Microsoft

Diversifying Sample Generation for Accurate Data-Free Quantization

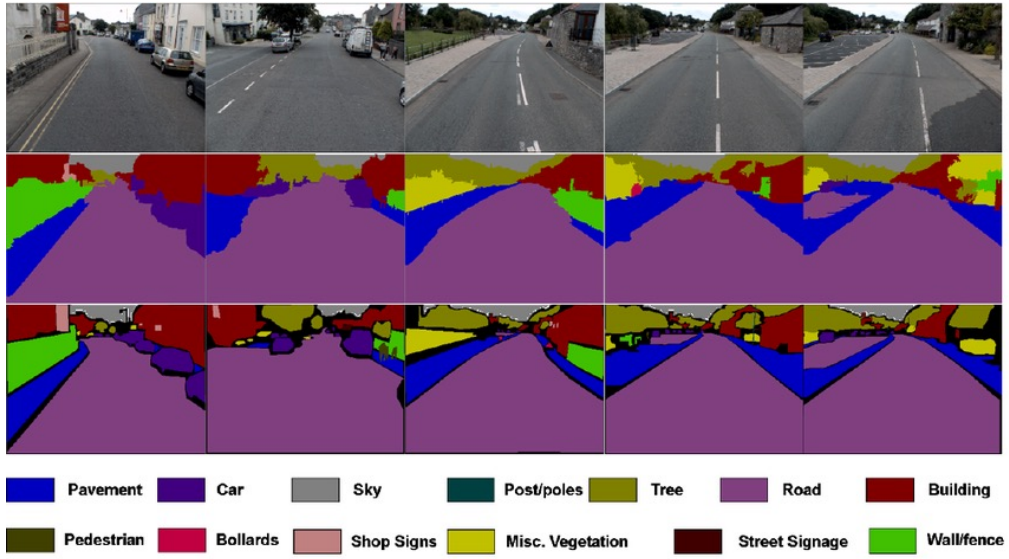
Xiangguo Zhang*[1], Haotong Qin*[1], Yifu Ding[1], Ruihao Gong[3, 4],
Qinghua Yan[1], Renshuai Tao[1], Yuhang Li[2], Fengwei Yu[3, 4], Xianglong Liu[1†]
[1]Beihang University [2]Yale University [3]SenseTime Research [4]Shanghai AI Laboratory
{xiangguozhang, zjdyf, yanqh, rstao}@buaa.edu.cn, yuhang.li@yale.edu,
{qinhaotong, xlliu}@nlsde.buaa.edu.cn, {gongruihao, yufengwei}@sensetime.com

Haotong Qin
Beihang University

Background: neural networks

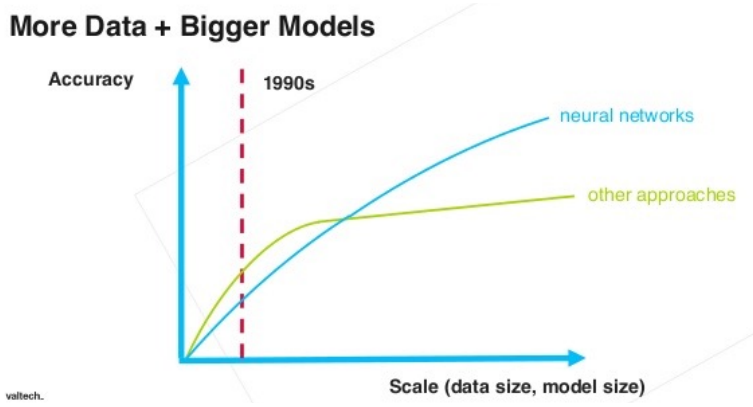
- **Image**
  - Classification
  - Detection
  - Localization
  - Segmentation
- **Audio**
  - Speech recognition
  - Language understanding

…



Deployment Challenges

- Millions of parameters
- Limited computing resources
- Short response time



Background: model compression

What is the meaning of model compression?

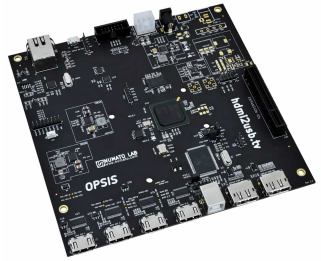
Memory Usage Compression

Inferred Acceleration

Easy to Deploy

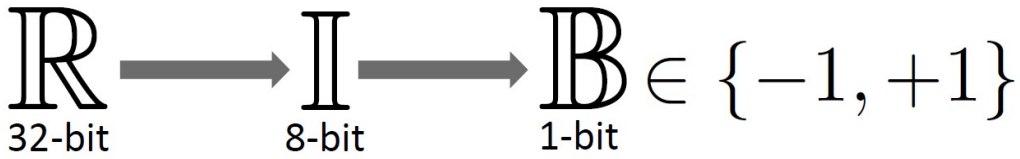
Easy to Transfer

Compressed models can be more widely deployed and updated more easily.



