Finding the Task-Optimal Low-Bit Sub-Distribution in Deep Neural Networks

Runpei Dong, Zhanhong Tan, Mengdi Wu, Linfeng Zhang and Kaisheng Ma

Xi'an Jiaotong University & Tsinghua University – ArChip Lab

International Conference on Machine Learning (ICML)
July 17th – 23rd, 2022

Summary – Main Contributions

ArChip-Lab
Algorithm ARchitecture Chip

A novel quantization method: DGMS

- The task-optimal latent low-bit sub-distribution guiding the quantization
- Trainable distributions and weights using a self-adaptive and end-to-end fashion

Promising transfer ability of the found sub-distribution

- Domain-invariant and model-inherent sub-distribution

Remarkable compression and generalization performance

- Negligible accuracy loss for 4-bit model on classification and object detection

An efficient TVM-based deployment flow Q-SIMD for DGMS

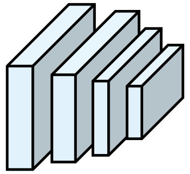
- Up to 7.46X speedup on mobile CPUs

Motivations – Quantization Problem

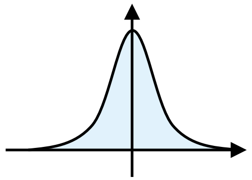
Quantization Target

- Eliminating the representative redundancy via shortened bit-width (less memory, I/O)
- Reduce computational cost

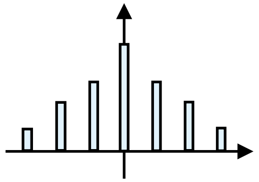
Full-precision DNN



Weight Distribution



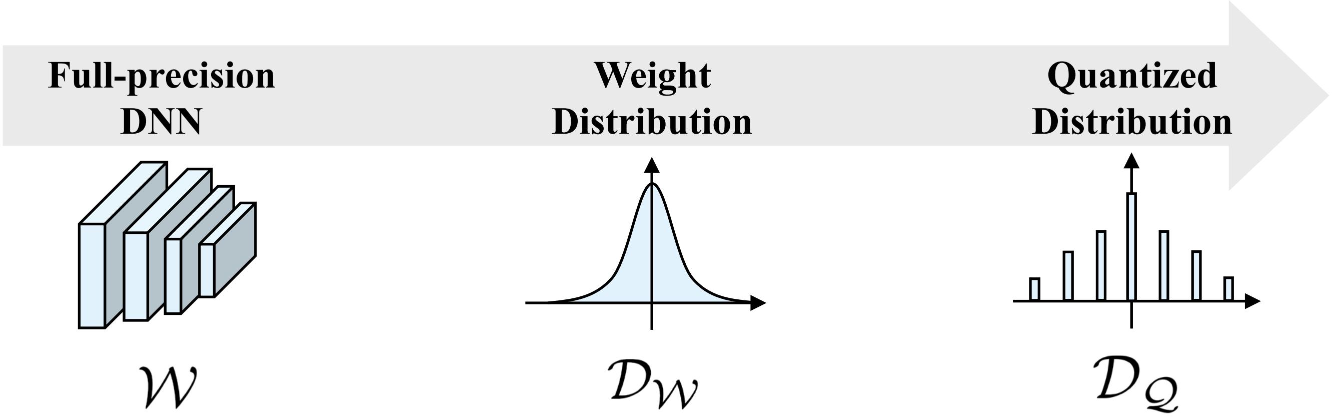
Quantized Distribution



Motivations – Quantization Problem

Quantization Target

- Eliminating the representative redundancy via shortened bit-width (less memory, I/O)
- Reduce computational cost



