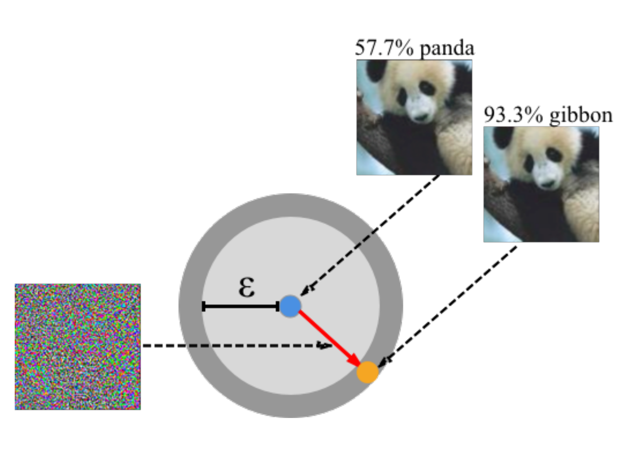
Model Compression with Adversarial Robustness: A Unified Optimization Framework

Shupeng Gui[1,†], Haotao Wang[2,†], Haichuan Yang[1], Chen Yu[1], Zhangyang Wang[2], Ji Liu[3]
[1]University of Rochester, [2]Texas A&M University, [3]Ytech Seattle AI lab, FeDA lab, AI platform, Kwai Inc
† Equal contribution

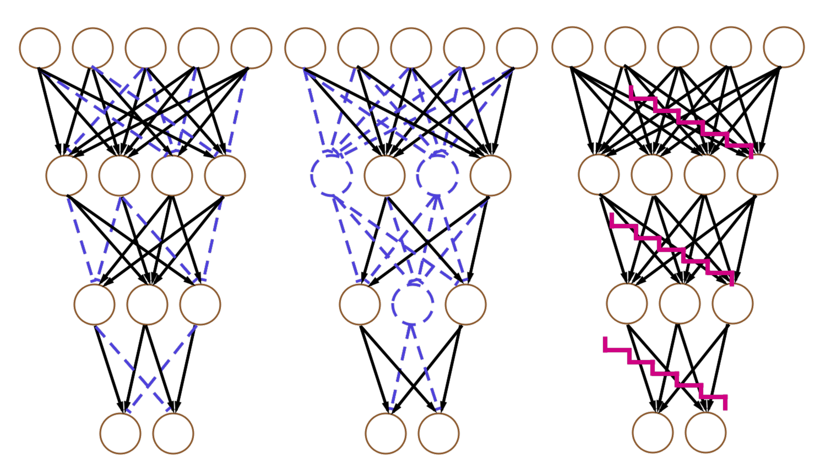
Highlights

- Model compression without hurting their robustness to adversarial attacks, in addition to maintaining accuracy
- Adversarially Trained Model Compression (ATMC) framework
- Integration of pruning, factorization, and quantization into constraints
- An extensive group of experiments demonstrate that ATMC obtains remarkably more favorable trade-off among model size, accuracy and robustness, over currently available alternatives in various settings.

Motivation: Compression & Robustness



Adversarial Robustness



Weight Pruning Node Pruning Quantization

- Highly non-straightforward and contextually varying w.r.t different means of compression

Adversarial Training + Model Compression = ?

We propose

A Unified Robust Model Compression Framework -- ATMC

Adversarial Training Model Compression

Adversarially Trained Model Compression

The overall objective:

$$\min_{\theta} \sum_{(x,y)\in\mathcal{Z}} f^{adv}(\theta; x, y)$$

DNN parameters — Adversarial Training Loss

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

Sparsity Constraint

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

Quantization Constraint

Unify Sparsification & Channel Pruning

$$W = UV + C, \|U\|_0 + \|V\|_0 + \|C\|_0 \le k$$

Non-uniform Quantization Strategy

$$\left|U^{(l)}\right|_0 \le 2^b, \left|V^{(l)}\right|_0 \le 2^b, \left|C^{(l)}\right|_0 \le 2^b$$