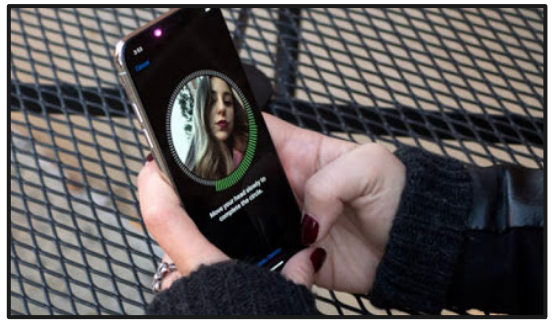
Quantization: overview

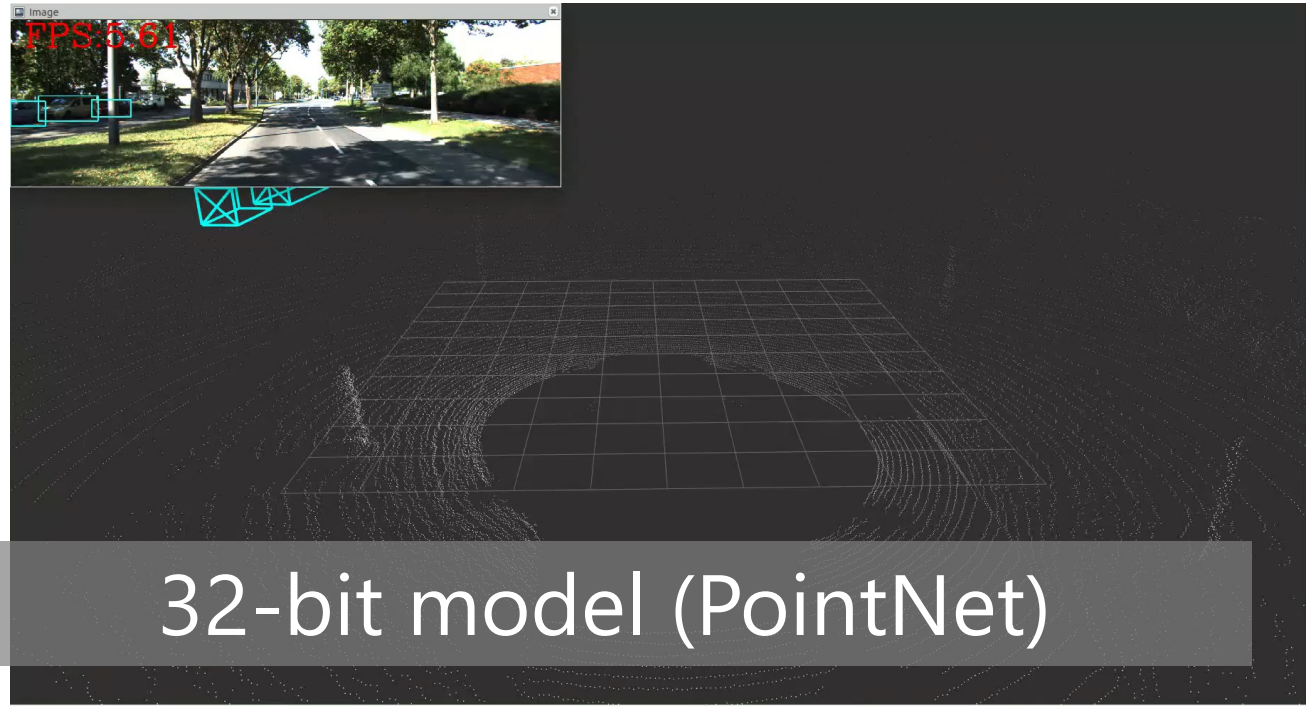
Quantization and Binarization

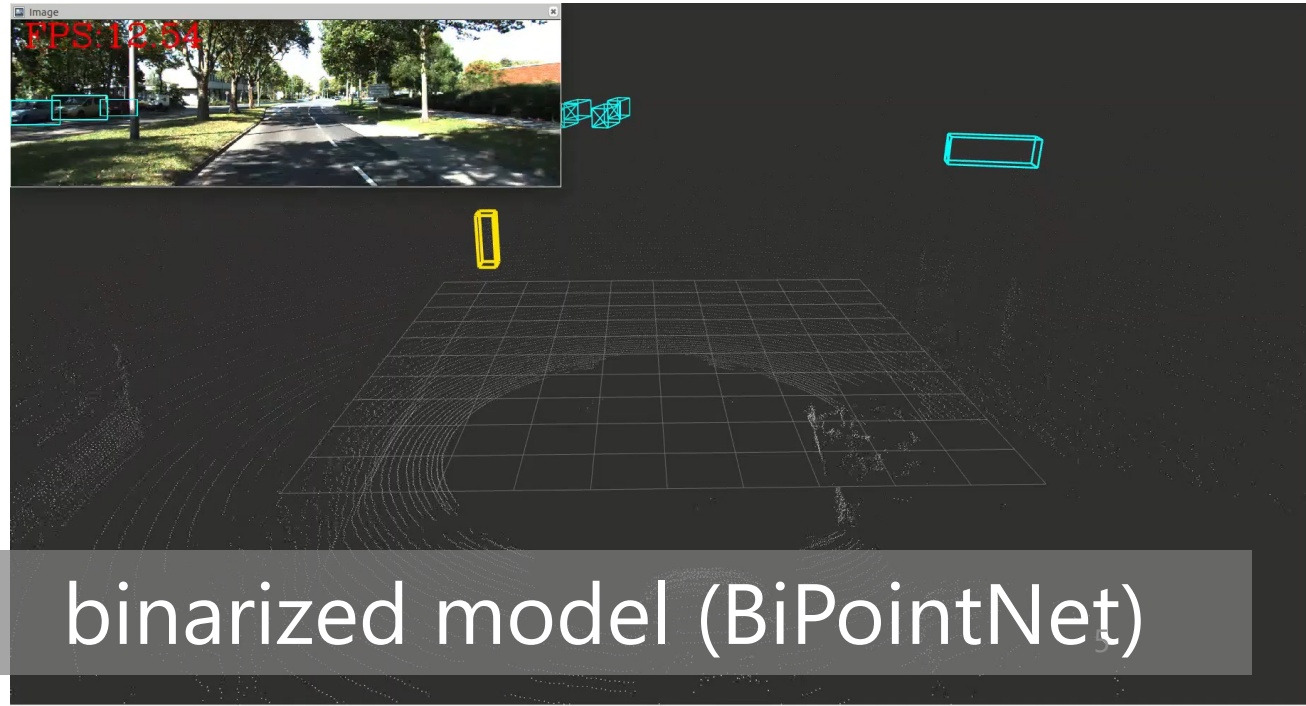


Full-Precision Neural Networks	Massive Parameters	Complex Computation	High Power Consumption	High Accuracy
Low-Bit Quantized Neural Networks	Quantized Parameters	Efficient instructions	Low Power Consumption	Drop of Accuracy

Quantization: applications





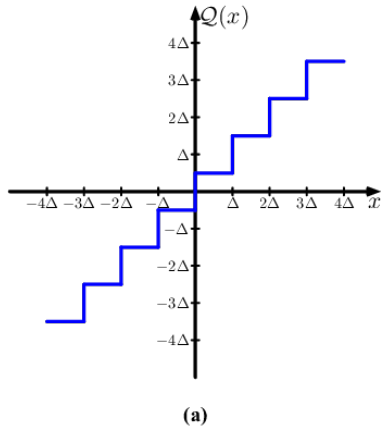


BiPointNet: Binary Neural Network for Point Clouds, Haotong Qin, et al., *ICLR 2021*

Quantization: formulation

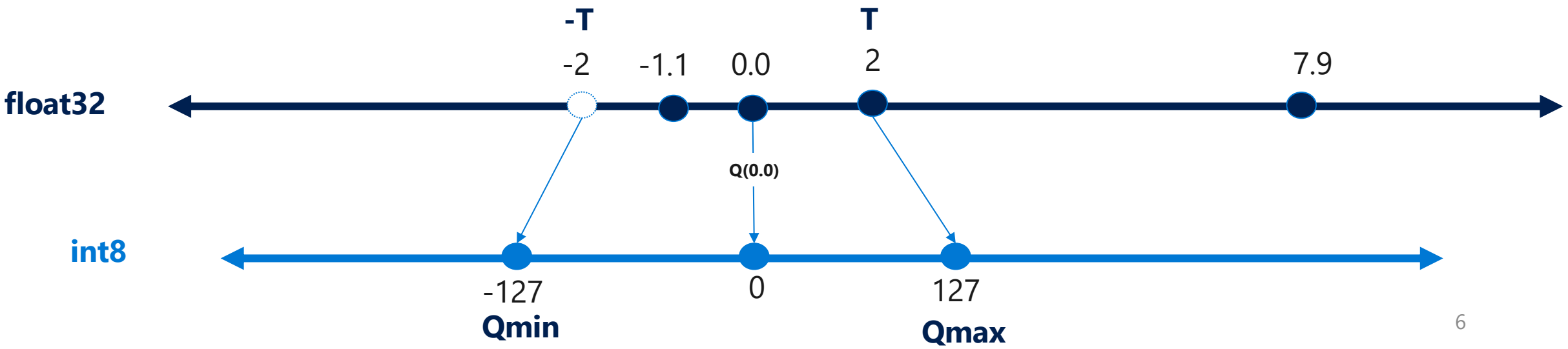
$$Q(x) = \text{round}\left(\frac{x}{\Delta}\right)\Delta$$

$$\Delta = \frac{2T}{2^b - 1}$$



(a)

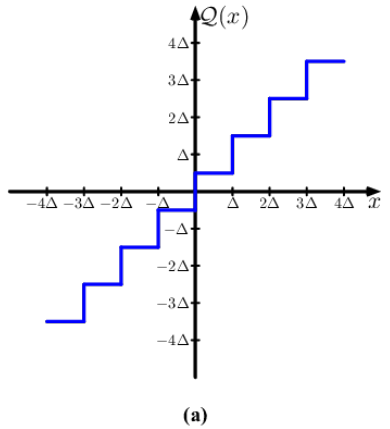
1. Network training to optimize x
2. Determining threshold T to optimize quantizer



