Model Compression with Adversarial Robustness: A Unified Optimization Framework

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

[1]Department of Computer Science, University of Rochester

[2]Department of Computer Science and Engineering, Texas A&M University

[3]Ytech Seattle AI lab, FeDA lab, AI platform, Kwai Inc

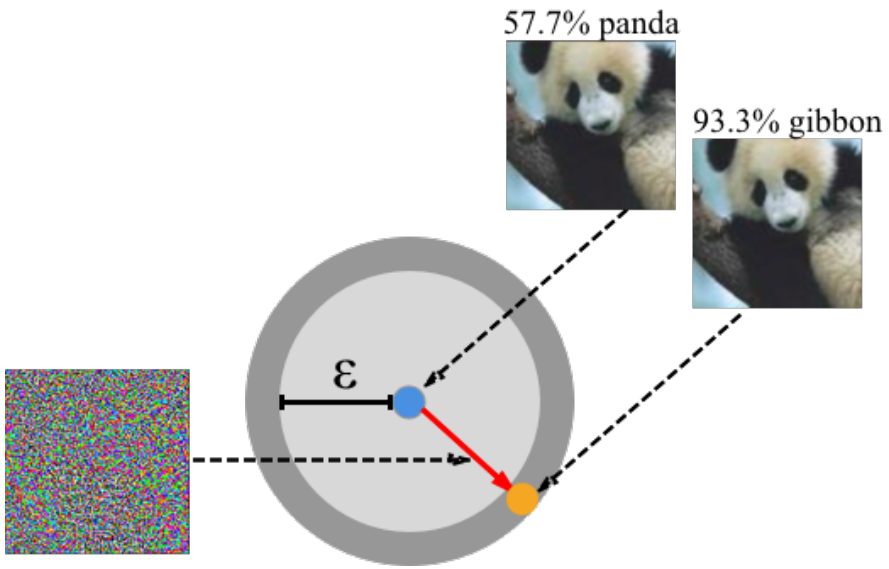
† Equal contribution

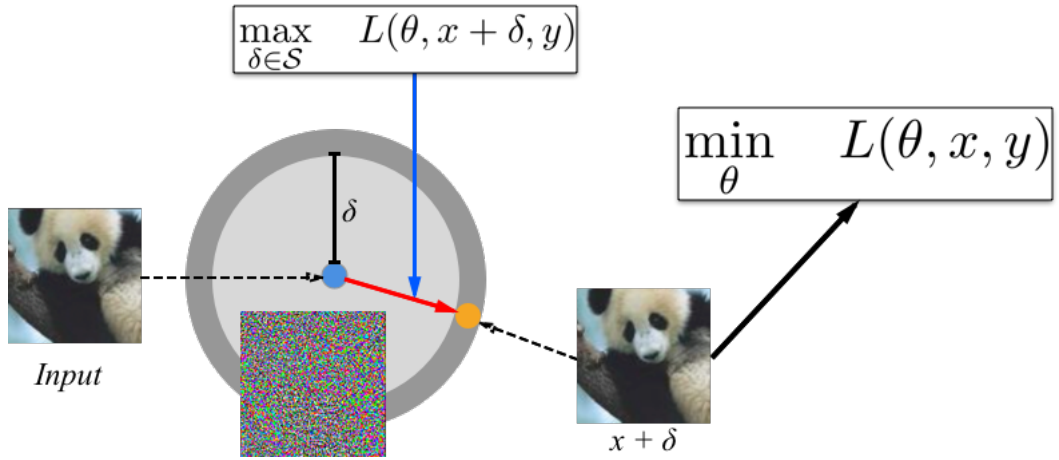
NeurIPS 2019, Vancouver CANADA

Highlights

- Compressing models without hurting their robustness to adversarial attacks, in addition to maintaining accuracy
- Adversarially Trained Model Compression (ATMC) framework
- Integrating 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.

