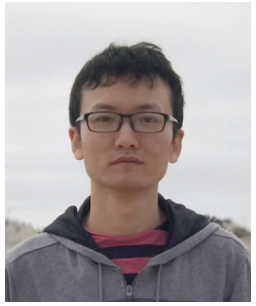
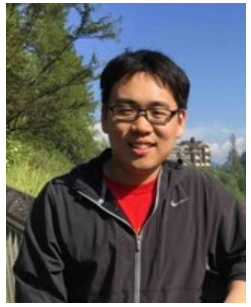
Automatic Neural Network Compression by Sparsity-Quantization Joint Learning: A Constrained Optimization-based Approach

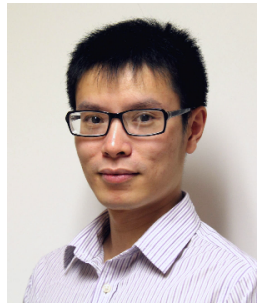


Haichuan Yang[1]



Shupeng Gui[1]

Yuhao Zhu[1]

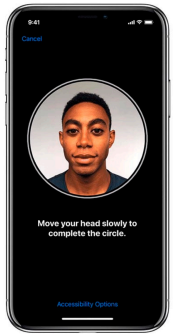


Ji Liu[2]

[1]University of Rochester

[2]Kwai Inc.

DNN on resource-constrained devices



Face verification

Speech recognition

Self driving*

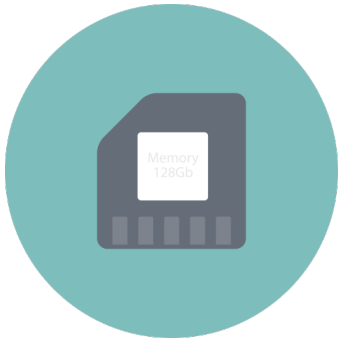
VR/AR

*Image from https://www.roboticsbusinessreview.com/unmanned/consumer-acceptance-of-self-driving-cars-soars-study-says/

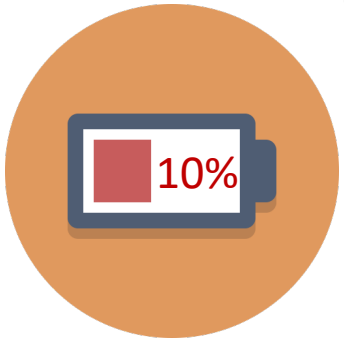
DNN Compression: maximize accuracy within resource constraint

Accuracy (Recall, IoU, etc.)

Hardware Resources



Model size



Inference energy

Inference latency

The overall objective

$$\min_{W} \ell(W)$$

$$\text{s.t.} \sum_{i=1}^{k} b(W^{(i)}) \left\| W^{(i)} \right\|_0 \leq B$$

Bitwidth of $W^{(i)}$

Model size budget

Jointly consider pruning and quantization
Each layer can have different bitwidth / sparsity

Pros:

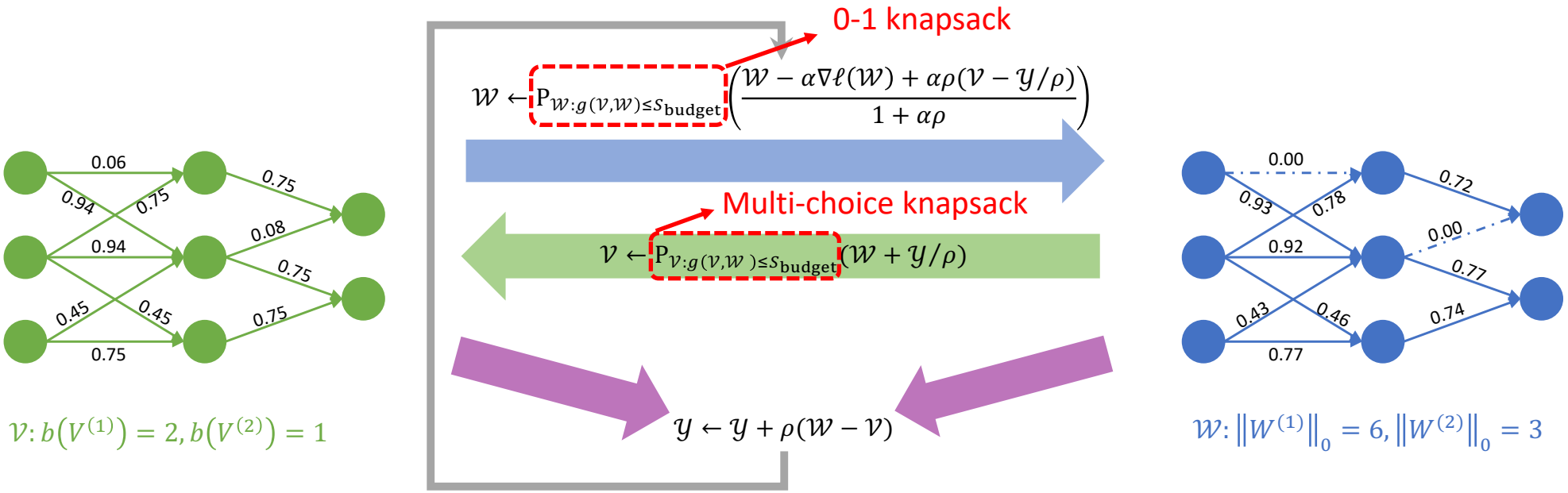
- Directly formulated from the original problem
- No introduced hyper-parameter (*no need to set layer-wise sparsity / bitwidth*)