Quantization: formulation

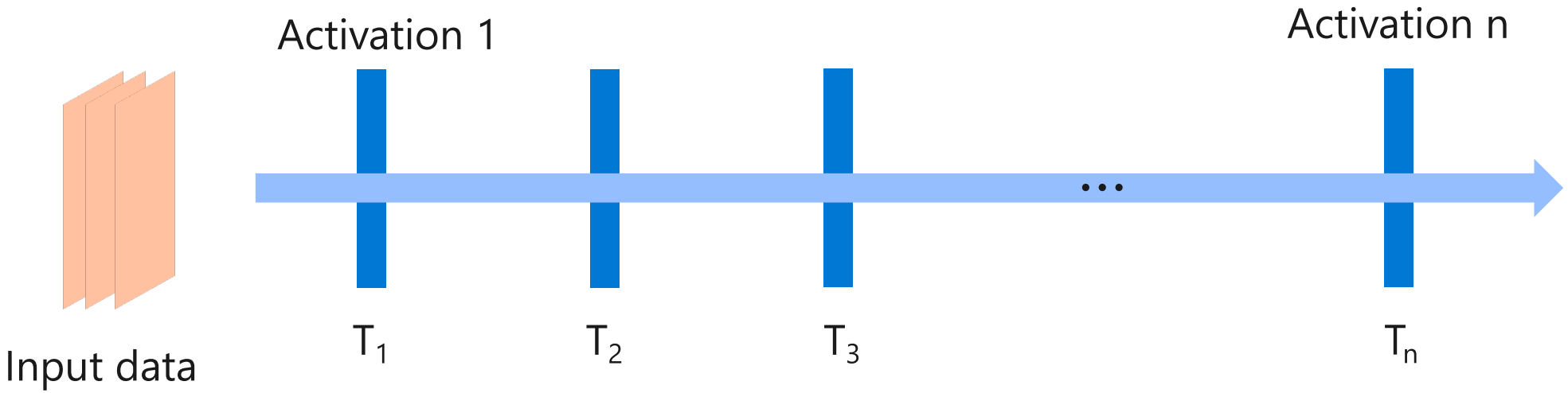
$$Q(x) = \text{round}\left(\frac{x}{\Delta}\right)\Delta$$

$$\Delta = \frac{2T}{2^b - 1}$$



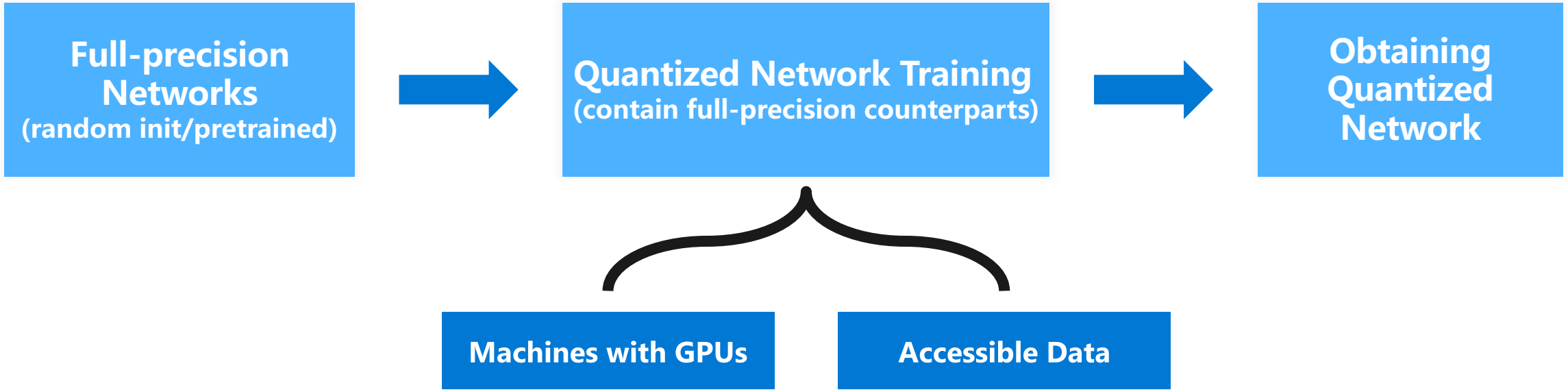
(a)

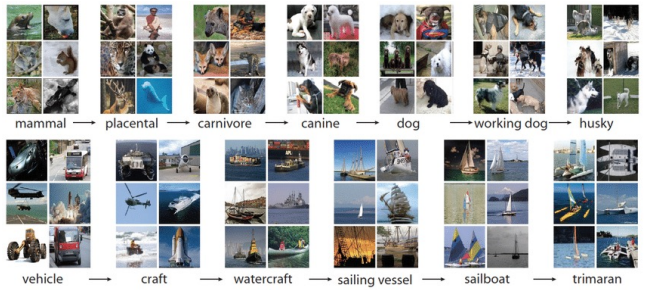
1. Network training to optimize x
2. Determining threshold T to optimize quantizer



Quantization: pipeline

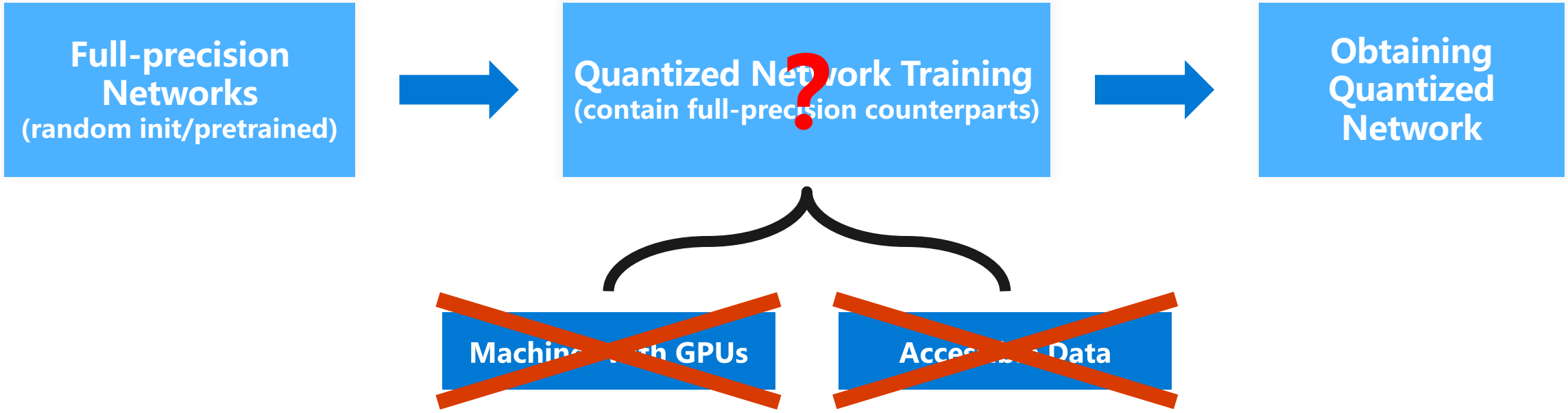
Traditional Quantization (Quantization-Aware Training):





Quantization: pipeline

Traditional Quantization (Quantization-Aware Training):

