Improving Adversarial Robustness in Weight-quantized Neural Networks

Chang Song[1], Elias Fallon[2], Hai Li[1]

[1]Duke University, [2]Cadence Design Systems, Inc.

AAAI 2021 - RSEML Workshop

Background - Overview

- With more layers and more complex structures, modern neural networks can achieve near or even beyond human-level accuracy in solving classification problems.
- Security industry has also adopted deep learning techniques in many fields, including surveillance, authentication, facial recognition, etc.
- However, a recent research[1] discovered that neural networks are vulnerable to some delibrately-perturbed examples, though the perturbation is imperceptible to humans. These examples are called *adversarial examples.*

[1] Szegedy et al., Intriguing Properties of Neural Networks, arXiv, 2013.

Background – Decision Space

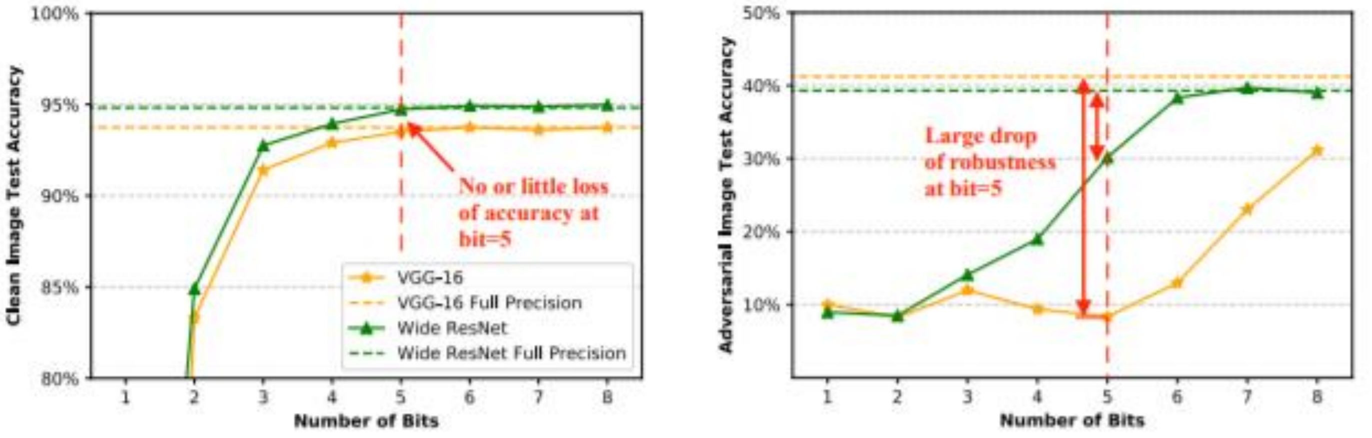
- Decision space: a vector space where all input samples lie in.
- Decision boundaries: hyper-surfaces that partition the decision space.
- In classification problems, we can define decision boundaries as sets of data points with tied highest score for multiple classes. Or, when a sample moves in one direction until being misclassified, that point will be on a decision boundary.
- In fact, decision boundaries are vague and data points near decision boundaries may not have any physical meaning.
- Adversarial examples are carefully sought points that cross boundaries with minimum effort.

Background – Nonlinearity and Robustness

- Model linearity leads to high success rate of adversarial attacks.
- Error amplification effect: Feature space distances between normal samples and adversarial examples increase layer by layer.
- Three ways to introduce nonlinearity:
  - Activation: But sigmoid and ReLU are mainly used in linear regions;
  - Pooling (max pooling, ~~average pooling~~);
  - Weight mapping: hard to be integrated in training, easy to map after training.

Related Works

- Quantized neural network are more vulnerable to adversarial attack[1].



(a) Quantization preserves the accuracy till 4-5 bits on clean image.

(b) Quantization no longer preserves the accuracy under adversarial attack (same legend as left).

- Use the Lipschitz constant to upper-bound the model's sensitivity to adversarial examples[2].
- Error amplification effect: smaller Lipschitz constant could control the adversarial perturbation not to be amplified.

[1] Lin et al., Defensive quantization: When efficiency meets robustness, ICLR, 2019.
[2] Cisse et al., Parseval networks: improving robustness to adversarial examples, ICML, 2017.

