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

Summary

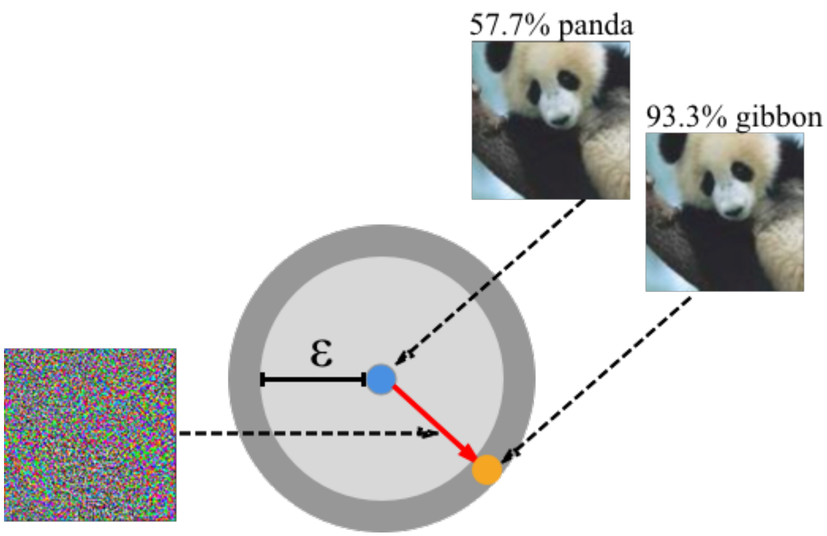
- Goals
  - Model Compression
  - Adversarial Robustness
  - Accuracy
- **A**dversarially **T**rained **M**odel **C**ompression (**ATMC**) Framework
  - Adversarial objective
  - Integrating three compression ways simultaneously (Pruning, Factorization, and Quantization)
- Experiments: Better trade-off among model size, accuracy, and robustness, over currently available alternatives in various settings.

Highlights of Contributions

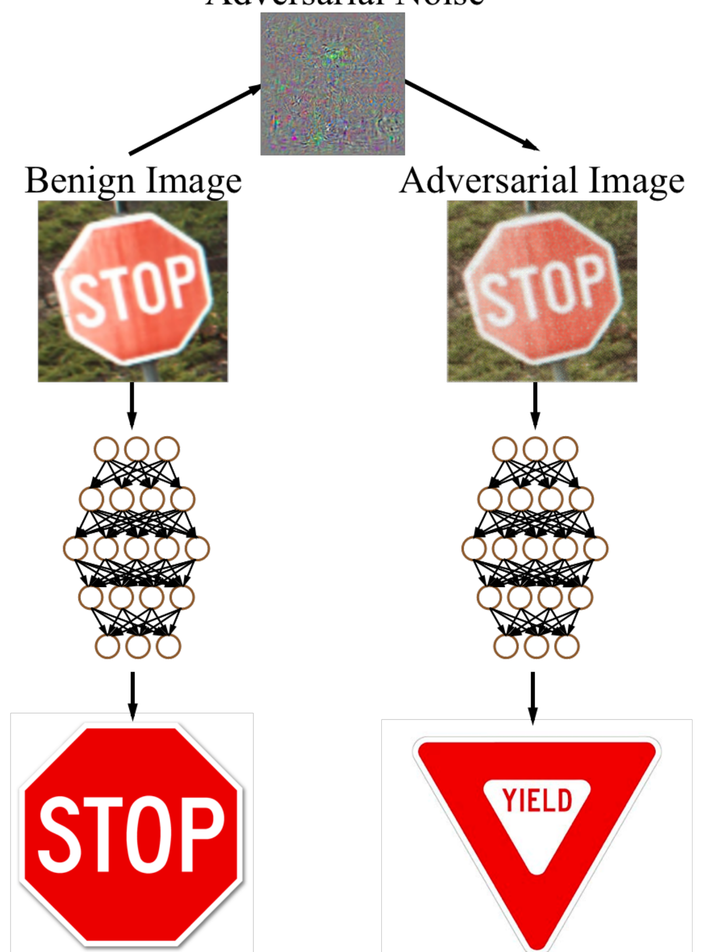
- First framework jointly optimizing
  - Model compression
  - Adversarial robustness
- First framework unifies all existing compression methods
  - Pruning
  - Factorization
  - Quantization

Why Adversarial Robustness?