Projection Definition

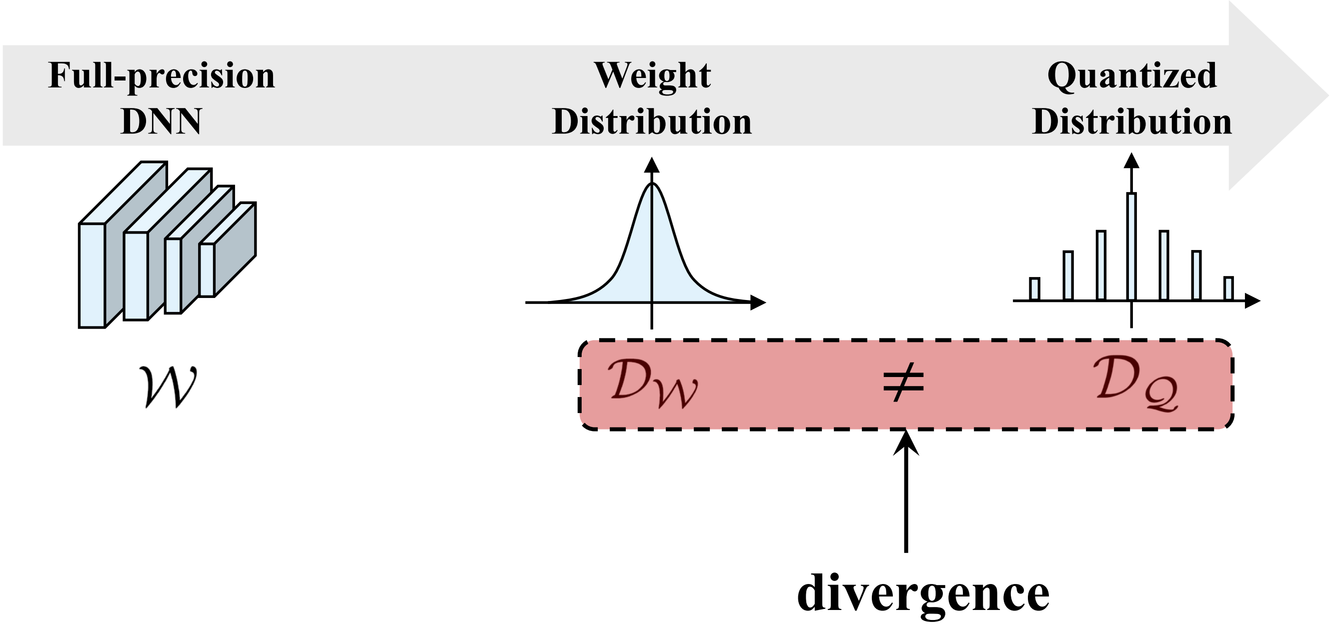
- Projection: $Q : \mathcal{W} \in \mathbb{R} \rightarrow \mathcal{Q} = \{q_0, q_1, \ldots, q_K\}$
  - $\mathcal{W} \sim \mathcal{D}_{\mathcal{W}}$: real-valued FP32 preimage
  - $\mathcal{Q} \sim \mathcal{D}_{\mathcal{Q}}$: compressed discrete representation

Motivations – Quantization Problem

ArChip-Lab
Algorithm ARchitecture Chip

Quantization Target

- Eliminating the representative redundancy via shortened bit-width (less memory, I/O)
- Reduce computational cost



Projection Definition

- Projection: $Q : \mathcal{W} \in \mathbb{R} \rightarrow \mathcal{Q} = \{q_0, q_1, \ldots, q_K\}$
  - $\mathcal{W} \sim \mathcal{D}_{\mathcal{W}}$: real-valued FP32 preimage
  - $\mathcal{Q} \sim \mathcal{D}_{\mathcal{Q}}$: compressed discrete representation

Distribution Divergence Problem

- $\mathcal{D}_{\mathcal{Q}}$ suffers a distributional divergence from the preimage $\mathcal{D}_{\mathcal{W}}$