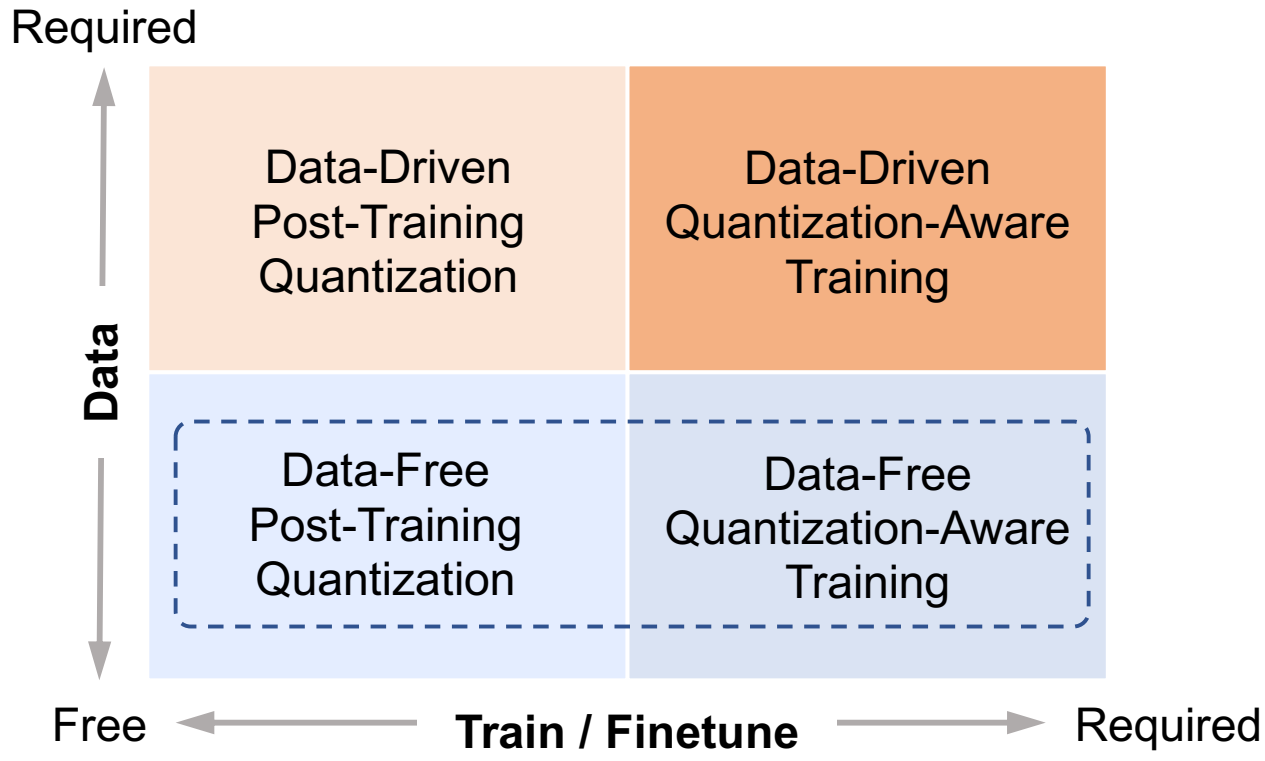


Quantization: classification

The Goal of the Paper:

No Real-Data

No Model-Training/Fine-tuning Process



Quantization in Different Scenarios

Diversifying Sample Generation for Accurate Data-Free Quantization

Xiangguo Zhang*[1], Haotong Qin*[1], Yifu Ding[1], Ruihao Gong[3, 4],
Qinghua Yan[1], Renshuai Tao[1], Yuhang Li[2], Fengwei Yu[3, 4], Xianglong Liu[1†]
[1]Beihang University [2]Yale University [3]SenseTime Research [4]Shanghai AI Laboratory
{xiangguozhang, zjdyf, yanqh, rstao}@buaa.edu.cn, yuhang.li@yale.edu,
{qinhaotong, xlliu}@nlsde.buaa.edu.cn, {gongruihao, yufengwei}@sensetime.com

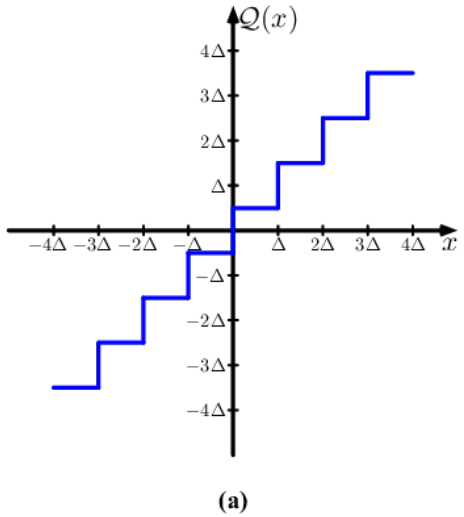
Fast, **Private**, **Accurate** data-free quantization

CVPR 2021, Oral

Data-Free Quantization: formulation

$$Q(x) = \text{round}\left(\frac{x}{\Delta}\right)\Delta$$

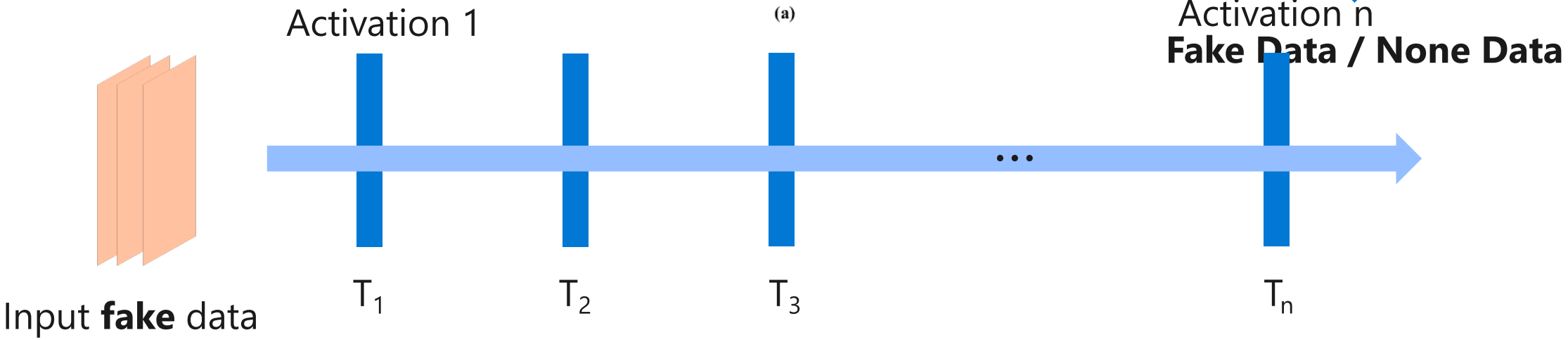
$$\Delta = \frac{2T}{2^b - 1}$$



(a)