Motivation

- The difference in the output of one specific layer:

$$\delta = \underbrace{(W + \Delta W)}_{Quant.Weight} \cdot \underbrace{(x + \Delta x)}_{Adv.Input} - Wx = \underbrace{W\Delta x}_{Adv.Loss} + \underbrace{\Delta W x}_{Quant.Loss} + \Delta W \Delta x$$

- Adversarial loss: can be measured by the accuracy drop
- Quantization loss: depends on both weights and inputs, we need an input-independent criterion to evaluate the quantization process.
- The (quantization) error amplification effect[1]: small residual perturbation is amplified to a large magnitude in top layers of a model and finally leads to a wrong prediction.
- The Lipschitz Constant of ΔW:

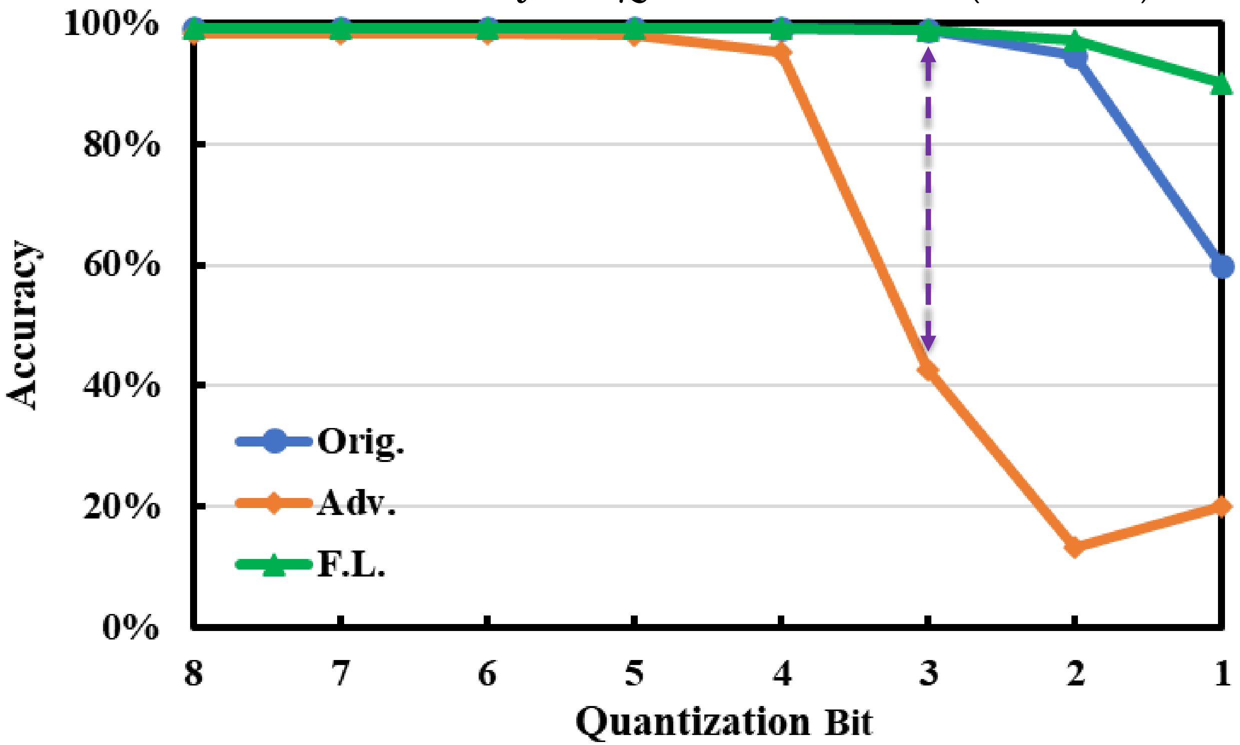
$$\|\Delta W\|_p = \sup_{z:\, \|z\|_p = 1} \|\Delta W z\|_p$$

[1] Liao, et al. Defense against adversarial attacks using high-level representation guided denoiser. CVPR, 2018.