- Easy to fool normal DNN classifiers

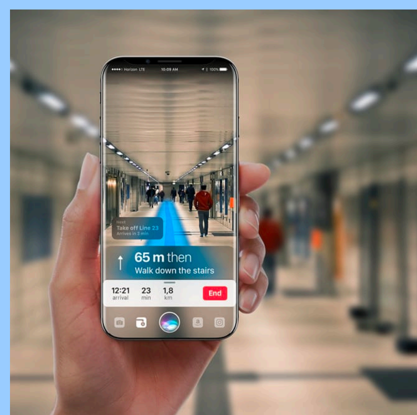


Goodfellow et al, "Explaining and Harnessing Adversarial Examples", ICLR 2015.

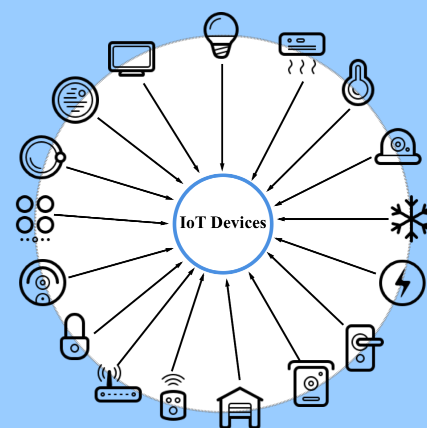


Why Model Compression?

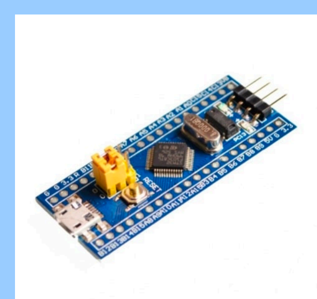
- Efficient inference on edging devices



App & Battery

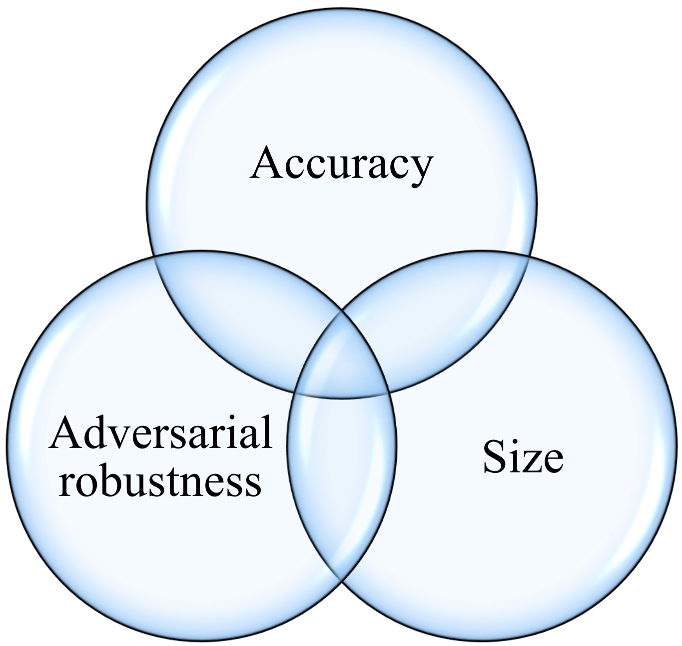


IoT & Limitation



https://www.cultofmac.com/492544/iphone-8-pack-lasers-improved-ar-autofocus/
https://www.digitaltrends.com/mobile/how-to-save-battery-life-on-your-smartphone/
https://images.app.goo.gl/92RrppKfD3bUYEwX8

Our Work: Optimize Three Goals Simultaneously



Our Framework: Adversarially Trained Model Compression

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

Accuracy + Robustness

$$\text{s.t.} \sum_{l} \left\|U^{(l)}\right\|_0 + \left\|V^{(l)}\right\|_0 + \left\|C^{(l)}\right\|_0 \leq k$$