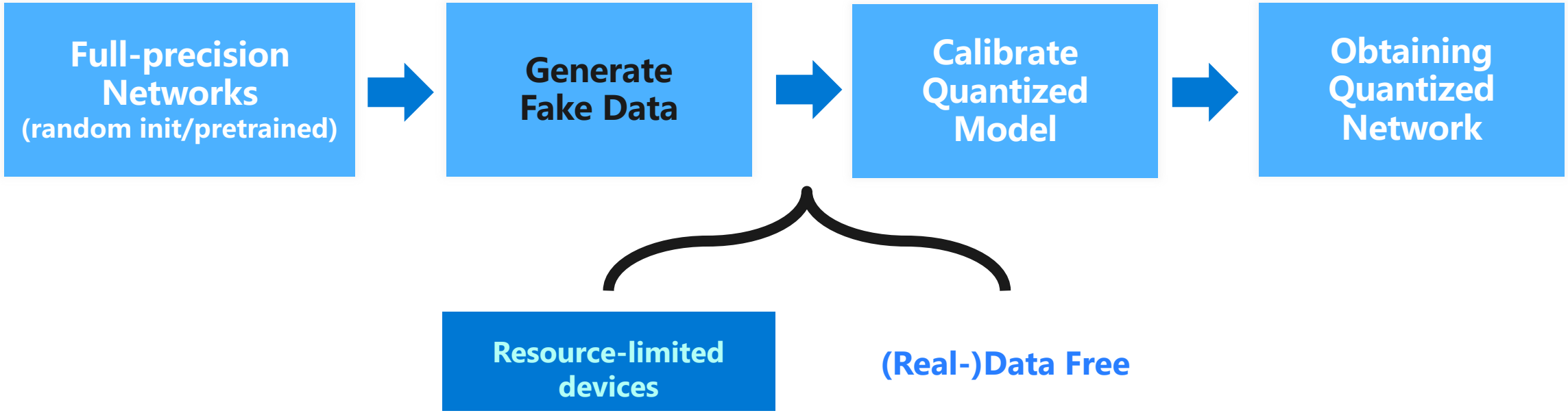
~~1. Network training to optimize x~~

2. Determining threshold T to optimize quantizer



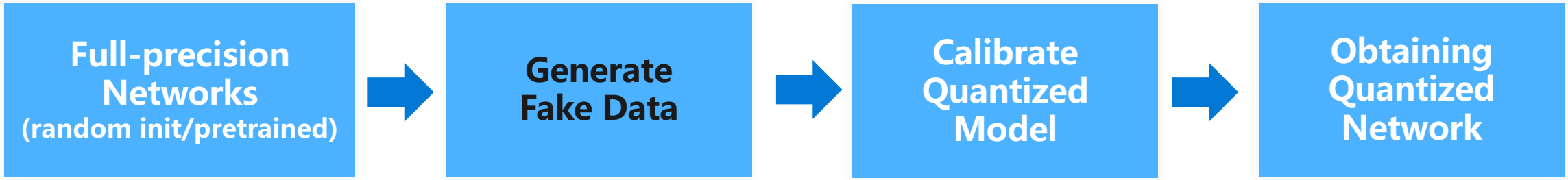
Data-Free Quantization: pipeline

Generative Data-Free Quantization: quantize network by generated fake data



Data-Free Quantization: challenge

Generative Data-Free Quantization: quantize network by generated fake data



Challenge: Significant Accuracy Drop

4-bit weight / 4-bit activation (ImageNet)

ResNet18: **71.47** to **26.04**

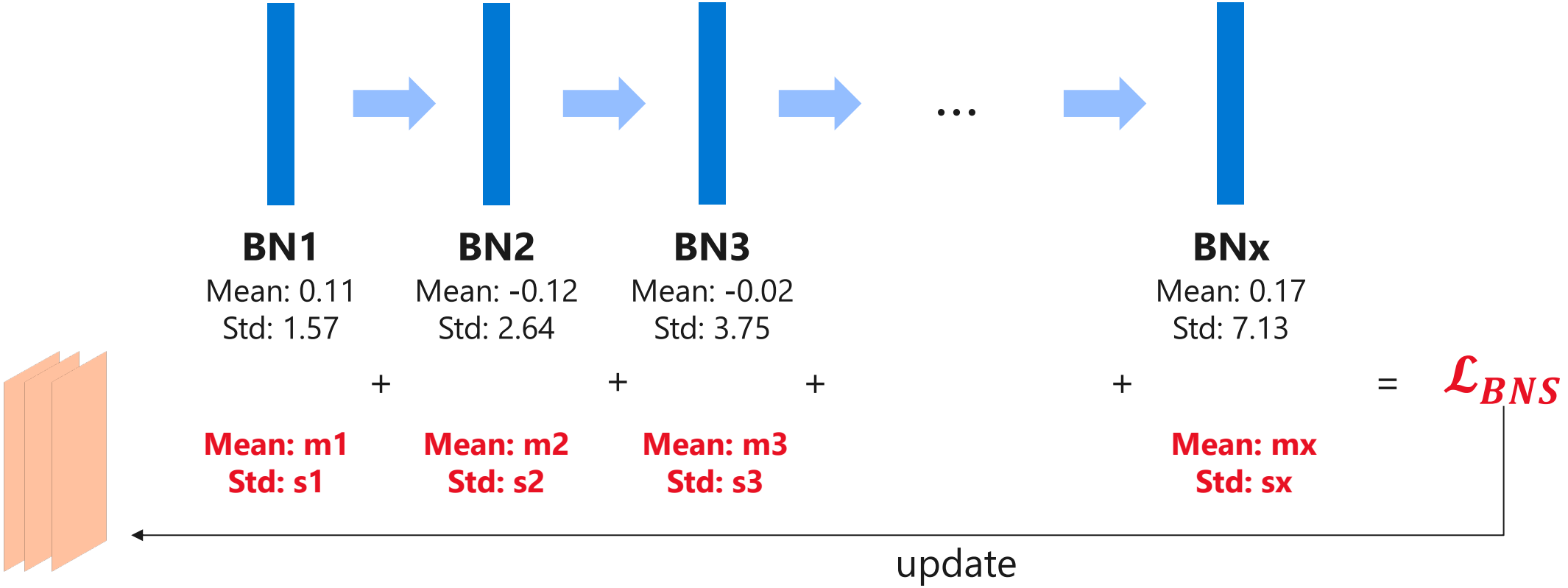
ResNet50: **77.72** to **12.73**

SqueezeNext : **69.38** to **19.15**

InceptionV3 : **78.80** to **12.00**

Diversifying Sample Generation: preliminaries

- Using **Batch Normalization Statistics** (BNS) loss generate images



$$\min_{x^r} \sum_{i=0}^{L} \|\tilde{\mu}_i^r - \mu_i\|_2^2 + \|\tilde{\sigma}_i^r - \sigma_i\|_2^2,$$

Diversifying Sample Generation: observation

- Fake (generated) data are different from real data!

Visually

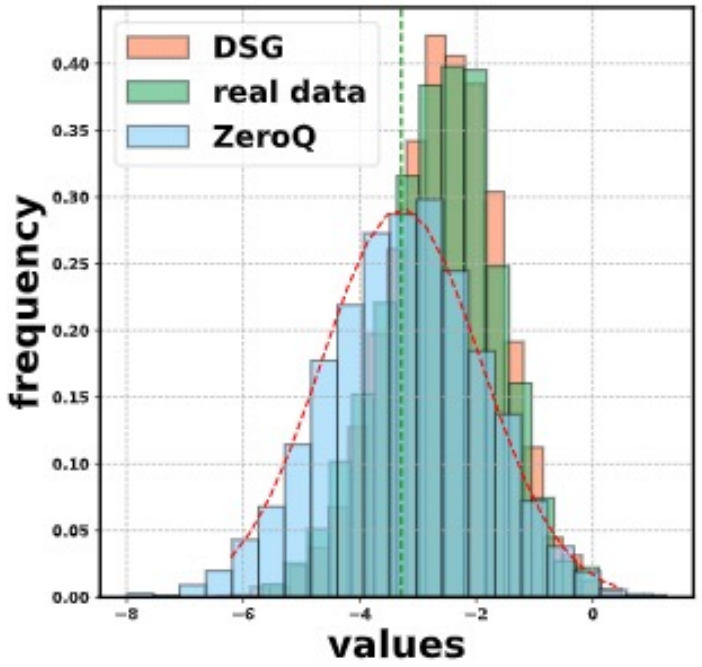
Real Data
Clear
Semantic

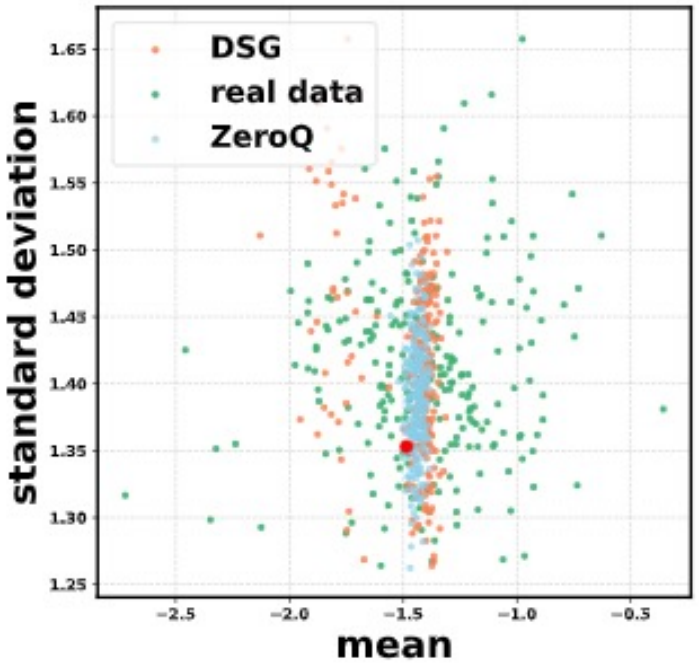
Fake Data
Fuzzy
Non-semantic

Diversifying Sample Generation: homogenization

- Fake (generated) data are different from real data!