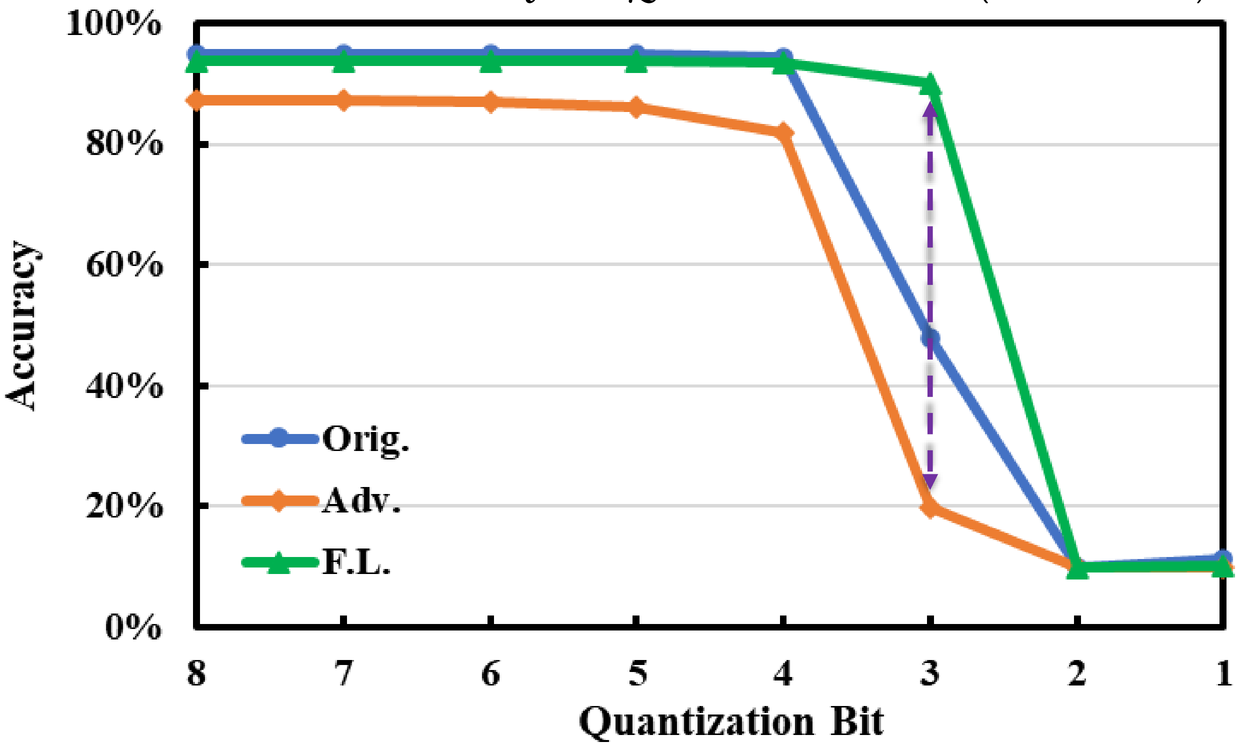
Motivation

- Adversarial training is more vulnerable to quantization.
- Here F.L. is a boundary-based training method[1].

Clean Accuracy vs. Quantization bits (MNIST)



Clean Accuracy vs. Quantization bits (CIFAR-10)

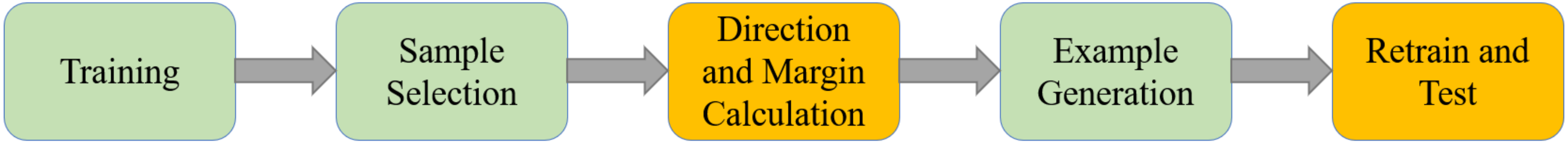


[1] Song et al., Feedback Learning for Improving the Robustness of Neural Networks, ICMLA, 2019.

Motivation (cont.)

