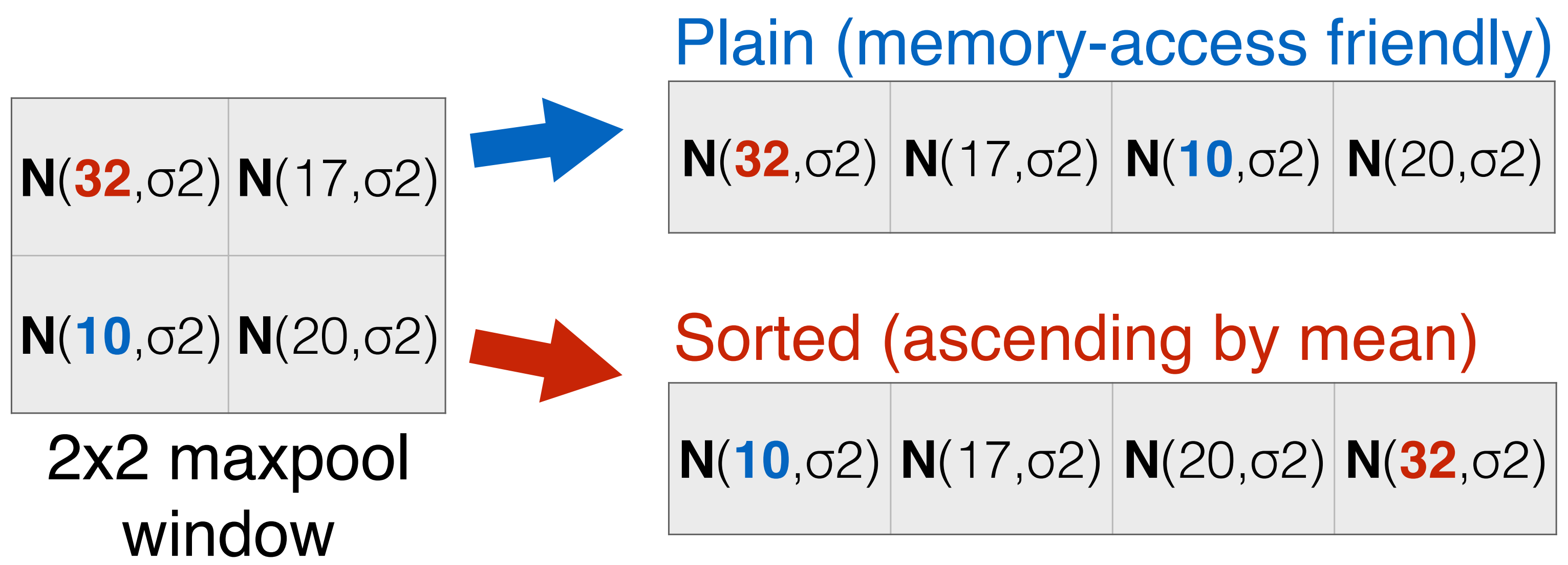
$$Y = max(X, \theta)$$



(non-Gaussian output)

Gaussian approximation for max-pooling (and ReLU)

- low error when max-pooled in ascending order by mean



N(**32**,σ2)	**N**(17,σ2)
N(**10**,σ2)	**N**(20,σ2)