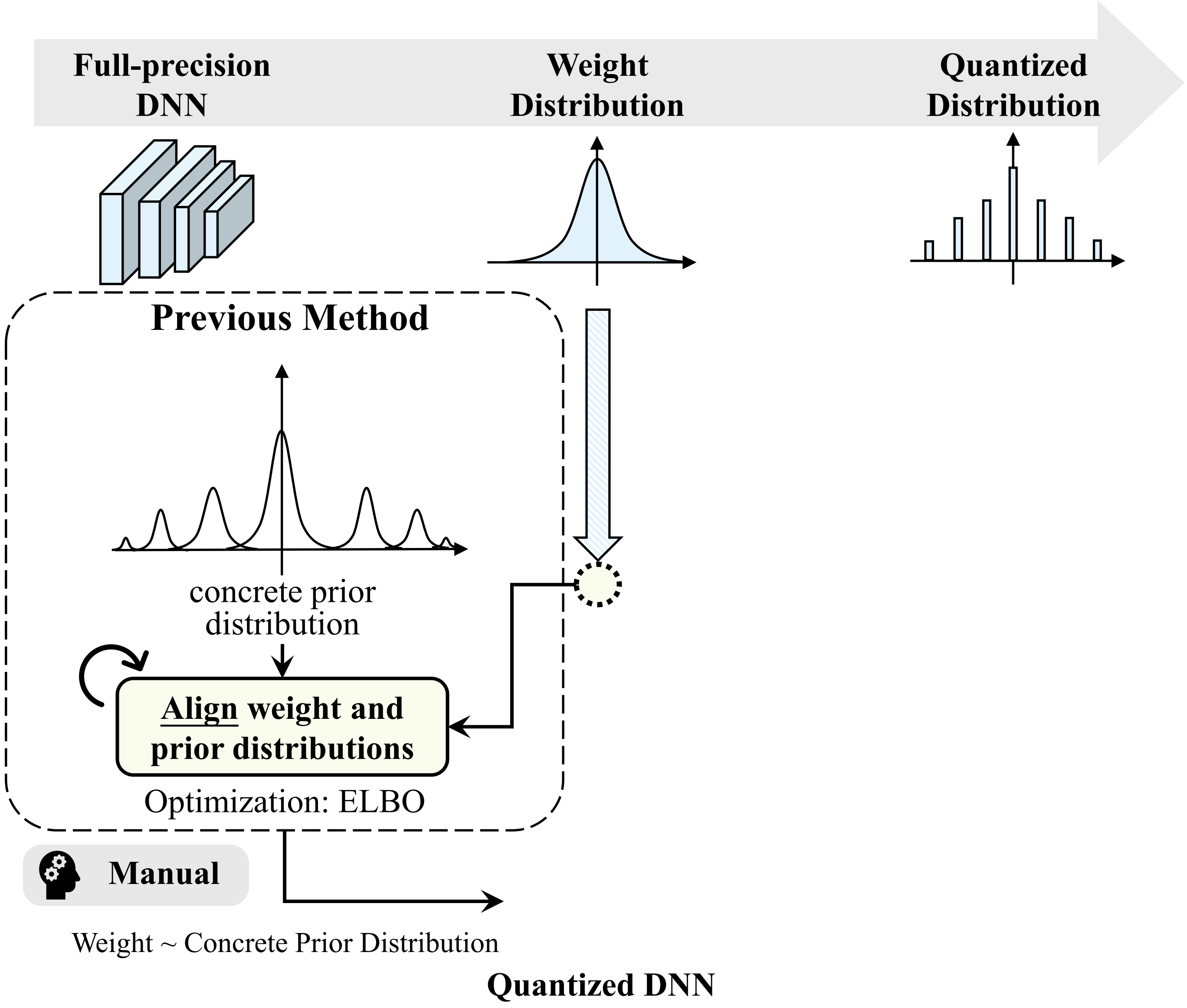
Motivations – Previous Methods Issues

Discontinuous-Mapping

- Rounding operation
- Require fixed configurations (e.g., centroids & stepsize)
- Inaccurate pseudo-gradient with STE
- Ignore global statistical information

Continuous-Mapping

- Adaptive-mapping
- Require concrete prior distribution $\mathcal{D}_{\mathcal{P}}$
- Require non-trivial & non-optimal manual configurations (e.g., $\mathcal{D}_{\mathcal{P}}$ hyperparameters)
- Optimize ELBO target
- Large memory footprint for MCMC
- Unapplicable to advanced DNN