- Larger margin between samples and decision boundaries is needed for tolerating the quantization process. Boundary-based training (F.L.) gives more (margin) tolerance to quantization loss.
- Problems with Adversarial training (AdvT):
- AdvT has worse performance against white-box attacks than black-box attacks (same attack strength), as white-box attacks are more fatal.
  - But relatively speaking, WB are easier to defend than BB.
  - BB need larger strength to downgrade accuracy (transferability matters).
- AdvT doesn't cooperate well with other techniques (quantization-aware training or regularization) w/ or w/o quantization.
  - The objective functions/goals are different or even in opposite directions.

Methodology – Feedback Learning[1]



- Classes are categorized into three robustness levels:

- **High-level**: top 20% of all classes, 20 samples are selected for each class.
- **Low-level**: bottom 50% of all classes, 150 samples are selected for each class.
- **Medium-level**: all remaining classes, 100 samples are selected for each class.

- Generated example: direction with top-40 minimum margins, 1.5x-2.0x margins to cross boundaries.
- All parameters here are empirical.

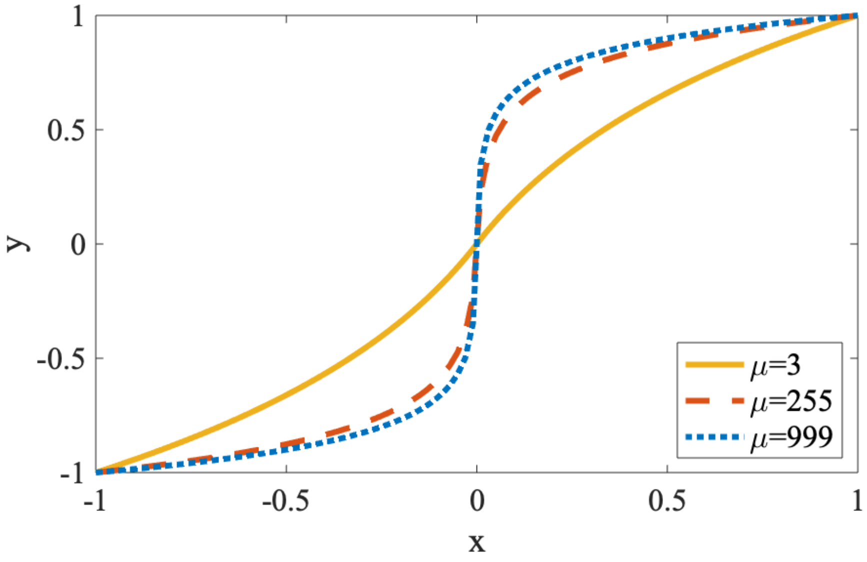
[1] Song et al., Feedback Learning for Improving the Robustness of Neural Networks, ICMLA, 2019.

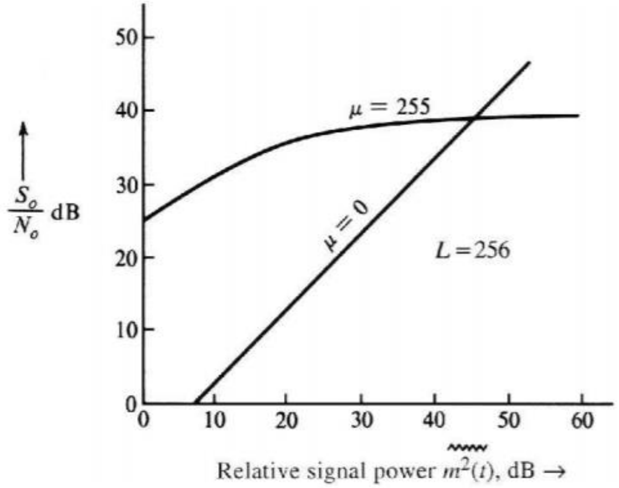
Methodology – Nonlinear Mapping

- μ-law algorithm: adopted from wireless communication, mainly to save bandwidth and improve SNR (signal-to-noise ratio).

$$F(x) = sgn(x)\frac{\ln(1+\mu|x|)}{\ln(1+\mu)}, -1 \leq x \leq 1$$

- Here, we can regard adversarial perturbations as noises, higher SNR means original components (signals) are more significant.