Defensive adversarial training objective

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

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

DNN Optimization for ATMC

Basic Idea: ADMM + Minimax Optimization

Update Adversarial Sample:

$$x^{adv} \leftarrow \mathrm{Proj}_{B_\infty^\Delta(x)}\{x + \alpha \nabla_x f(\theta; x, y)\}$$

Duplicate Weights θ (sparsity), θ' (quantization)

1) Update Sparse:

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

2) Update Quantized:

$$\min_{\theta'} \|\theta' - (\theta + u)\|_F^2, \text{ s.t. } \theta' \in \mathcal{Q}_b$$

Essentially solving:

$$\min_{\theta, \{a_k\}_{k=1}^{2^b}} \|\theta + u - \theta'\|_F^2, \text{s.t. } \theta_{i,j} \in \{0, a_1, a_2, \cdots, a_{2^b}\}$$

Lloyd's Algorithm:

$$\theta' \leftarrow \mathrm{ZeroKmeans}_{2^b}(\theta + u_\theta)$$

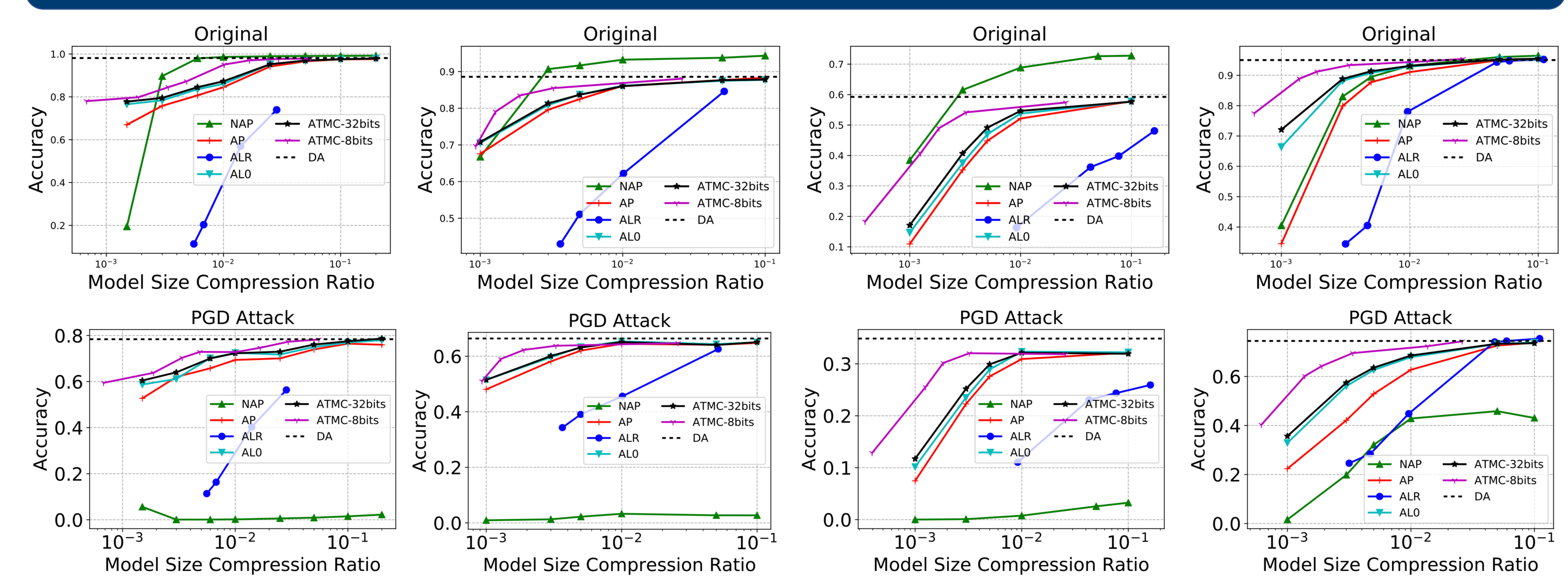
Algorithm 1 ZeroKmeans$_B(\bar{U})$

Input: a set of real numbers $\bar{U}$, number of clusters B.
Output: quantized tensor U.
Initialize $a_1, a_2, \cdots, a_B$ by randomly picked nonzero elements from $\bar{U}$.
$Q := \{0, a_1, a_2, \cdots, a_B\}$
repeat
  for $i = 0$ to $|\bar{U}| - 1$ **do**
    $\delta_i \leftarrow \arg\min_j (\bar{U}_i - Q_j)^2$
  end for
  Fix $Q_0 = 0$
  for $j = 1$ to B **do**
    $a_j \leftarrow \frac{\sum_i \mathbf{I}(\delta_i = j)\bar{U}_i}{\sum_i \mathbf{I}(\delta_i = j)}$