Motivations – Previous Methods Issues

ArChip-Lab
Algorithm ARchitecture Chip

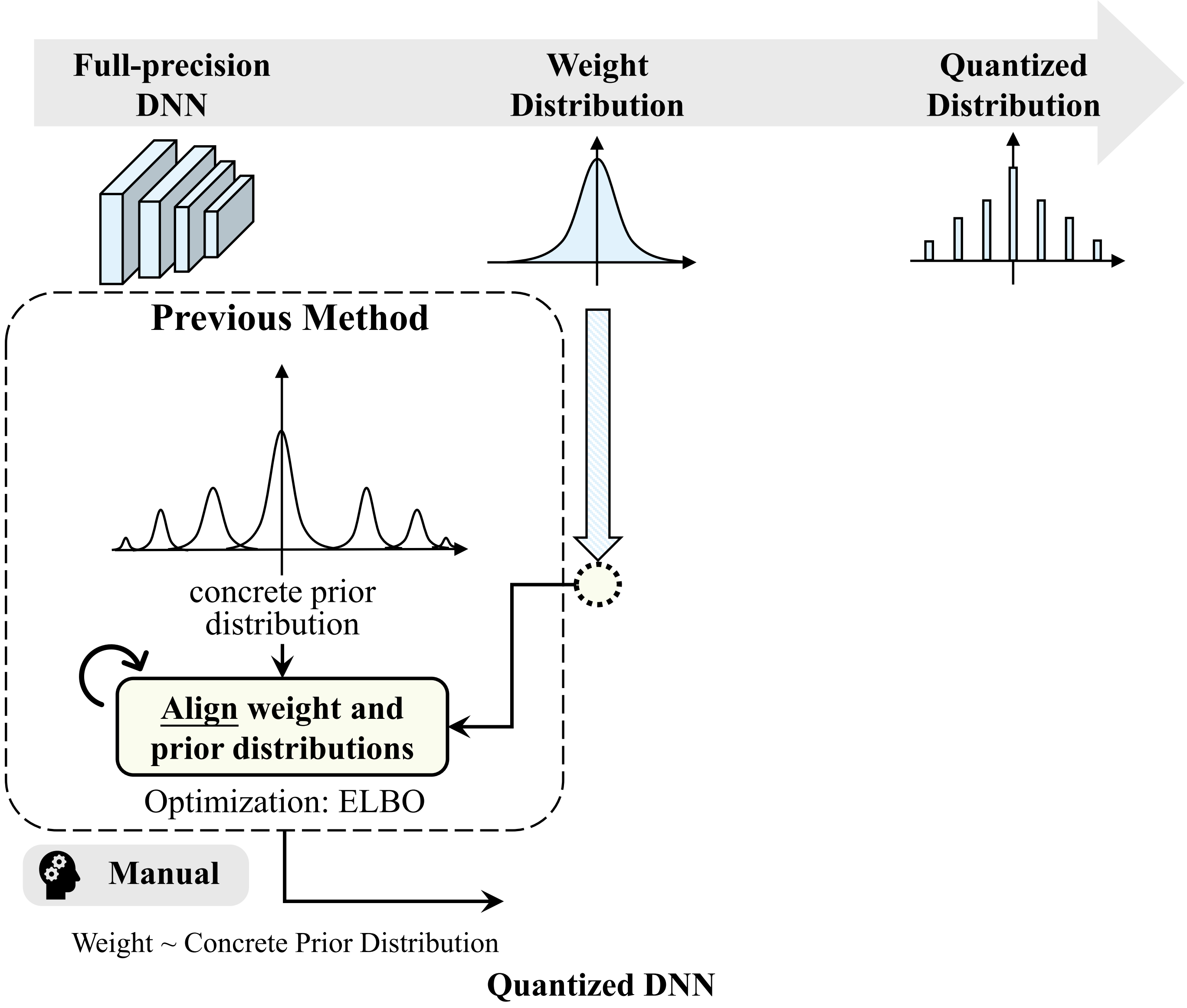
Discontinuous-Mapping

- Rounding operation
- Require fixed configurations (e.g., centroids & stepsize)
- Inaccurate pseudo-gradient with STE
- Ignore global statistical information