Statistics





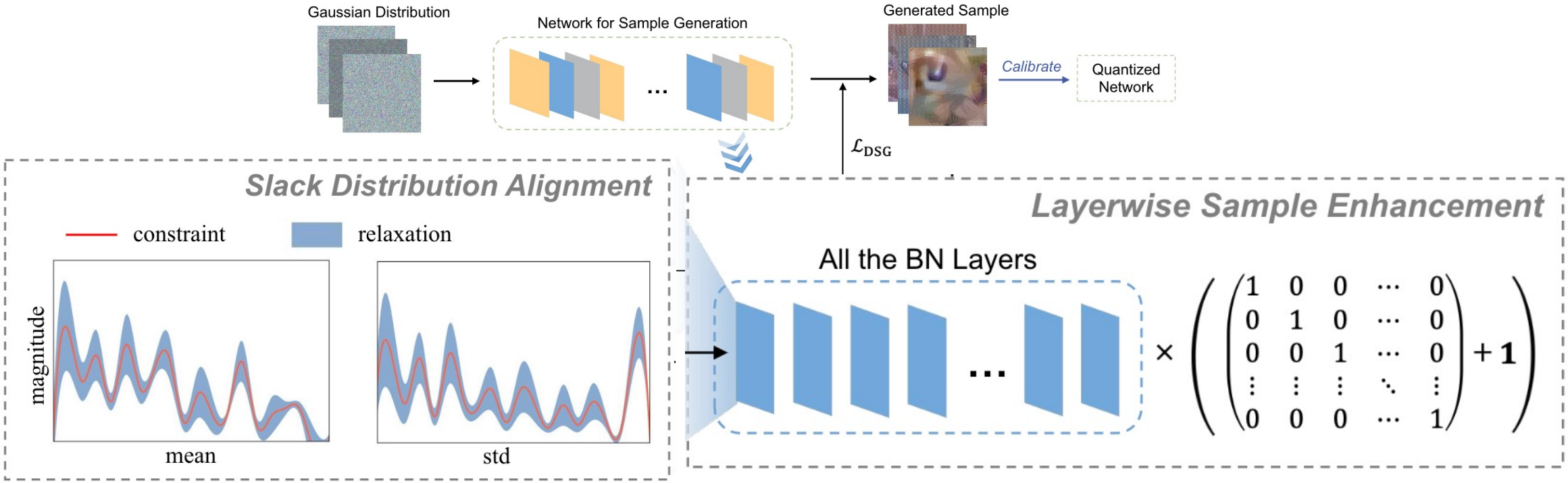
Homogenization

Distribution Level

Sample Level

Diversifying Sample Generation: overview

- We proposed **Diversifying Sample Generation (DSG)** to address the homogenization.



Diversifying Sample Generation

Slack Distribution Alignment (SDA)

- BNS

$$\min_{x^r} \sum_{i=0}^{L} \|\tilde{\mu}_i^r - \mu_i\|_2^2 + \|\tilde{\sigma}_i^r - \sigma_i\|_2^2,$$

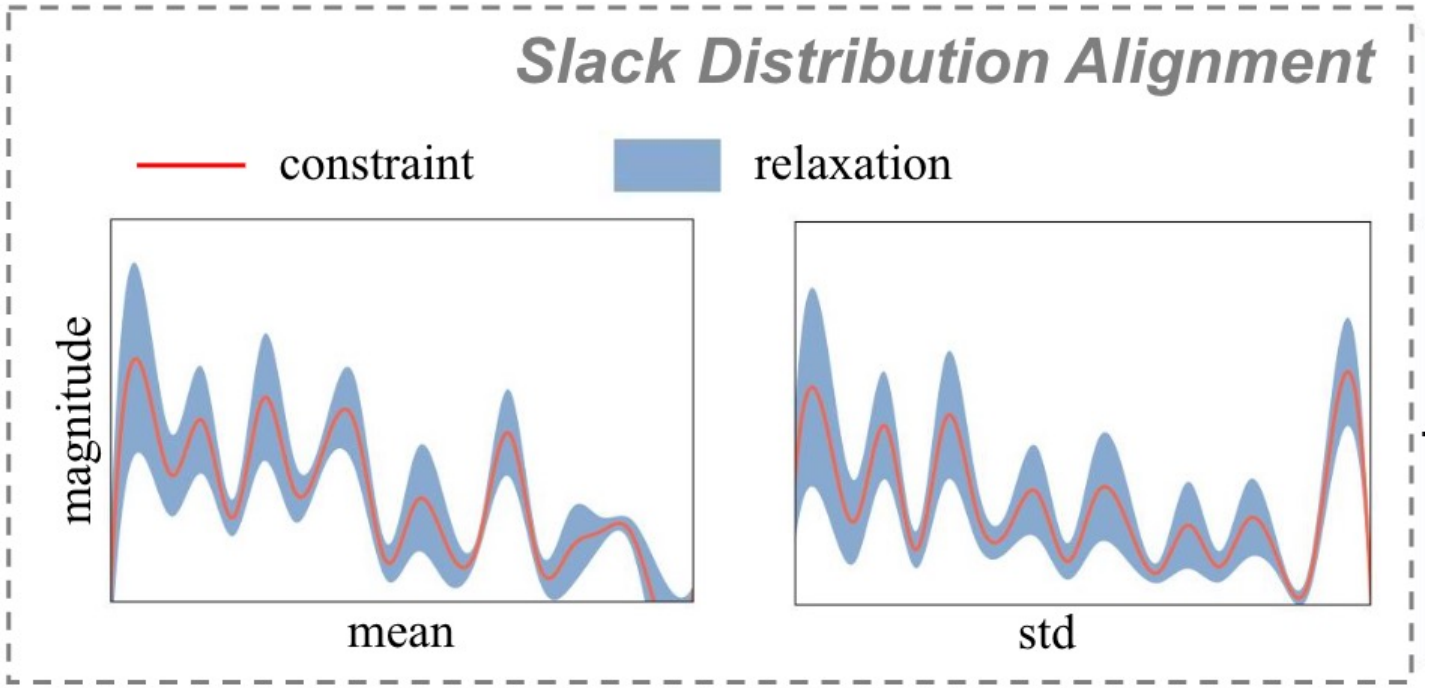
- **SDA (Ours)**

$$l_{\mathrm{SDA}_i} = \|\max\left(|\tilde{\boldsymbol{\mu}}_i^s - \boldsymbol{\mu}_i| - \delta_i, 0\right)\|_2^2 + \|\max\left(|\tilde{\boldsymbol{\sigma}}_i^s - \boldsymbol{\sigma}_i| - \gamma_i, 0\right)\|_2^2$$

Diversifying Sample Generation

Slack Distribution Alignment (SDA)