https://web.stanford.edu/class/ee179/lectures/notes12.pdf

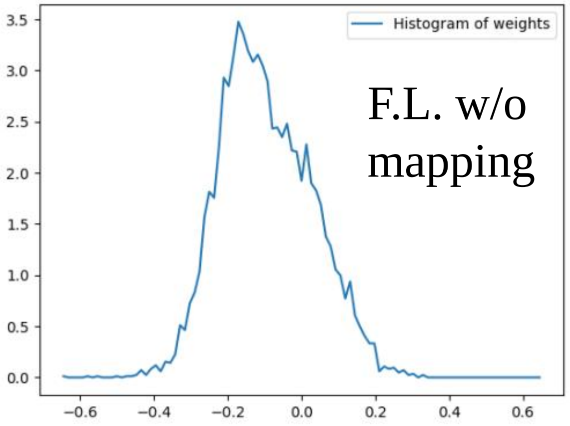
Methodology – Nonlinear Mapping (cont.)

- Procedures of combining nonlinear mapping with training:
  1) Training with other defensive techniques
  2) Post-training weight nonlinear mapping
- Which layers to map? Increasing nonlinearity vs. accuracy loss.
  - Mapping more layers means higher nonlinearity level, but…
  - Mapping feature extractors (convolutional layers) introduces more accuracy loss than mapping classifiers (FC layers)[1].
  - Adversarial perturbations have larger impact on models' decision-making than feature extraction.

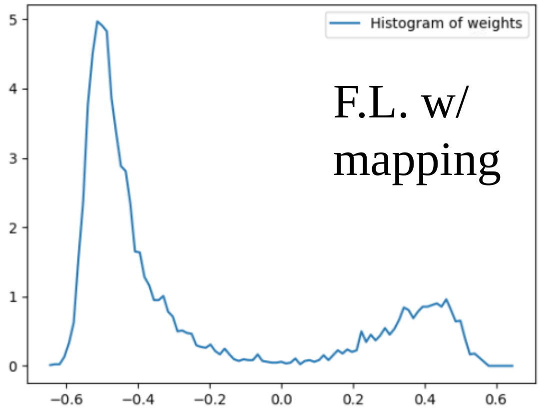
[1] Inkawhich et al., Feature Space Perturbations Yield More Transferable Adversarial Examples, CVPR, 2019.

Experimental Results

- Datasets: MNIST (4-layer CNN) and CIFAR-10 (wide ResNet-32).
- Models: Orig., Adv. (adversarially-trained model), F.L. (feedback learning).
- Attacks (adversarial and non-adversarial): clean image, CW-L2, FGSM, PGD, BIM, Momentum IM, normal noise, uniform noise; white-box and black-box attacks.
- 3-bit quantization, post-training weight quantization only.
- Nonlinear mapping only the last few layers.



CIFAR-10, last layer



Experimental Results – Accuracy on MNIST

- White-box accuracy: ~20% improvement on F.L. model, no improvement on Orig. and Adv. models.
  - F.L. model has better tolerance to error introduced by quantization and nonlinear mapping.
- Black-box accuracy: same robustness after mapping.

Table 1: The accuracy of white-box attacks on MNIST models.

Models	Clean	CW-L2	FGSM (w)	FGSM (s)	PGD	BIM	MIM
Orig.	99.17%	39.40%	73.53%	7.67%	4.38%	5.68%	6.77%
Orig. (Q)	98.97%	36.98%	68.70%	7.40%	2.63%	3.53%	4.27%
Adv.	98.40%	94.51%	98.01%	96.24%	97.77%	97.41%	97.32%
Adv. (Q)	42.69%	25.56%	37.28%	32.28%	33.78%	31.44%	30.72%
F.L.	99.17%	51.60%	89.69%	39.43%	39.92%	41.42%	43.25%
F.L. (Q)	**98.99%**	**49.49%**	87.93%	38.36%	35.35%	36.48%	38.33%
Orig.+mu	99.06%	34.97%	78.55%	6.32%	7.25%	8.61%	9.04%
Orig.+mu (Q)	98.94%	33.09%	73.78%	5.95%	5.21%	6.32%	6.82%
Adv.+mu	97.97%	91.77%	97.00%	95.18%	96.79%	95.99%	95.90%
Adv.+mu (Q)	37.12%	28.20%	35.35%	31.15%	34.29%	32.64%	32.15%
F.L.+mu	99.11%	48.08%	89.25%	70.86%	57.39%	64.53%	64.92%
F.L.+mu (Q)	98.93%	47.65%	**88.31%**	**69.45%**	**55.24%**	**62.64%**	**62.92%**

Table 2: The accuracy of black-box attacks and noises on MNIST models.