Robustness for CNN



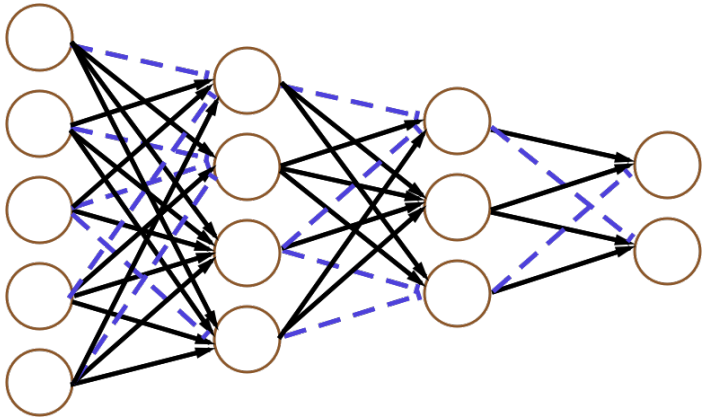


$$\min_{\theta} \rho(\theta), \quad \text{where } \rho(\theta) = \mathbb{E}_{(x,y)\in\mathcal{D}} \left[\max_{\delta} L(\delta, x+\delta, y) \right]$$

Goodfellow et al, "Explaining and Harnessing Adversarial Examples", ICLR 2015.

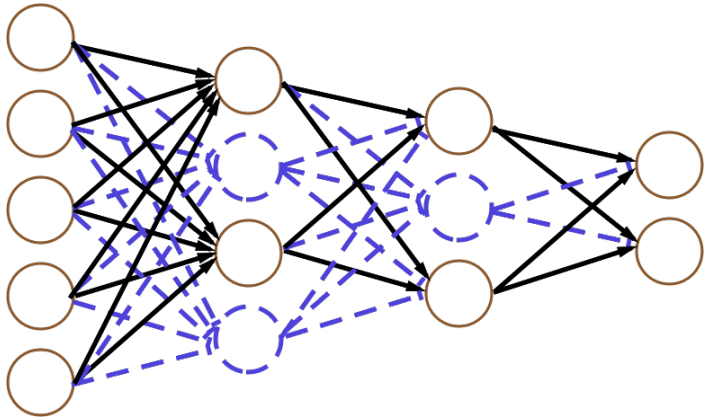
Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks.", ICLR 2018

Compression for CNN



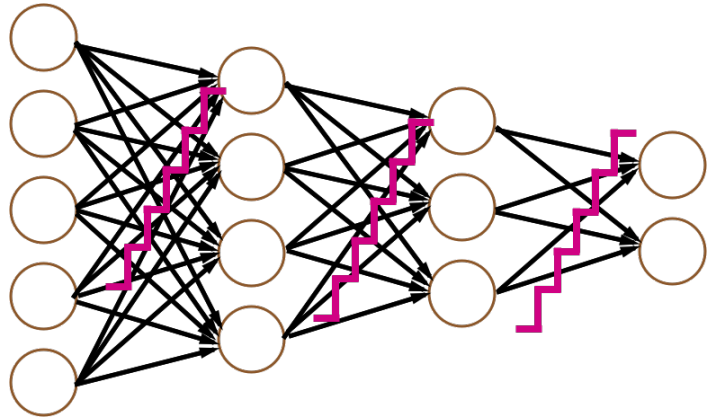
Weight Pruning

[Han et al., NIPS 15]



Node Pruning

[Li et al., ICLR 17]
[He et al., ICCV 17]
[Porikli et al., ECCV 16]



Quantization

[Han et al., ICLR 16]