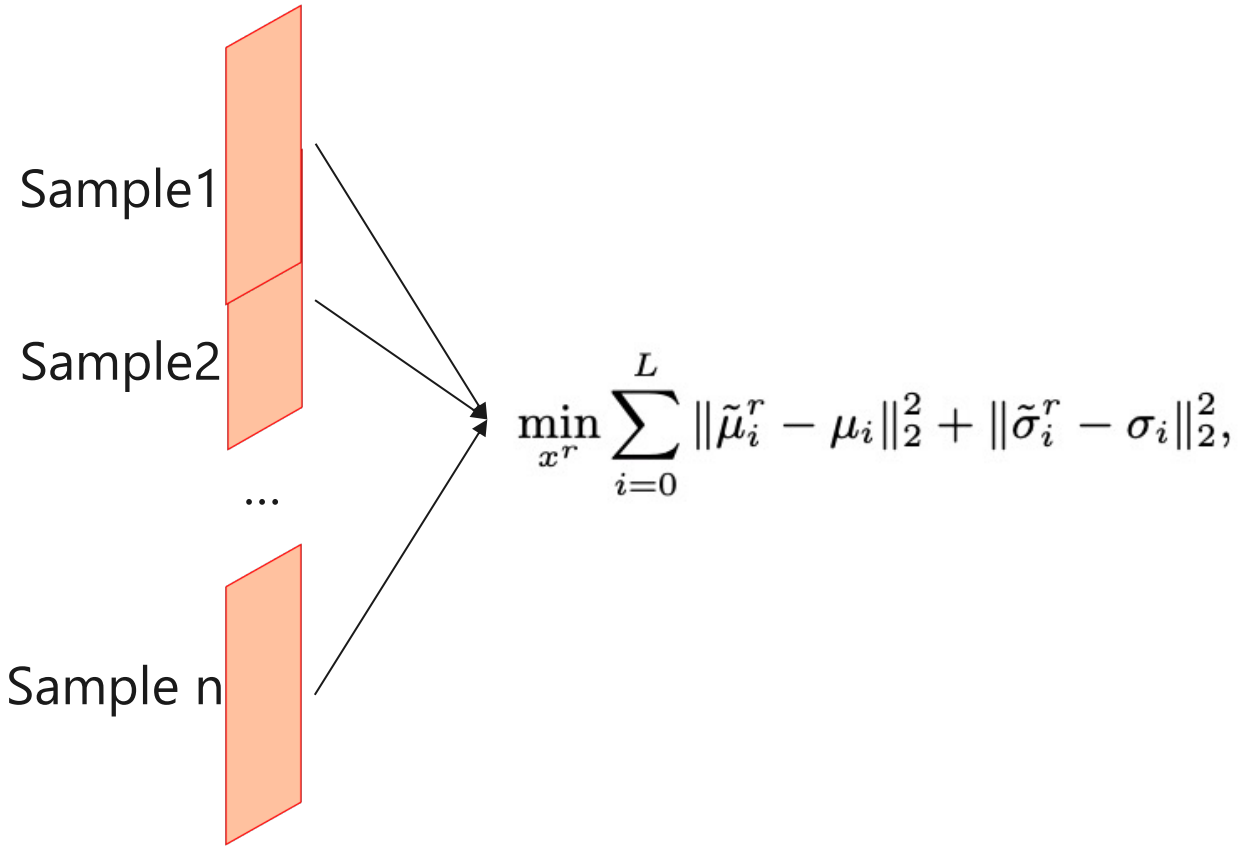
- SDA

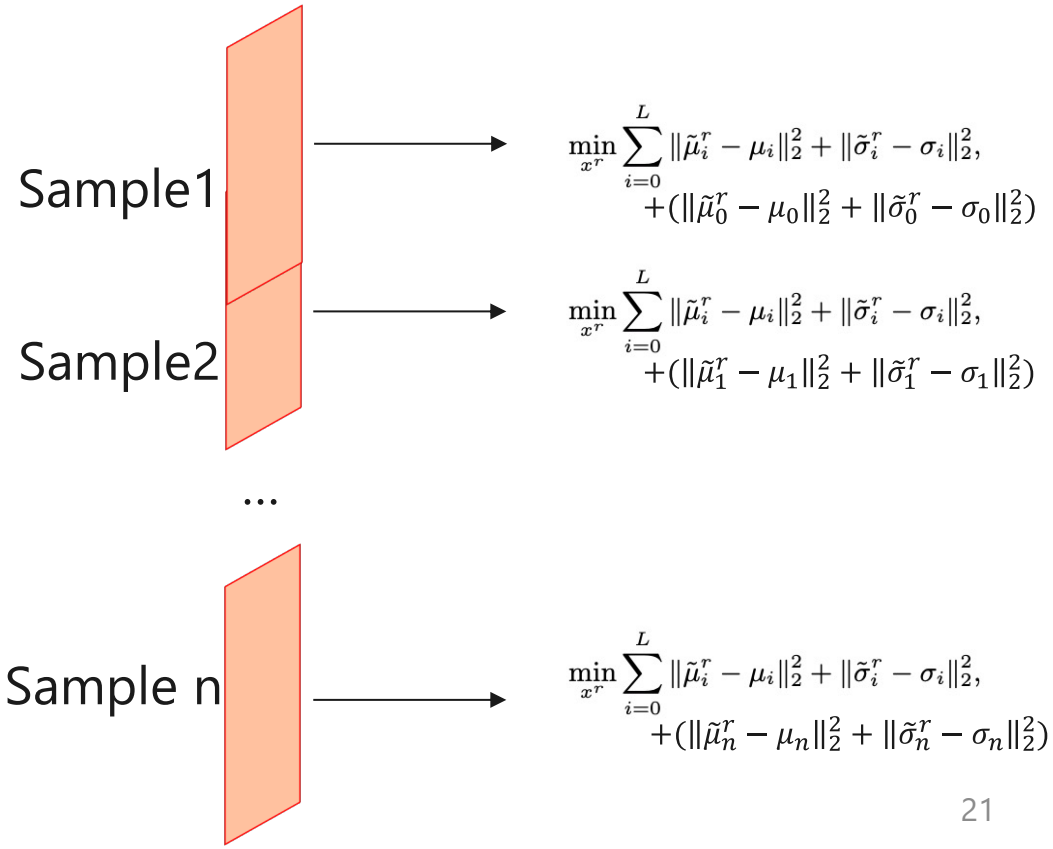


Diversifying Sample Generation

Layerwise Sample Enhancement (LSE)