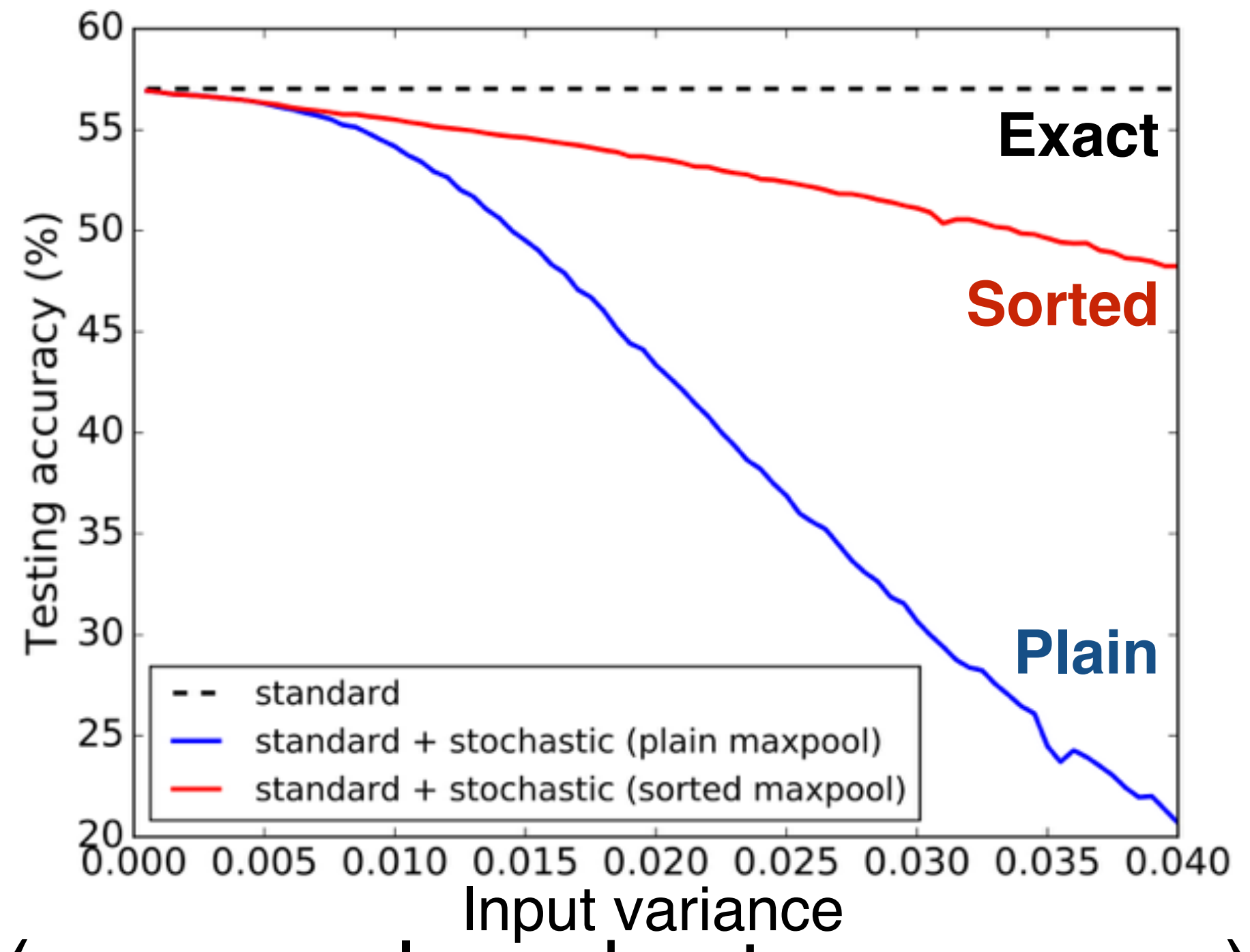
2x2 maxpool window

Plain (memory-access friendly)

N(**32**,σ2)	**N**(17,σ2)	**N**(**10**,σ2)	**N**(20,σ2)

Sorted (ascending by mean)