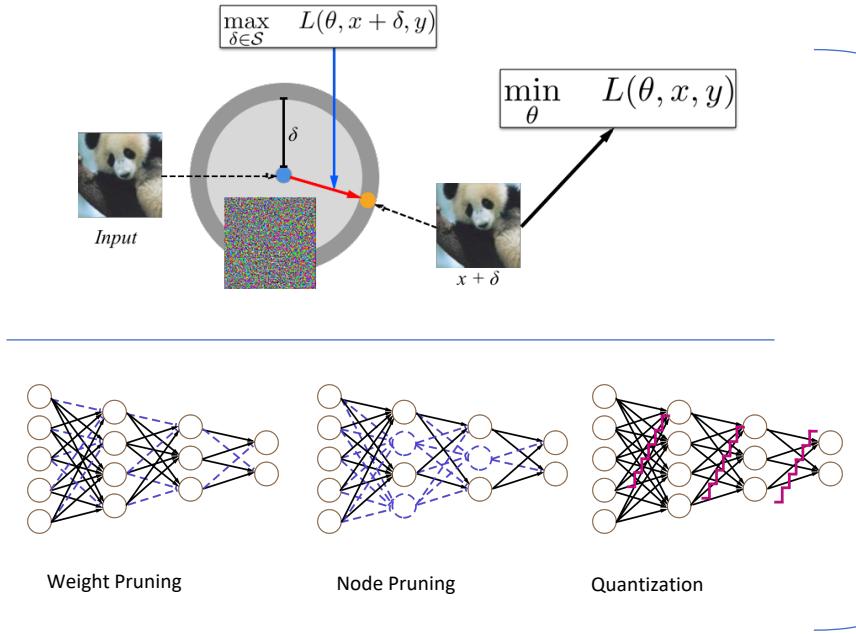
Compression v.s. Robustness

- Highly non-straightforward and contextually varying w.r.t different means of compression
- An appropriately higher CNN model sparsity led to better robustness, whereas over-sparsification could in turn cause more fragility [1]
- [2] showed that sparse algorithms are not stable:
  - If an algorithm promotes sparsity, then its sensitivity to small perturbations fo the input data remains bounded away from zero
  - Quantization seems to reduce the Minimum Description Length and might potentially make the algorithm more robust.
- [3] argued that the tradeoff between robustness and accuracy may be inevitable for the classification task.

[1] Guo et al, "Sparse dnns with improved adversarial robustness", NeurIPS 2018
[2] Xu et al, "Sparse algorithms are not stable: A no-free-lunch theorem". IEEE t-PAMI 2012
[3] Tsipras et al, "Robustness may be at odds with accuracy". Stat, 1050:11, 2018

Adversarially Trained Model Compression (Overall)

- A unified optimization framework for both compression and robustness on CNN



$$\min_{\boldsymbol{\theta}} \quad \sum_{(x,y)\in\mathcal{Z}} f^{\text{adv}}(\boldsymbol{\theta}; x, y) \tag{4}$$

$$\text{s.t.} \quad \underbrace{\sum_{i=1}^{l} \|\boldsymbol{U}^{(i)}\|_0 + \|\boldsymbol{V}^{(i)}\|_0 + \|\boldsymbol{C}^{(i)}\|_0}_{\|\boldsymbol{\theta}\|_0} \leq k, \ \text{(sparsity constraint)}$$

$$\boldsymbol{\theta} \in \mathcal{Q}_b := \{\boldsymbol{\theta} : |\boldsymbol{U}^{(l)}|_0 \leq 2^b, \ |\boldsymbol{V}^{(l)}|_0 \leq 2^b, \ |\boldsymbol{C}^{(l)}|_0 \leq 2^b \ \forall l \in [L]\}. \text{(quantization constraint)}$$

Adversarially Trained Model Compression (Unified Pruning Constraints)

- Sparsification and channel pruning into one constraint

$$\boldsymbol{W} = \boldsymbol{U}\boldsymbol{V} + \boldsymbol{C}, \quad \|\boldsymbol{U}\|_0 + \|\boldsymbol{V}\|_0 + \|\boldsymbol{C}\|_0 \leq k$$

where $||\cdot||_0$ denotes the number of nonzeros of the augment matrix

- The above enforces a novel, compound (including both multiplicative and additive) sparsity structure on W
- UV indicates a natural form of channel pruning
- Such form can be used in the model acceleration with other norm such as L21