Model size

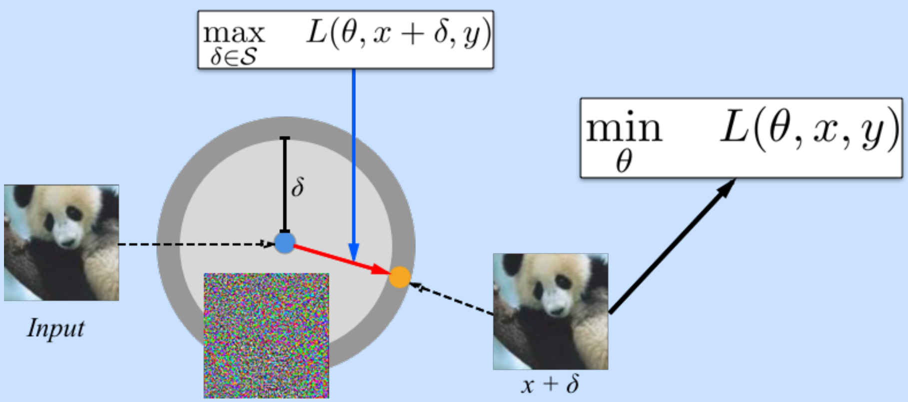
$$W \in \mathcal{Q}_b := \left\{W : \left|U^{(l)}\right|_0 \leq 2^b, \left|V^{(l)}\right|_0 \leq 2^b, \left|C^{(l)}\right|_0 \leq 2^b, \forall l \in [L]\right\}$$

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

$\|\cdot\|_0$: L-0 norm

$|\cdot|_0$: # Unique scalars

Adverbial Training Loss



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

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

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

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

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

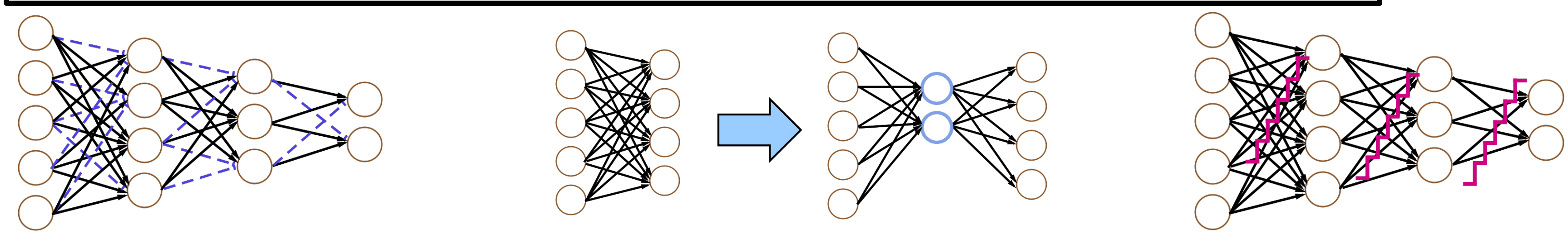
Model Size Compression: Combine Three Compression Ways

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

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

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

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



Weight Pruning

Factorization

Quantization

ATMC: Optimization

- Duplicate Variables
  - $\min_{\substack{\|W\|_0 \leq k, \\ W' \in \mathcal{Q}_b}} \sum_{(x,y) \text{ in data set}} f^{\text{adv}}(W; x, y)$ s.t. $W = W'$
  - $\|W\|_0 \coloneqq \sum_l \left\|U^{(l)}\right\|_0 + \left\|V^{(l)}\right\|_0 + \left\|C^{(l)}\right\|_0$
- Removing the equality constraint $W = W'$
  - $\min_{\substack{\|W\|_0 \leq k, \\ W' \in \mathcal{Q}_b}} \max_{u} \sum_{(x,y) \text{ in data set}} f^{\text{adv}}(W; x, y) + \rho\langle u, W - W'\rangle + \frac{\rho}{2}\|W - W'\|_F^2$
  - $\rho > 0$ as predefined positive number in ADMM