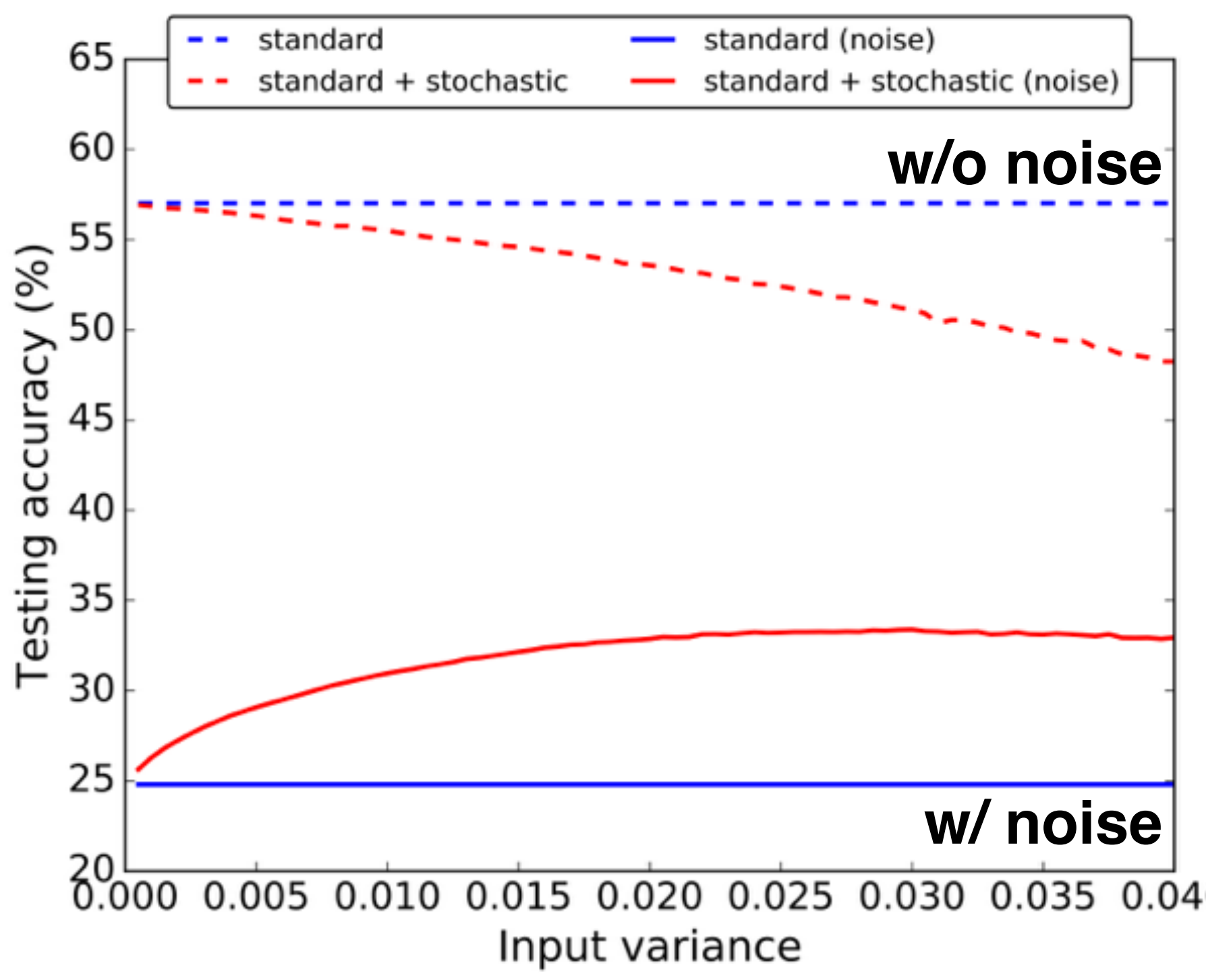
N(**10**,σ2)	**N**(17,σ2)	**N**(20,σ2)	**N**(**32**,σ2)