Models	CW-L2	FGSM (w)	FGSM (m)	FGSM (s)	Normal	Uniform
Orig.	97.56%	98.95%	97.80%	93.30%	97.19%	98.85%
Orig. (Q)	97.47%	98.47%	96.26%	90.08%	95.50%	98.38%
Adv.	97.28%	98.30%	98.22%	96.17%	77.16%	98.37%
Adv. (Q)	39.42%	45.09%	43.14%	28.02%	17.62%	42.99%
F.L.	97.04%	98.90%	97.36%	94.99%	97.01%	98.67%
F.L. (Q)	96.38%	**98.54%**	**96.84%**	**94.38%**	**96.58%**	**98.44%**
Orig.+mu	97.31%	98.72%	97.16%	90.61%	96.16%	98.69%
Orig.+mu (Q)	96.83%	98.31%	96.15%	88.69%	95.16%	98.27%
Adv.+mu	97.44%	97.83%	97.62%	94.09%	74.06%	97.81%
Adv.+mu (Q)	38.02%	40.06%	39.60%	24.48%	15.69%	37.32%
F.L.+mu	97.47%	98.70%	96.72%	93.76%	96.64%	98.58%
F.L.+mu (Q)	**97.68%**	98.46%	96.44%	93.54%	96.36%	98.21%

Experimental Results – Accuracy on CIFAR-10

- Similar results as MNIST with more significant improvement.
  - Adv. model suffers more from quantization.
  - White-box robustness improved by mapping in the Orig. model.
- Mapping the last three layers introduce more nonlinearity to models.

Table 3: The accuracy of white-box attacks on CIFAR-10 models.

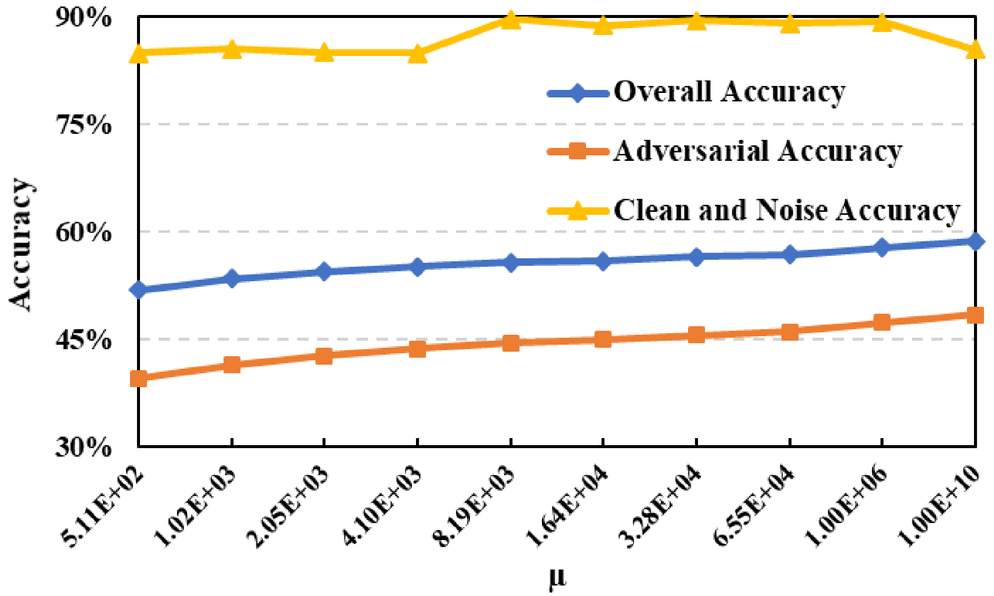
Models	Clean	CW-L2	FGSM (w)	FGSM (s)	PGD	BIM	MIM
Orig.	95.00%	9.30%	20.90%	10.60%	2.20%	2.60%	2.50%
Orig. (Q)	47.92%	13.60%	16.80%	11.90%	11.10%	17.80%	17.70%
Adv.	87.27%	54.20%	74.70%	36.80%	66.80%	57.60%	59.70%
Adv. (Q)	19.84%	15.80%	17.50%	10.90%	17.90%	18.20%	17.70%
F.L.	93.77%	20.30%	39.70%	27.50%	4.00%	4.00%	4.00%
F.L. (Q)	90.14%	21.30%	42.60%	28.70%	5.90%	5.90%	5.80%
Orig.+mu	94.05%	5.30%	95.30%	94.90%	64.40%	95.30%	95.30%
Orig.+mu (Q)	51.55%	11.60%	45.10%	46.80%	30.80%	49.50%	49.40%
Adv.+mu	85.70%	51.90%	83.30%	83.20%	81.60%	83.30%	83.30%
Adv.+mu (Q)	16.80%	17.00%	16.70%	16.70%	17.00%	17.30%	17.50%
F.L.+mu	93.80%	20.70%	92.80%	92.30%	89.50%	92.80%	92.80%
F.L.+mu (Q)	**92.20%**	**23.10%**	**90.80%**	**90.70%**	**86.90%**	**90.80%**	**90.80%**