ATMC: Optimization

- E.g., given U in an arbitrary layer
- Update u:
  - $u_{t+1} = u_t + (U - U')$
- Update x^{adv}:
  - $x^{adv} \leftarrow \text{Proj}_{\|x'-x\|_\infty \leq \Delta}\{x + \alpha \nabla_x f(W; x, y)\}$
- Update U:
  - $U \leftarrow \text{Proj}_{\|U''\|_0 \leq k}\{U - \gamma \nabla_U [f\left(U; x^{\text{adv}}, y\right) + \frac{\rho}{2}\|U - U' + u\|_F^2]\}$
- Update U':
  - Solving the following projection problem for θ'
    - $\min_{U'} \|U' - (U + u)\|_F^2, \quad s.t. |U'|_0 \leq 2^b$
  - Lloyd's algorithm

Experiment: CNNs

- 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%

Experiment: Settings

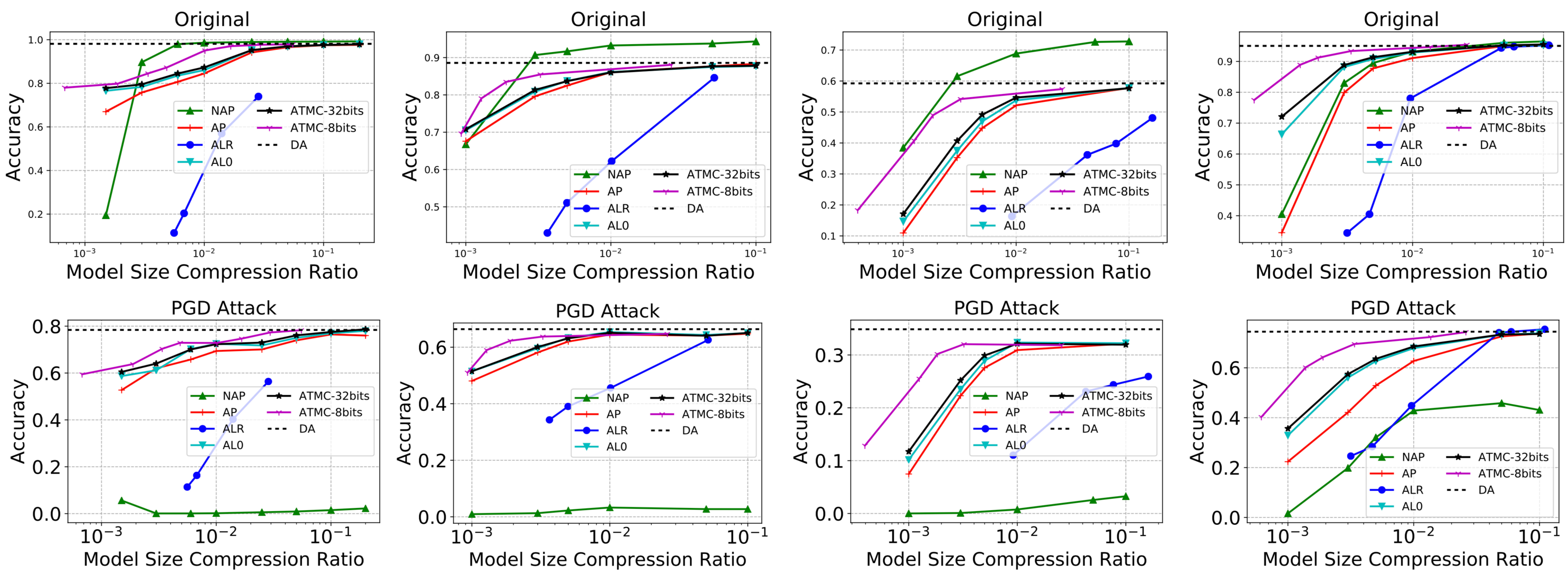
- Evaluation Metrics
  - Classification Accuracy on both benign and adversarial perturbed test set
- Model Size
  - # Non-zero elements × the bit-width of each layer
- Compression Ratio
  - the ratio between the compressed and original model sizes
- We apply PGD attack mainly for the test of adversarial robustness