Continuous-Mapping

- Adaptive-mapping
- Require concrete prior distribution $\mathcal{D}_{\mathcal{P}}$
- Require non-trivial & non-optimal manual configurations (e.g., $\mathcal{D}_{\mathcal{P}}$ hyperparameters)
- Optimize ELBO target
- Large memory footprint for MCMC
- Unapplicable to advanced DNN

Question: *Whether there exists a task-optimal latent **sub-distribution** for quantization as the **sub-network** proposed by the "Lottery Ticket Hypothesis"?*



Our Method

Key Idea Comparison

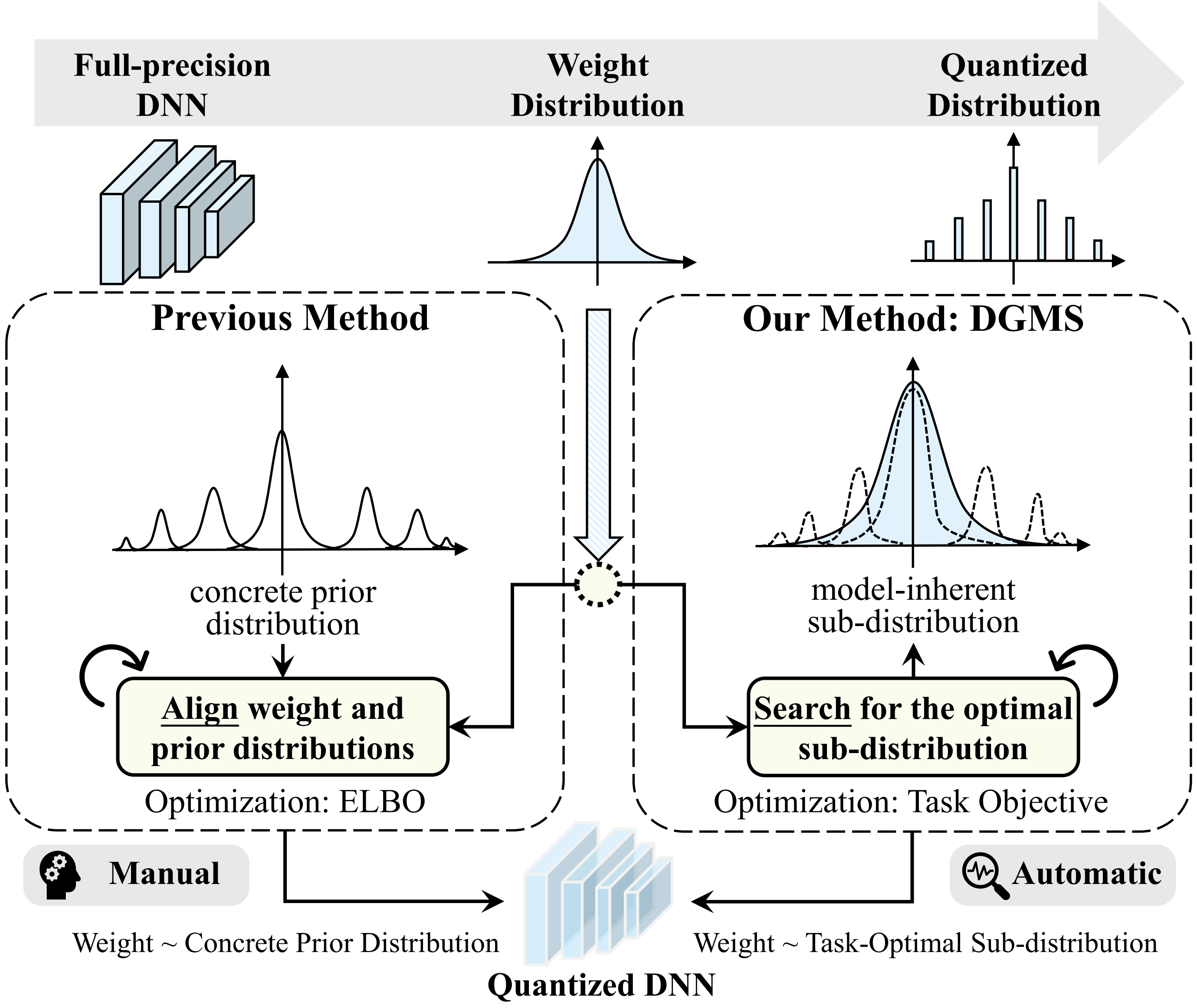
- Automatic searching
- Model-inherent latent sub-distribution
- Non-manual quantization configs
- Directly optimize task-objective
- Task-optimal quantization

What is sub-distribution?

