Model Compression with Adversarial Robustness: A Unified Optimization Framework

Shupeng Gui[1,†], Haotao Wang[2,†], Haichuan Yang[1], Chen Yu[1], Zhangyang Wang[2], Ji Liu[3] †*Equal contribution*

[1]University of Rochester, [2]Texas A&M University, [3]Ytech Seattle AI lab, FeDA lab, AI platform, Kwai Inc

NeurIPS | 2019

INTRODUCTION

➤ Model Compression encounters Robustness

Can a Compression Algorithm lead to compressed models that are not only ACCURATE, but also ROBUST?

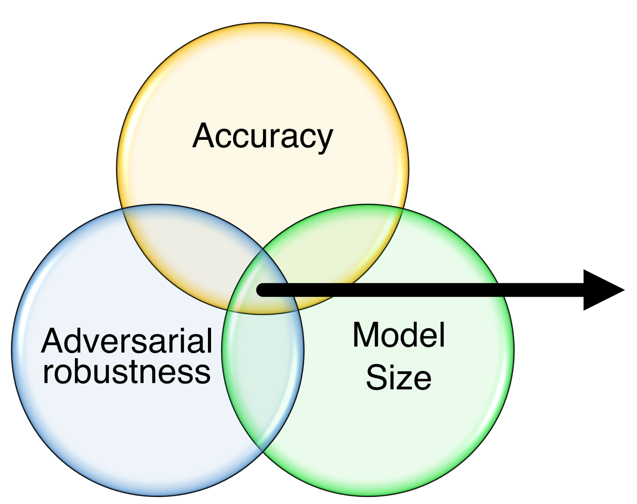
- Tsipras et al.[3] argued that the tradeoff between robustness and accuracy may be inevitable for the classification task.
- Nakkiran [2] showed theoretical examples implied that a both accurate and robust classifier might exist, given sufficiently large model capacity.
- Guo et al.[1] discovered that an appropriately higher CNN model sparsity led to better robustness, whereas over-sparsification could cause more fragility.

➤ Highlights of Contributions:

- First framework jointly optimizing **Model Compression** & **Adversarial Robustness**
- First framework unifies all existing compression methods **Pruning**, **Factorization** & **Quantization**

ADVERSARIALLY TRAINED MODEL COMPRESSION

➤ Optimize Three Goals Simultaneously



Unified Robust Model Compression ATMC

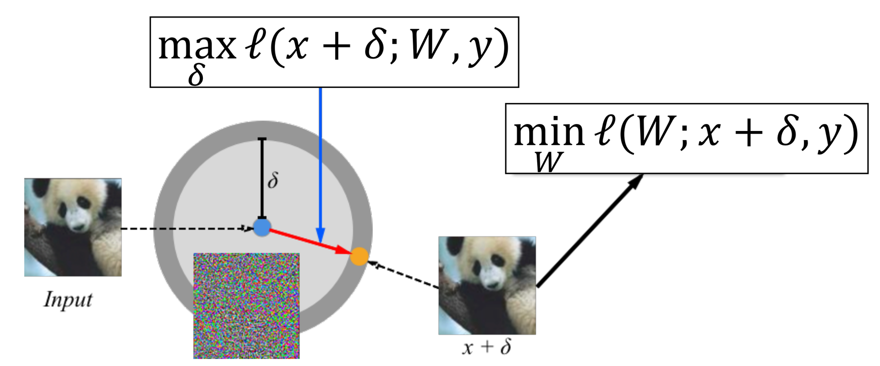
➤ The Overall Objective

$$\min_W \sum_{(x,y)\text{ in data set}} f^{\mathrm{adv}}(W;x,y) \qquad \text{Accuracy + Robustness}$$

$$\text{s.t.} \sum_l \|U^{(l)}\|_0 + \|V^{(l)}\|_0 + \|C^{(l)}\|_0 \le k \qquad \text{Model size}$$

$$W \in \mathcal{Q}_b := \left\{W: |U^{(l)}|_0 \le 2^b, |V^{(l)}|_0 \le 2^b, |C^{(l)}|_0 \le 2^b, \forall l \in [L]\right\}$$

➤ Robustness: Adversarial Training Loss



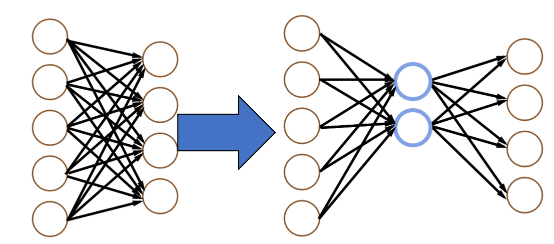
$$\min_W \sum_{(x,y)\text{ in data set}} f^{\mathrm{adv}}(W;x,y)$$