Experiment: Baselines and Alternatives

- Non-adversarial Pruning (NAP) [Han et al., NIPS 15]:
  - Pure pruning method without adversarial training
- Dense Adversarial Training (DA) [Madry et al., ICLR 18]:
  - Pure adversarial training without model compression
- Adversarial Pruning (AP):
  - NAP + Adversarial Training
- Adversarial l0 Pruning (Al0)
  - L0-projection based Pruning + Adversarial Training
- Adversarial Low-Rank Decomposition (ALR)
  - Low-rank weight decomposition + Adversarial Training
- ATMC (8bits, 32 bits)
  - Our method with two bit-width settings

Experiment: Results

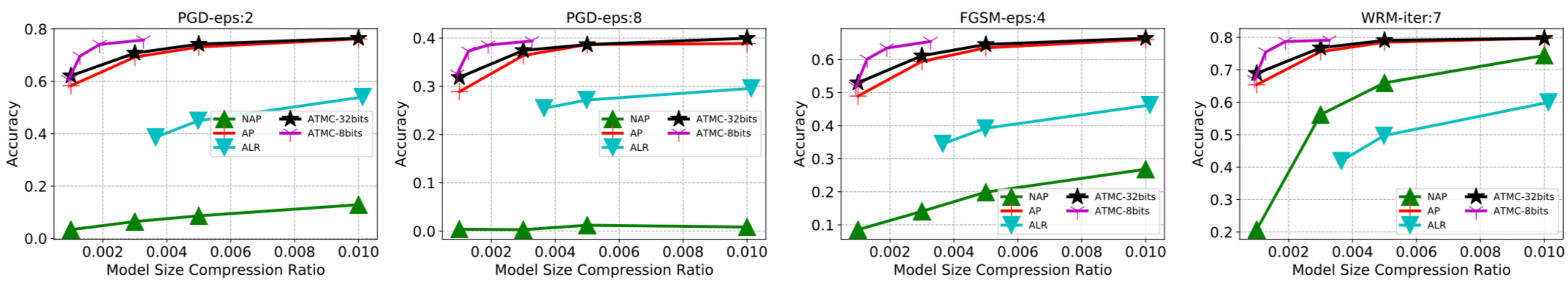
- Outstanding Performance on Trade-off between Compression and robustness for ATMC



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

Experimental results

- Consistent Adversarial Robustness under various attack settings
  - Different perturbation magnitude, e.g., 2, 8
  - Different adversarial attack methods, e.g., FGSM, WRM



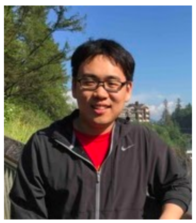
(a) PGD, perturbation=2 (b) PGD, perturbation=8 (c) FGSM, perturbation=4 (d) WRM, penalty=1.3, iteration=7

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

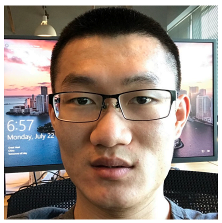
Conclusion

- ATMC, First algorithmic framework, optimizing:
  - Model Compression
  - Adversarial Robustness
- Unifying the existing compression methods:
  - Pruning
  - Factorization
  - Quantization
- Effectiveness of ATMC
  - General outstanding trade-off between model compression and robustness
  - Consistent robustness under various adversarial settings

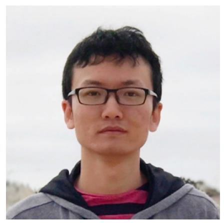
Thanks



Shupeng Gui

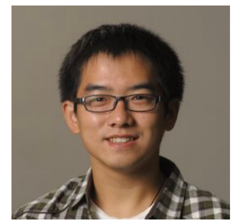


Haotao Wang

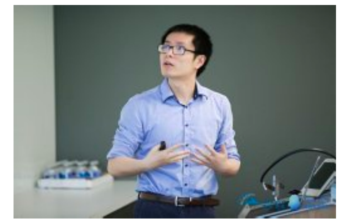


Haichuan Yang

Chen Yu



Zhangyang Wang



Ji Liu

References

- 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.