Adversarially Trained Model Compression (Non-uniform Quantization)

- Each nonzero element of the DNN parameter can only be chosen from a set of a few values
- These values are not necessarily evenly distributed and need to be optimized
- With respect to the non-uniform quantization strategy, we introduce the quantization constraints

$$|\boldsymbol{U}^{(l)}|_0 \leq 2^b, \ |\boldsymbol{V}^{(l)}|_0 \leq 2^b, \ |\boldsymbol{C}^{(l)}|_0 \leq 2^b \quad \forall l \in [L].$$

Adversarially Trained Model Compression (Adversarial Robustness)

- White-box attack setting: an adversary to eavesdrop the optimization and gradients of the learning model
- When an “clean" image x comes to a target model, “perturb” the image with an adversarial perturbation with bounded magnitudes.

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

- The training objective for the attacker

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

- Defensive objective for ATMC

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

Optimization for ATMC

- Apply ADMM to solve the optimization problem
- Update x^{adv}: $x^{\text{adv}} \leftarrow \text{Proj}_{\{x':\|x'-x\|_\infty \leq \Delta\}} \{x + \alpha \nabla_x f(\boldsymbol{\theta}; x, y)\}$
- Update θ: $\min_{\boldsymbol{\theta}} \quad f(\boldsymbol{\theta}; x^{\text{adv}}, y) + \frac{\rho}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}' + \boldsymbol{u}\|_F^2 \quad \text{s.t. } \|\boldsymbol{\theta}\|_0 \leq k.$
  - Only related to the sparsity constraint

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

- Update θ': $\min_{\boldsymbol{\theta}'} \quad \|\boldsymbol{\theta}' - (\boldsymbol{\theta} + \boldsymbol{u})\|_F^2, \quad \text{s.t. } \boldsymbol{\theta}' \in \mathcal{Q}_b$
  - Only related to quantization constraint
  - For example, essentially solving

$$\min_{\boldsymbol{U}, \{\boldsymbol{a}_k\}_{k=1}^{2^b}} \quad \|\boldsymbol{U} - \bar{\boldsymbol{U}}\|^2 \quad \text{s.t.} \quad \boldsymbol{U}_{i,j} \in \{0, \boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_{2^b}\}$$

  - Lloyd's algorithm [1]

$$\boldsymbol{U}_t'^{(l)} = \text{ZeroKmeans}_{2^b}(\boldsymbol{U}^{(l)} + \boldsymbol{u}_{\boldsymbol{U}^{(l)}})$$

[1] Lloyd, "Least squares quantization in pcm", IEEE transactions on information theory, 1982