DNN compression framework

$$\min_{W} \ell(W) \quad \text{s.t.} \sum_{i=1}^{k} b(W^{(i)}) \left\| W^{(i)} \right\|_0 \le B$$

$$\min_{W,V} \max_{U} \ell(W) + \langle U, W - V \rangle + \frac{\rho}{2} \|W - V\|^2$$
$$\text{s.t.} \sum_{i=1}^{k} b(V^{(i)}) \left\| W^{(i)} \right\|_0 \le B$$

Alternatively update W, V, and U



$\mathcal{V}: b(V^{(1)}) = 2, b(V^{(2)}) = 1$

$\mathcal{W}: \|W^{(1)}\|_0 = 6, \|W^{(2)}\|_0 = 3$

Experiment Results

- Table: Comparison across different compression methods on ImageNet.

Model	Method	Automated	Pruning	Quantization	NZ%	Ave. bits	Comp. rate	Acc.-1↓	Acc.-5↓
MobileNet	Uniform Baseline [17]	✗	✓	✗	61%	-	1.6×	2.50%	1.70%
	Uniform Baseline [17]	✗	✓	✓	61%	8	6.6×	4.10%	2.90%
	Deep Compression [11]	✗	✗	✓	-	2	16×	33.28%	25.59%
	HAQ [42]	✓	✗	✓	-	2	16×	13.76%	8.03%
	Ours	✓	✗	✓	-	2	16×	7.10%	4.40%
	Deep Compression [11]	✗	✗	✓	-	3	10.7×	4.97%	3.05%
	HAQ [42]	✓	✗	✓	-	3	10.7×	3.24%	1.69%
	Ours	✓	✗	✓	-	3	10.7×	**1.19%**	**0.76%**
	Ours	✓	✓	✓	42%	2.8	**26.7**×	4.41%	2.61%
AlexNet	Constraint-Aware [3]	✓	✓	✗	4.9%	-	20×	2.57%	-
	Deep Compression [11]	✗	✓	✓	11%	5.4	54×	0.00%	-0.03%
	CLIP-Q [41]	✓	✓	✓	8%	3.3	119×	-0.70%	-
	Ours	✓	✓	✓	7.4%	3.7	118×	**-1.00%**	**-1.15%**
	Ye et al. [53]	✗	✓	✓	4%	4.1	**210**×	0.10%	-
	Ours	✓	✓	✓	5%	3.1	205×	-0.08%	-0.56%
MnasNet	Fixed-Bitwidth	✓	✓	✓	50%	4	16×	3.14%	1.86%
	Ours	✓	✓	✓	50%	3.7	17.1×	**1.66%**	**0.92%**
	Ours	✓	✓	✓	30%	3.0	**35.6**×	5.82%	3.23%
ProxylessNAS-mobile	Fixed-Bitwidth	✓	✓	✓	50%	4	16×	3.17%	1.73%
	Ours	✓	✓	✓	51%	3.8	16.8×	**2.13%**	**1.16%**
	Ours	✓	✓	✓	31%	2.9	**35.6**×	5.21%	2.84%

Table: Comparison results on LeNet-5@MNISTs.

Method	Automated	NZ%	Avg. bits	Comp. rate	Acc.↓
Deep Compression [11]	✗	8.3%	5.3	70×	0.1%
BC-GNJ [33]	✗	0.9%	5	573×	0.1%
BC-GHS [33]	✗	0.6%	5	771×	0.1%
Ye et al. [53]	✗	0.6%	2.8	1,910×	0.1%
Ours	✓	1.0%	1.46	**2,120**×	**0.0%**

Table: Comparison results on CIFAR-10.

Model	Method	Pruning	Quantization	NZ%	Ave. bits	Comp. rate	Acc.↓
ResNet-20	ReLeQ [50]	✗	✓	-	2.8	11.4×	0.12%
	Ours	✗	✓	-	2	16×	**0.00%**
	Ours	✓	✓	46%	1.9	**35.4**×	0.14%
ResNet-50	AMC [14]	✓	✗	60%	-	1.7×	-0.11%
	Ours	✓	✗	50%	-	2×	**-1.51%**
	Ours	✓	✓	4.2%	1.7	462×	-1.25%
	Ours	✓	✓	3.1%	1.9	565×	-0.90%
	Ours	✓	✓	2.2%	1.8	**836**×	0.00%

Summary

- We automatically allocate the layer-wise sparsity and bit-width and jointly prune and quantize a DNN with the given model size budget.