Table 4: The accuracy of black-box attacks and noises on CIFAR-10 models.

Models	CW-L2	FGSM (w)	FGSM (m)	FGSM (s)	Normal	Uniform
Orig.	58.90%	55.07%	46.87%	41.12%	21.40%	43.80%
Orig. (Q)	23.00%	22.60%	20.64%	19.17%	19.30%	21.80%
Adv.	76.44%	75.82%	74.61%	73.48%	70.30%	84.90%
Adv. (Q)	19.38%	19.32%	18.92%	18.55%	15.60%	17.80%
F.L.	64.70%	61.82%	57.12%	53.68%	79.10%	85.50%
F.L. (Q)	**62.99%**	**60.30%**	**56.07%**	**52.44%**	**72.40%**	**81.90%**
Orig.+mu	55.95%	52.58%	44.74%	38.62%	20.90%	41.00%
Orig.+mu (Q)	25.64%	24.53%	21.41%	20.01%	15.80%	19.30%
Adv.+mu	73.24%	72.79%	71.52%	69.90%	68.20%	82.30%
Adv.+mu (Q)	15.74%	15.67%	15.23%	14.69%	11.10%	12.10%
F.L.+mu	63.69%	60.37%	55.58%	52.04%	73.60%	84.00%
F.L.+mu (Q)	62.54%	59.65%	55.03%	51.69%	72.20%	**81.90%**

Experimental Results – Ablation Study

- Nonlinearity vs. robustness: CIFAR-10, map only the last layer.
- As μ increases, adversarial robustness is improved, while nonlinear mapping may marginally harm accuracies on non-adversarial attacks.
- These results align with our theoretical assumptions.



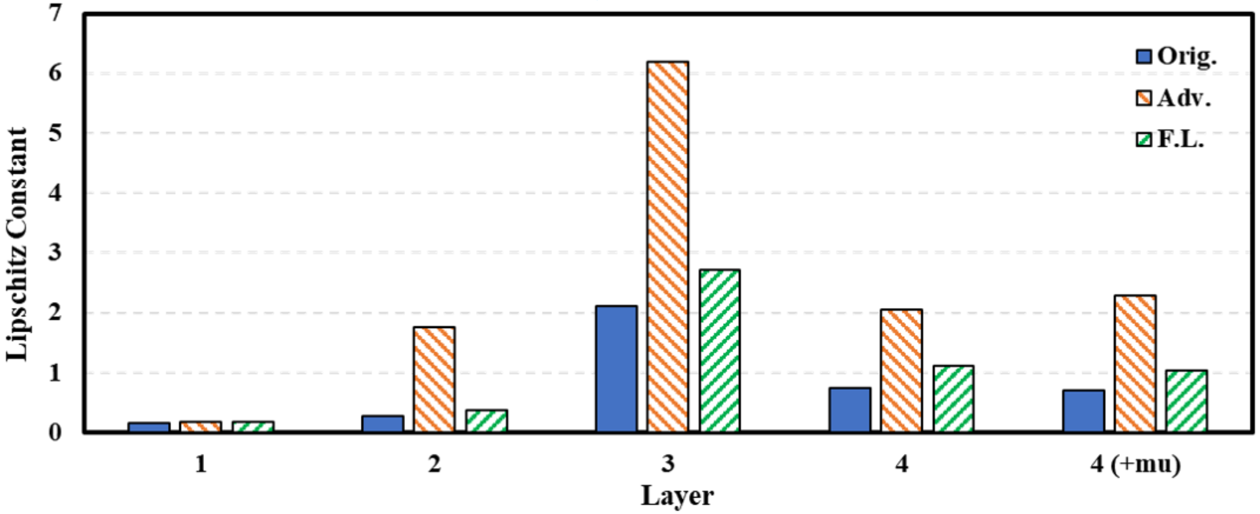
Experimental Results – Lipschitz Measurement