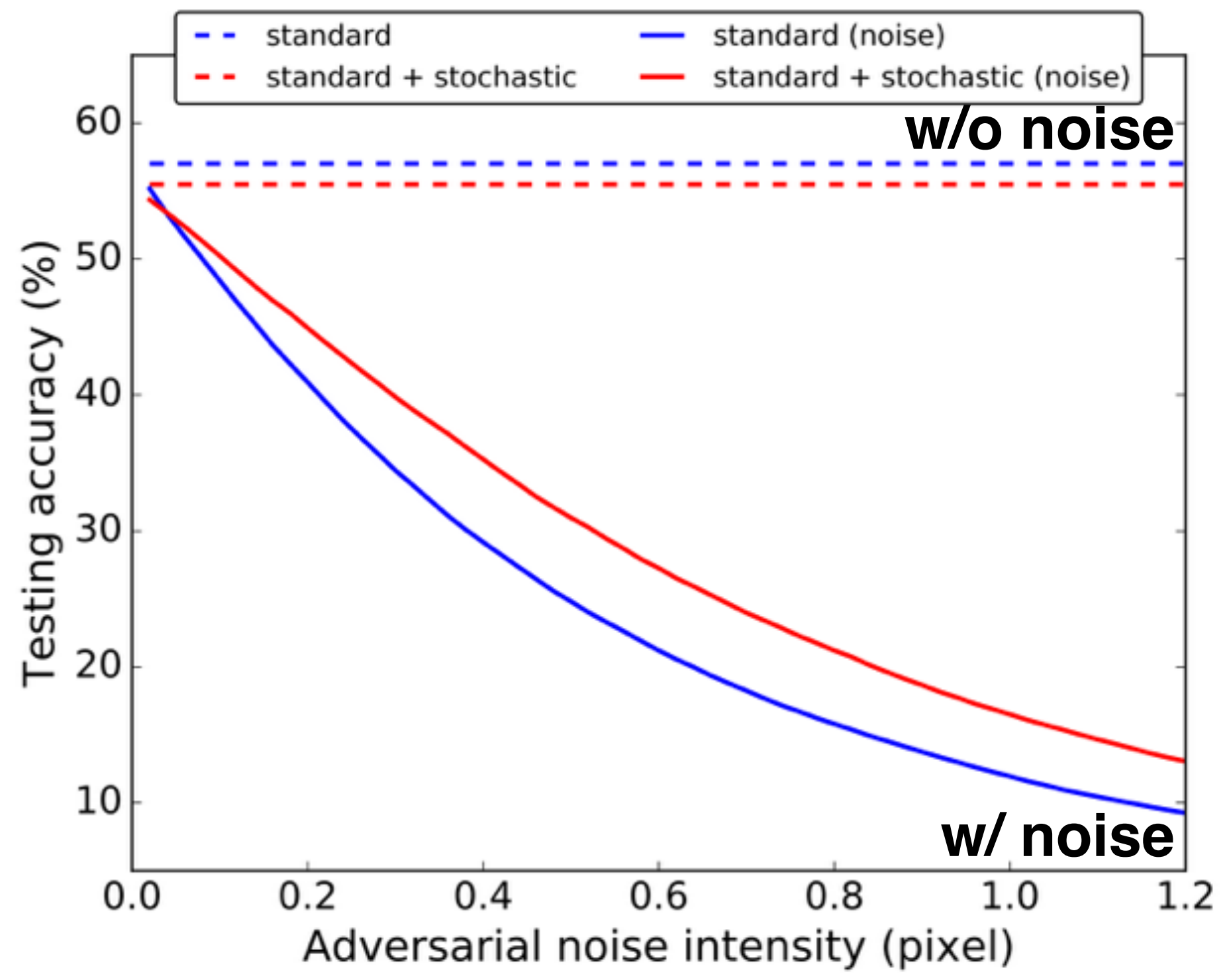
(accuracy loss due to approx err)

Parameter tuning (σ)

- trade-off (small classification loss <—> large gain with noise)



(fixed noise intensity)



(fixed input variance)

Demo

Code

jhjin/stochastic-cnn

Classification accuracy under noise

(higher is better)

Dataset	CIFAR-10		ImageNet			
Model	NIN		AlexNet			ResNet-101
Adversarial noise intensity [px]	0	0.5	0	0.01	0.5	1
Standard training	90.1	72.3	57.0	56.1	24.8	17.50
Standard training + stochastic (this work)	88.9	**78.1**	57.0	**56.2**	**33.4**	**39.24**
LWA + BN (Huang et al. 2016)	89.0	82.3	—	—	—	—
Adversarial training (Goodfellow et al. 2015)	88.7	82.1	43.0	42.9	∗	∗
Adversarial training + stochastic (this work)	88.7	**82.9**	43.0	42.9	∗	∗

(∗ : failed to converge)