Optimization for ATMC

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

Input: a set of real numbers $\bar{\boldsymbol{U}}$, number of clusters B.
Output: quantized tensor $\boldsymbol{U}$.
Initialize $\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_B$ by randomly picked nonzero elements from $\bar{\boldsymbol{U}}$.
$\boldsymbol{Q} := \{0, \boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_B\}$
repeat
  for $i = 0$ to $|\bar{\boldsymbol{U}}| - 1$ **do**
    $\delta_i \leftarrow \arg\min_j (\bar{\boldsymbol{U}}_i - \boldsymbol{Q}_j)^2$
  end for
  Fix $\boldsymbol{Q}_0 = 0$
  for $j = 1$ to B **do**
    $\boldsymbol{a}_j \leftarrow \frac{\sum_i \mathbf{I}(\delta_i = j)\bar{\boldsymbol{U}}_i}{\sum_i \mathbf{I}(\delta_i = j)}$
  end for
until Convergence
for $i = 0$ to $|\bar{\boldsymbol{U}}| - 1$ **do**
  $\boldsymbol{U}_i \leftarrow \boldsymbol{Q}_{\delta_i}$
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 $\boldsymbol{\theta}$
$\alpha \leftarrow 1.25 \times \Delta / n$
Initialize θ, let $\boldsymbol{\theta}' = \boldsymbol{\theta}$ and $\boldsymbol{u} = 0$
for $t = 0$ **to** $T - 1$ **do**
  Sample (x, y) from $\mathcal{Z}$
  for $i = 0$ to $n - 1$ **do**
    $x^{\text{adv}} \leftarrow \text{Proj}_{\{x' : \|x' - x\|_\infty \le \Delta\}} \{x + \alpha \nabla_x f(\boldsymbol{\theta}; x, y)\}$
  end for
  $\boldsymbol{\theta} \leftarrow \text{Proj}_{\{\boldsymbol{\theta}'' : \|\theta''\|_0 \le k\}} \big(\boldsymbol{\theta} - \gamma_t \nabla_{\boldsymbol{\theta}} [f(\boldsymbol{\theta}; x^{\text{adv}}, y) + \frac{\rho}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}' + \boldsymbol{u}\|_F^2]\big)$
  $\boldsymbol{\theta}' \leftarrow \text{ZeroKmeans}_{2^b}(\boldsymbol{\theta} + \boldsymbol{u})$
  $\boldsymbol{u} \leftarrow \boldsymbol{u} + (\boldsymbol{\theta} - \boldsymbol{\theta}')$
end for

Empirical Study

- Datasets and Benchmark Models
  - Four popular image classification datasets
  - Pick one top-performer CNN model on each

Table 1: The datasets and CNN models used in the experiments.