  end for
until Convergence
for $i = 0$ to $|\bar{U}| - 1$ **do**
  $U_i \leftarrow Q_{\delta_i}$
end for

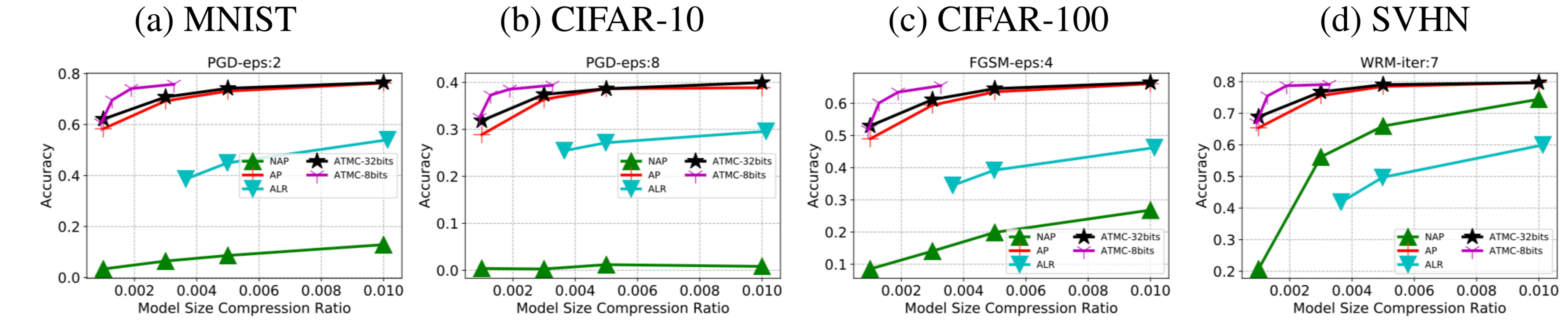
Algorithm 2 ATMC

Input: dataset $\mathcal{Z}$, stepsize sequence $\{\gamma_t > 0\}_{t=0}^{T-1}$, update steps n and T, hyper-parameter ρ, k, and b, Δ
Output: model θ
$\alpha \leftarrow 1.25 \times \Delta/n$
Initialize θ, let $\theta' = \theta$ and $u = 0$
for $t = 0$ **to** $T - 1$ **do**
  Sample (x, y) from $\mathcal{Z}$
  for $i = 0$ to $n - 1$ **do**
    $x^{adv} \leftarrow \mathrm{Proj}_{\{x' : \|x' - x\|_\infty \le \Delta\}}\{x + \alpha \nabla_x f(\theta; x, y)\}$
  end for
  $\theta \leftarrow \mathrm{Proj}_{\{\theta'' : \|\theta''\|_0 \le k\}}(\theta - \gamma_t \nabla_\theta [f(\theta; x^{adv}, y) + \frac{\rho}{2}\|\theta - \theta' + u\|_F^2])$
  $\theta' \leftarrow \mathrm{ZeroKmeans}_{2^b}(\theta + u)$
  $u \leftarrow u + (\theta - \theta')$
end for

Experiment Results



(a) MNIST (b) CIFAR-10 (c) CIFAR-100 (d) SVHN



(a) PGD, perturbation=2 (b) PGD, perturbation=8 (c) FGSM, perturbation=4 (d) WRM, penalty=1.3, iteration=7

Conclusion

- Propose ATMC by integrating Model Compression and Adversarial Robustness;
- Unify pruning and quantization in one stage problem;
- Endorse the effectiveness of ATMC by a series of experiments;