- The Lipschitz constant of the quantization weight loss (ΔW):

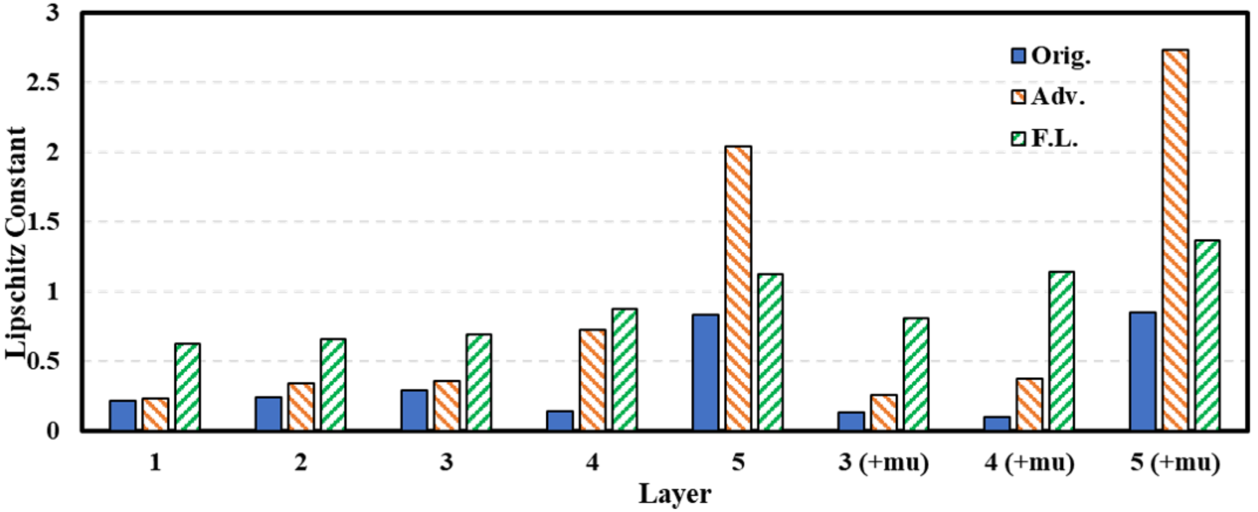
$$\|\Delta W\|_p = \sup_{z:\|z\|_p=1} \|\Delta W z\|_p$$

- When $p = 2$, $\|\Delta W\|_2$ is the maximum singular value of ΔW. $\|\Delta W\|_2 > 1$ means quantization error may be amplified in this layer.
- The adv model has weak tolerance to quantization.

$\|\Delta W\|_2$ of each layer in MNIST models.



$\|\Delta W\|_2$ of the last five layers in CIFAR-10 models.



Conclusions

- We observe that adversarially-trained neural networks are vulnerable to quantization loss.
- We theoretically analyze both adversarial and quantization losses and come up with criteria to measure the two losses. We also propose a solution to minimize both losses at the same time.
- The results show that our method is capable of defending both black-box and white-box gradient-based adversarial attacks and minimizing the quantization loss, showing an average accuracy improvement against adversarial attacks of 7.55% on MNIST and 27.84% on CIFAR-10 compared to the next best approach studied.

Thanks for your attention!
Q&A