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%

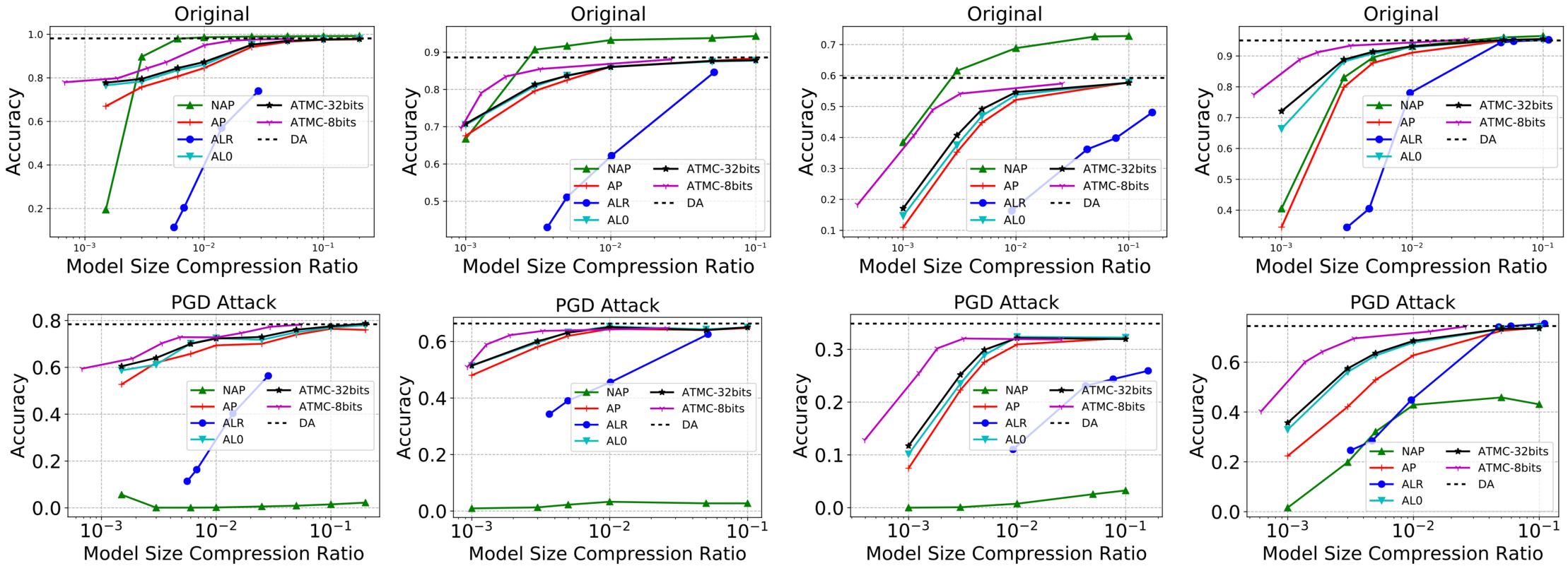
Empirical Study

- Evaluation Metrics
  - Classification accuracies on both benign and on attacked testing sets are reported
- Model Size
  - multiplying the quantization bit per element with the total number of non-zero elements, added with the storage size for the quantization thresholds
- Compression Ratio
  - the ratio between the compressed and original model sizes
- Adversarial Attack Settings
  - PGD attack with the same settings as used in adversarial training on testing sets to evaluate model robustness

Empirical Study

- Baselines
  - Non-adversarial Pruning (NAP)
  - Dense Adversarial Training (DA)
  - Adversarial Pruning (AP)
  - Adversarial l0 Pruning (Al0)
  - Adversarial Low-Rank Decomposition (ALR)
  - ATMC (8bits, 32 bits)
- Comparison on same compression ratio

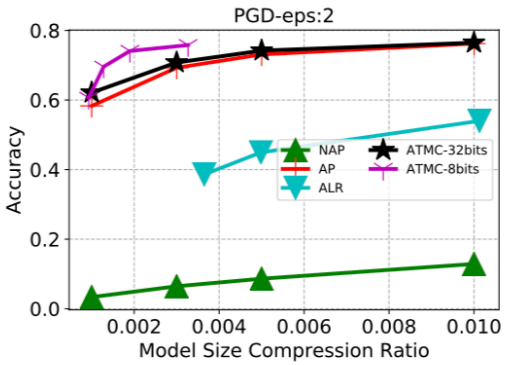
Empirical Study



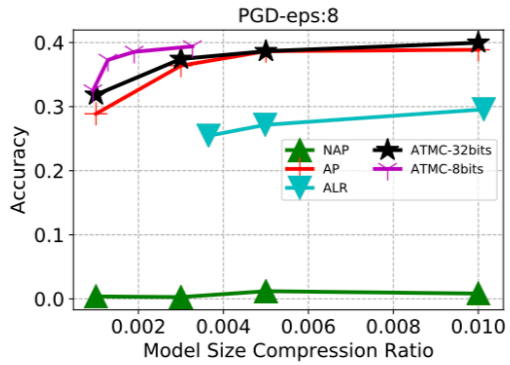
(a) MNIST (b) CIFAR-10 (c) CIFAR-100 (d) SVHN

Empirical Study

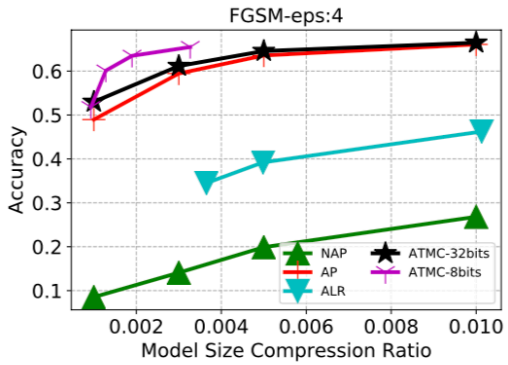
- ATMC also maintains robustness under different adversarial attacks