Def: $f^{adv}(W;x,y) = \max_{x' \in B_\infty^\Delta(x)} \ell(x';W,y)$

$B_\infty^\Delta(x) := \{x' \| x'-x\|_\infty \le \Delta\}$

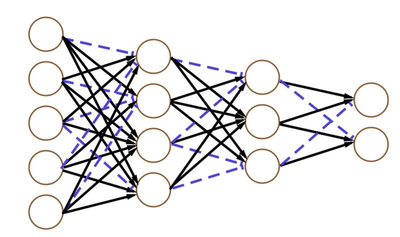
➤ Efficiency: Model Size Compression

$$W := \{W^{(l)}\}_{l\in[L]}, W^{(l)} = U^{(l)}V^{(l)} + C^{(l)}$$



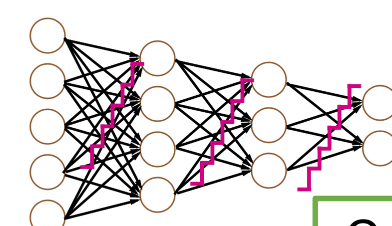
Factorization

$$\text{s.t.} \sum_l \|U^{(l)}\|_0 + \|V^{(l)}\|_0 + \|C^{(l)}\|_0 \le k$$



Weight Pruning

$$W \in \mathcal{Q}_b := \left\{W: |U^{(l)}|_0 \le 2^b, |V^{(l)}|_0 \le 2^b, |C^{(l)}|_0 \le 2^b, \forall l \in [L]\right\}$$



Quantization

ATMC: OPTIMIZATION

➤ Duplicate Variables

$$\min_{\substack{\|W\|_0 \le k,\\ W' \in \mathcal{Q}_b}} \sum_{(x,y)\text{ in data set}} f^{\mathrm{adv}}(W;x,y) \quad \text{s.t. } W = W'$$

Def: $\|W\|_0 := \sum_l \|U^{(l)}\|_0 + \|V^{(l)}\|_0 + \|C^{(l)}\|_0$

➤ Removing the Equality Constraint $W = W'$

$$\min_{\substack{\|W\|_0 \le k,\\ W' \in \mathcal{Q}_b}} \max_u \sum_{(x,y)\text{ in data set}} f^{\mathrm{adv}}(W;x,y) + \rho\langle u, W - W'\rangle + \frac{\rho}{2}\|W - W'\|_F^2$$

Def: $\rho > 0$ as predefined positive number in ADMM

➤ Given U in an arbitrary layer

Update u: $u_{t+1} = u_t + (U - U')$

Update x^{adv}: $x^{\mathrm{adv}} \leftarrow \mathrm{Proj}_{\|x'-x\|_\infty \le \Delta}\{x + \alpha\nabla_x f(W;x,y)\}$

Update U: $U \leftarrow \mathrm{Proj}_{\|U''\|_0 \le k}\{U - \gamma\nabla_U[f\left(U;x^{\mathrm{adv}},y\right) + \frac{\rho}{2}\|U - U' + u\|_F^2]\}$