Definition 1 (Sub-Distribution)

Given the preimage $\mathcal{W} \sim \mathcal{D}_\mathcal{W}$ and quantized data $\mathcal{Q} \sim \mathcal{D}_\mathcal{Q}$, the sub-distribution $\mathcal{D}_\mathcal{S}$ is defined as an estimation for $\mathcal{D}_\mathcal{W}$ (i.e., $\mathcal{D}_\mathcal{S} \approxeq \mathcal{D}_\mathcal{W}$ where $\approxeq$ denotes approximate equivalence), and under a parameter limitation τ, $\mathcal{D}_\mathcal{S}$ is approximately equivalent to $\mathcal{D}_\mathcal{Q}$ (i.e., $\lim_{\tau\to 0}\mathcal{D}_\mathcal{S} \approxeq \mathcal{D}_\mathcal{Q}$).

Key idea: *The model-inherent sub-distribution serves as a distributional bridge and automatically evolves*



Experimental Validation

Image Classification & Object Detection

Model	#Params	Bits	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	dog	horse	mbike	person	plant	sheep	sofa	table	train	tv
SSDLite -MBV2	3.39M	32	68.6	69.7	78.2	63.4	54.8	35.6	78.8	74.4	82.0	53.8	61.9	78.5	82.2	80.6	71.8	42.8	62.6	78.4	73.7	83.3	65.5
	2.68M	4	67.9	69.4	78.5	63.6	54.6	34.2	77.6	73.1	82.4	53.1	61.1	76.3	81.3	81.3	70.7	41.0	62.3	78.8	72.6	81.1	65.0
SSDLite -MBV3	2.39M	32	67.2	66.0	79.7	58.7	54.9	36.9	79.3	73.0	83.3	51.1	62.8	77.7	81.5	77.1	70.0	39.0	58.9	74.6	72.0	82.2	64.8
	1.26M	4	65.7	65.5	77.9	55.5	54.5	34.3	77.7	72.1	83.8	48.7	59.6	76.2	80.2	75.8	69.2	37.2	57.0	72.3	72.5	80.8	62.5

ResNet-50	Our FP32	×	FP32	76.15%	N/A
	UNIQ	×	4/32	75.09%	-0.93%
	HAQ	×	4/32 mixed	76.14%	-0.01%
	Ours	×	4/32	**76.28%**	**+0.13%**
	CLIP-Q♮	✓	3/32	73.70%	**+0.60%**
	HAQ	×	3/32 mixed	75.30%	-0.85%
	Ours	×	3/32	**75.91%**	-0.24%
	HAQ	×	2/32 mixed	70.63%	-5.52%
	Ours	×	2/32	**72.88%**	**-3.27%**
	UNIQ	×	4/8	74.37%	-1.65%
	TQT	×	4/8	74.40%	-1.00%
	HAWQV3	×	4/8 mixed	75.39%	-2.33%
	Ours	×	4/8	**76.22%**	**+0.07%**
	AutoQ	×	4/4	72.43%	-2.37%
	HAWQV3	×	4/4	74.24%	-3.48%
	Ours	×	4/4	**75.05%**	**-1.10%**