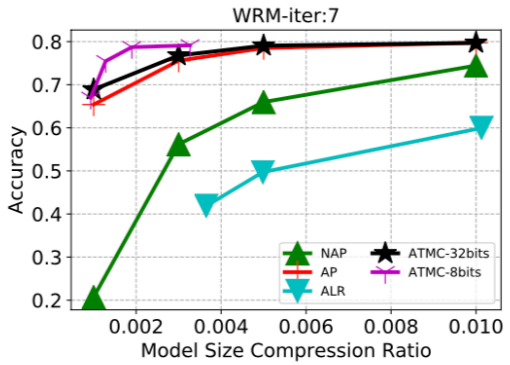
(a) PGD, perturbation=2



(b) PGD, perturbation=8



(c) FGSM, perturbation=4

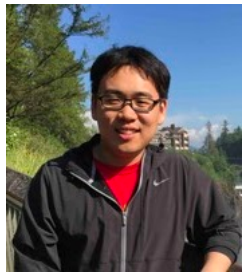


(d) WRM, penalty=1.3, iteration=7

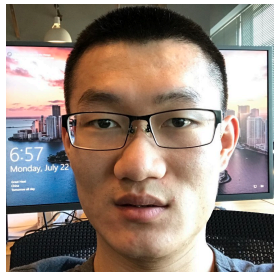
Conclusion

- We propose the ATMC framework, by integrating the two goals, model compression and robustness in one unified constrained optimization framework
- Our extensive expriments endorse the effectiveness of ATMC
  - Naïve model compression may hurt robustness, if the latter is not explicitly taken into account
  - A proper joint optimization could achieve both well
  - A properly compressed model could even maintain almost the same accuracy and robustness compared to the original one

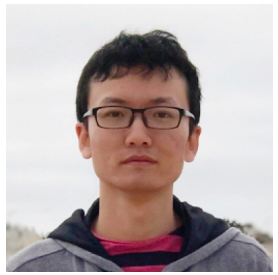
Thanks



Shupeng Gui

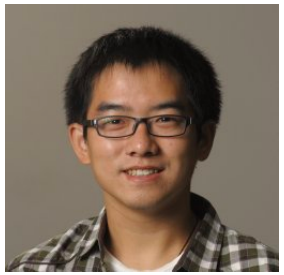


Haotao Wang



Haichuan Yang

Chen Yu



Zhangyang Wang

Ji Liu

UNIVERSITY of ROCHESTER

VITA

TEXAS A&M UNIVERSITY

Reference

- Guo, Yiwen, et al. "Sparse dnns with improved adversarial robustness." *Advances in neural information processing systems*. 2018.
- Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).
- Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." *arXiv preprint arXiv:1706.06083* (2017).
- Han, Song, et al. "Learning both weights and connections for efficient neural network." *Advances in neural information processing systems*. 2015.
- Xu, Huan, Constantine Caramanis, and Shie Mannor. "Sparse algorithms are not stable: A no-free-lunch theorem." *IEEE transactions on pattern analysis and machine intelligence* 34.1 (2011): 187-193.
- Tsipras, Dimitris, et al. "Robustness may be at odds with accuracy." *arXiv preprint arXiv:1805.12152* (2018).
- Li, Hao, et al. "Pruning filters for efficient convnets." *arXiv preprint arXiv:1608.08710* (2016).
- Lloyd, “Least squares quantization in pcm”, IEEE transactions on information theory, 1982
- Han, Song, Huizi Mao, and William J. Dally. "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding." *arXiv preprint arXiv:1510.00149* (2015).
- He, Yihui, Xiangyu Zhang, and Jian Sun. "Channel pruning for accelerating very deep neural networks." *Proceedings of the IEEE International Conference on Computer Vision*. 2017.
- Zhou, Hao, Jose M. Alvarez, and Fatih Porikli. "Less is more: Towards compact cnns." *European Conference on Computer Vision*. Springer, Cham, 2016.