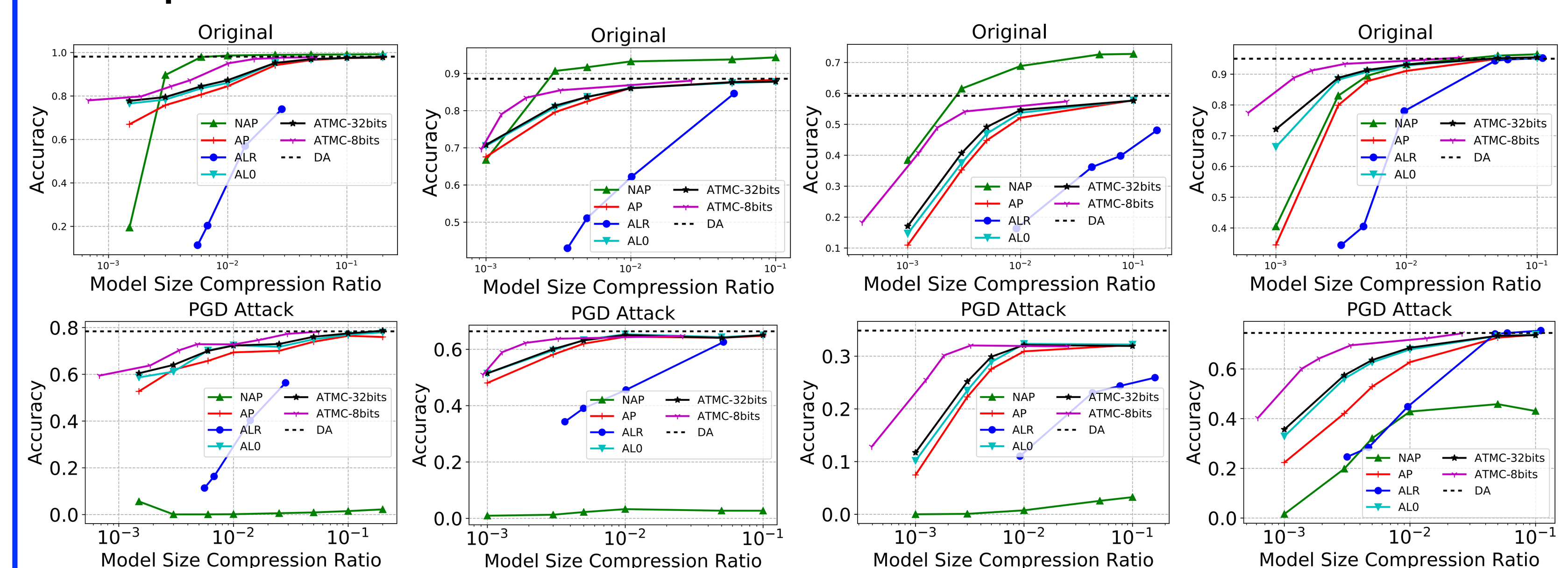
Update U': $U' = \underset{U'}{\mathrm{argmin}}\|U' - (U+u)\|_F^2$, s.t. $|U'|_0 \le 2^b$ (Lloyd's algorithm)

EXPERIMENTS: CNNs

➤ Datasets & CNN Models

Models	#Parameters	Bit width	Model Size (bits)	Dataset & Accuracy
LeNet	430K	32	13,776,000	MNIST: 99.32%
ResNet34	21M	32	680,482,816	CIFAR-10: 93.67%
ResNet34	21M	32	681,957,376	CIFAR-100: 73.16%
WideResNet	11M	32	350,533,120	SVHN: 95.25%

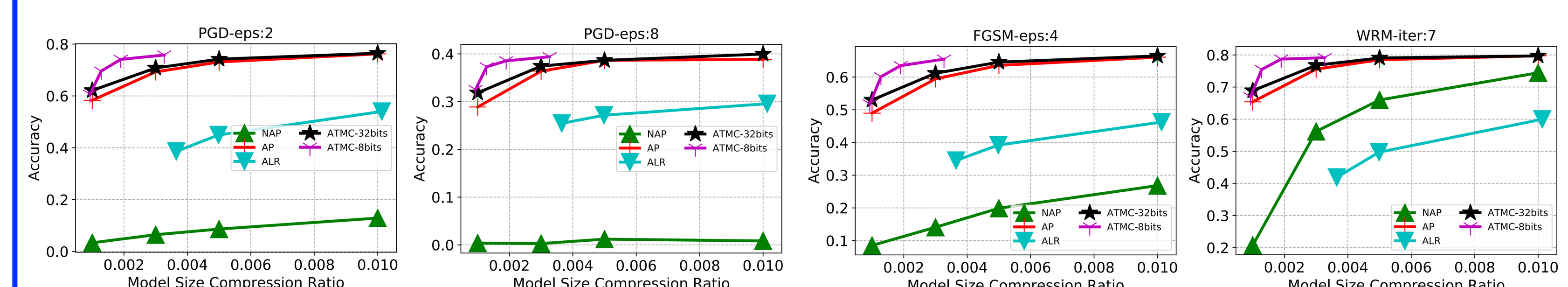
➤ Outstanding Performance on Trade-off between Compression and Robustness for ATMC



MNIST | CIFAR-10 | CIFAR-100 | SVHN

➤ Consistent Adversarial Robustness under Various Attack Settings

- Different perturbation magnitude, e.g., 2, 8
- Different adversarial attack methods, e.g., FGSM, WRM



PGD, Perturbation=2 | PGD, Perturbation=8 | FGSM, Perturbation=4 | WRM, Penalty=1.3, Iteration=7

More details https://github.com/TAMU-VITA/ATMC

REFFERENCE

[1] Guo et al, "Sparse DNNs with improved adversarial robustness", NeurIPS 2018

[2] Nakkiran, "Adversarial robustness may be at odds with simplicity". arxiv preprint

[3] Tsipras et al, "Robustness may be at odds with accuracy". STAT, 1050:11, 2018