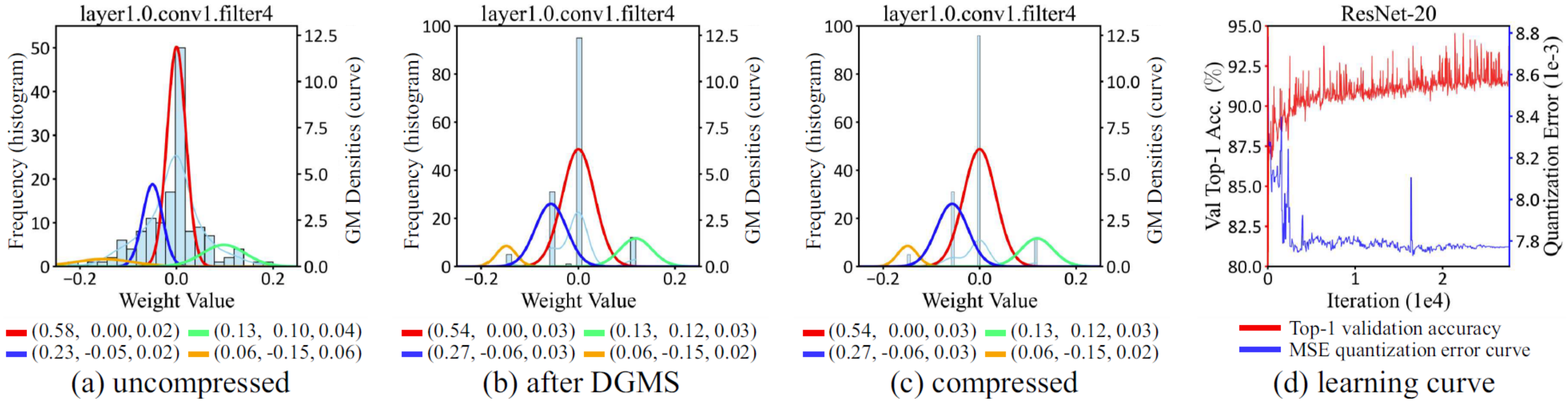
MnasNet-A1	Our FP32	×	FP32	73.51%	N/A
	ANNC♯	✓	4/32	71.80%	-1.66%
	Ours	×	4/32	**72.18%**	**-1.33%**
	Ours	×	4/8	**72.09%**	**-1.42%**
ProxylessNAS -Mobile	Our FP32	×	FP32	74.59%	N/A
	ANNC♯	✓	4/32	72.46%	-2.13%
	Ours	×	4/32	**73.85%**	**-0.74%**
	Ours	×	4/8	**73.24%**	**-1.35%**

Model	ResNet-18	ResNet-50	MnasNet-A1	ProxylessNAS	SSDLite◇
Baseline	67.09 ms	148.56 ms	18.27 ms	22.17 ms	43.95 ms
DGMS	**8.99 ms**	**34.26 ms**	**10.49 ms**	**13.34 ms**	**25.60 ms**
Speedup	7.46×	4.34×	1.74×	1.66×	1.72×

◇ SSDLite-MBV2 with MobileNetV2 as the backbone model.

Experimental Validation

Sub-distribution Evolution & Domain-invariance Study



Target		ImageNet			CUB200-2011			Stanford Cars			FGVC Aircraft		
FP32 Full-Model		69.76%			78.68%			86.58%			80.77%		
4-bit DGMS (w/o transfer)		70.25%			77.90%			86.39%			80.41%		
4-bit DGMS (w/ transfer)	Source	CUB	Cars	Air	Img	Cars	Air	Img	CUB	Air	Img	CUB	Cars
	ZERO-SHOT	34.69%	62.31%	35.09%	73.53%	74.29%	66.44%	82.28%	81.46%	71.75%	77.46%	74.97%	77.31%
	ONE-EPOCH	68.37%	69.13%	68.80%	77.70%	77.54%	77.50%	85.70%	85.84%	85.79%	79.90%	79.87%	80.14%

Finding the Task-Optimal Low-Bit Sub-Distribution in Deep Neural Networks

Runpei Dong, Zhanhong Tan, Mengdi Wu, Linfeng Zhang and Kaisheng Ma

Xi'an Jiaotong University & Tsinghua University – ArChip Lab

Thanks!

Please contact us if you have any questions.

runpei.dong@outlook.com

ArChip-Lab
Algorithm ARchitecture Chip

ICML
International Conference On Machine Learning