- BNS: **All** samples pay the **same** attention to **all** BNS



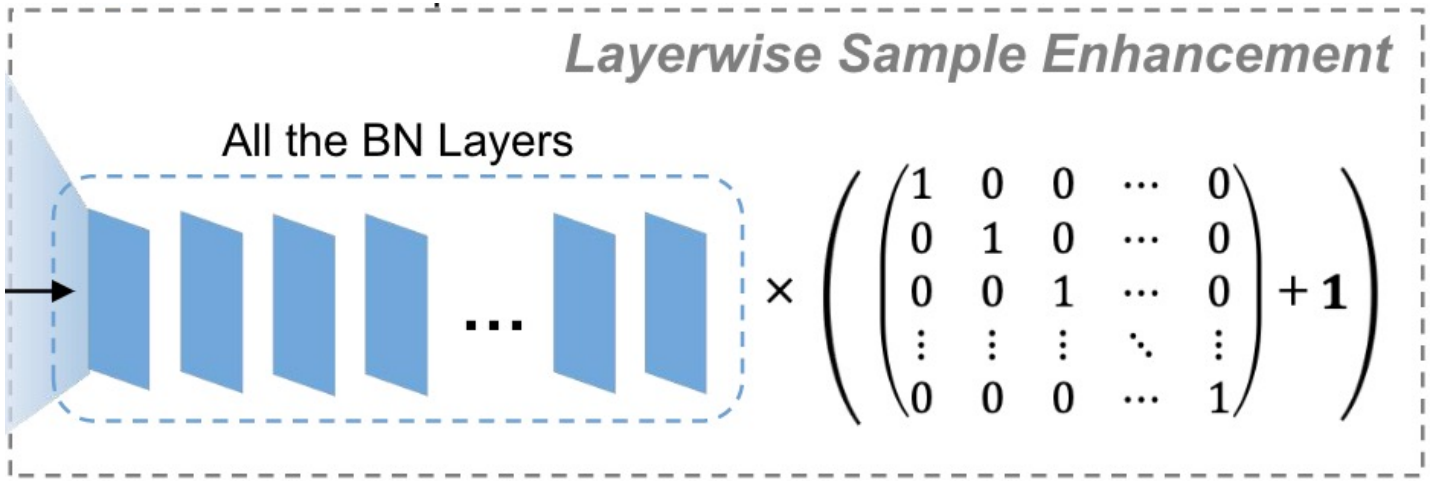


Diversifying Sample Generation

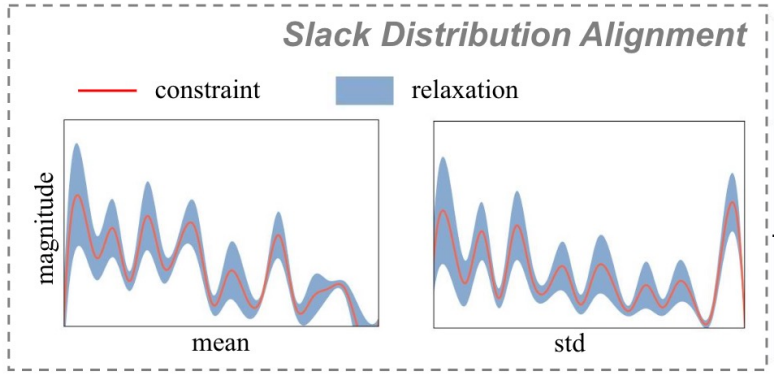
Layerwise Sample Enhancement (LSE)

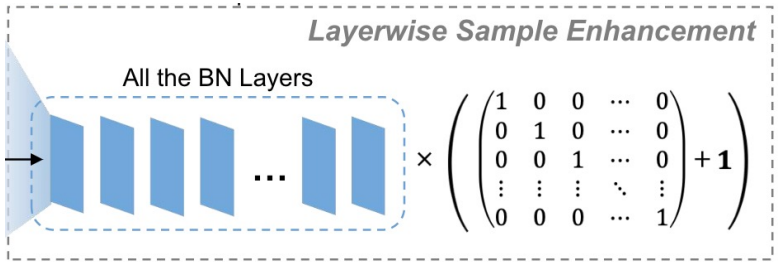
- Enhance the loss of particular layer for the corresponding sample using an enhancement matrix.

$$\mathcal{L} = \frac{1}{N} \cdot \mathbf{1}^T \left(\mathbf{X}_{\mathrm{LSE}} \mathbf{L} \right)$$



Diversifying Sample Generation





$$l_{\mathrm{SDA}_i} = \|\max\left(|\tilde{\boldsymbol{\mu}}_i^s - \boldsymbol{\mu}_i| - \delta_i, 0\right)\|_2^2 + \|\max\left(|\tilde{\boldsymbol{\sigma}}_i^s - \boldsymbol{\sigma}_i| - \gamma_i, 0\right)\|_2^2$$

$$\mathcal{L} = \frac{1}{N} \cdot \mathbf{1}^T \left(\mathbf{X}_{\mathrm{LSE}}\mathbf{L}\right)$$

$$\mathcal{L}_{\mathrm{DSG}} = \frac{1}{N} \cdot \mathbf{1}^T \left(\mathbf{X}_{\mathrm{LSE}}\mathbf{L}_{\mathrm{SDA}}\right)$$

Update generated data

Diversifying Sample Generation: experiment

Improvement

Visually

ZeroQ

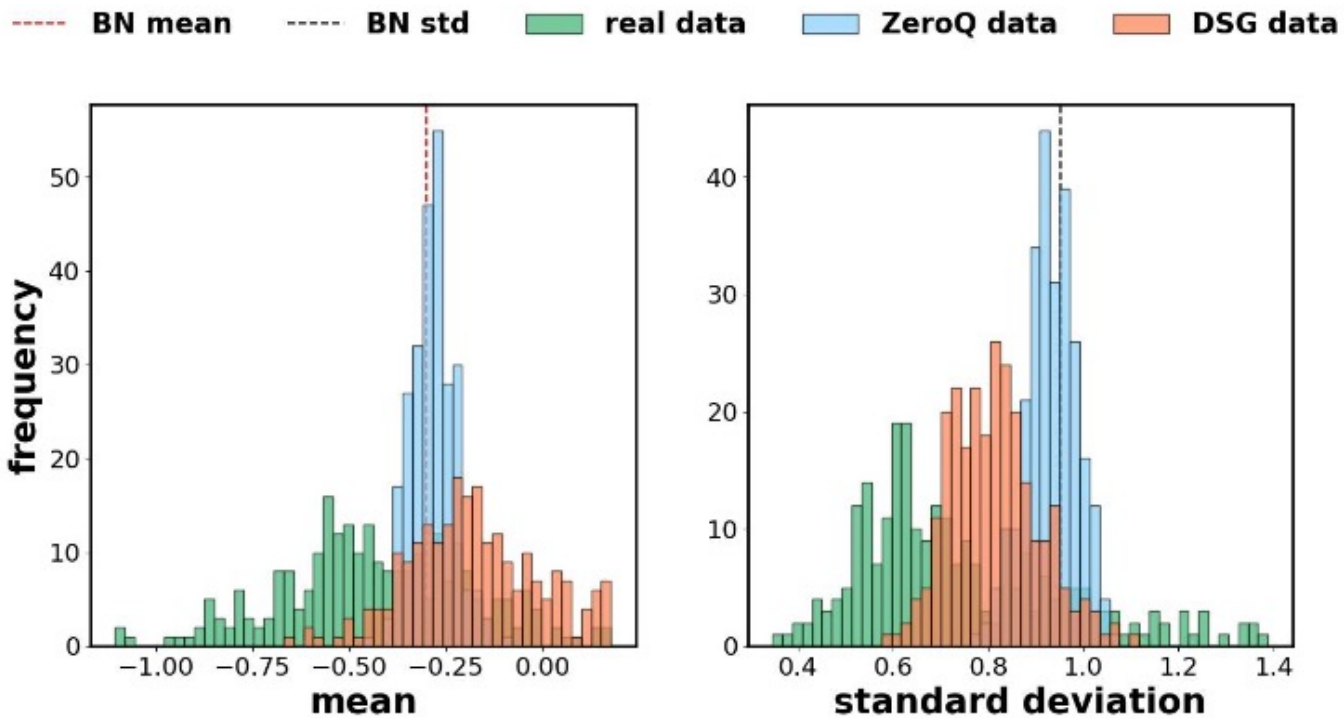
Homogeneous

DSG

Diverse

Diversifying Sample Generation: experiment

Improvement



Significant improvement in distribution
(**alleviation of homogenization**)

Diversifying Sample Generation: experiment

Improvement

(a) SqueezeNext

Method	No D	No FT	W-bit	A-bit	Top-1
Baseline	–	–	32	32	69.38
Real Data	✗	✓	6	6	62.88
ZeroQ	✓	✓	6	6	39.83
DSG (Ours)	✓	✓	6	6	**60.50**
Real Data	✗	✓	8	8	69.23
ZeroQ	✓	✓	8	8	68.01
DSG (Ours)	✓	✓	8	8	**69.27**

20% improvement (6W/6A)

(b) InceptionV3

Method	No D	No FT	W-bit	A-bit	Top-1
Baseline	–	–	32	32	78.80
Real Data	✗	✓	4	4	23.23
ZeroQ	✓	✓	4	4	12.00
DSG (Ours)	✓	✓	4	4	**34.89**
Real Data	✗	✓	6	6	77.96
ZeroQ	✓	✓	6	6	75.14
DSG (Ours)	✓	✓	6	6	**76.52**
Real Data	✗	✓	8	8	78.78
ZeroQ	✓	✓	8	8	78.70
DSG (Ours)	✓	✓	8	8	**78.81**

23% improvement (4W/4A)

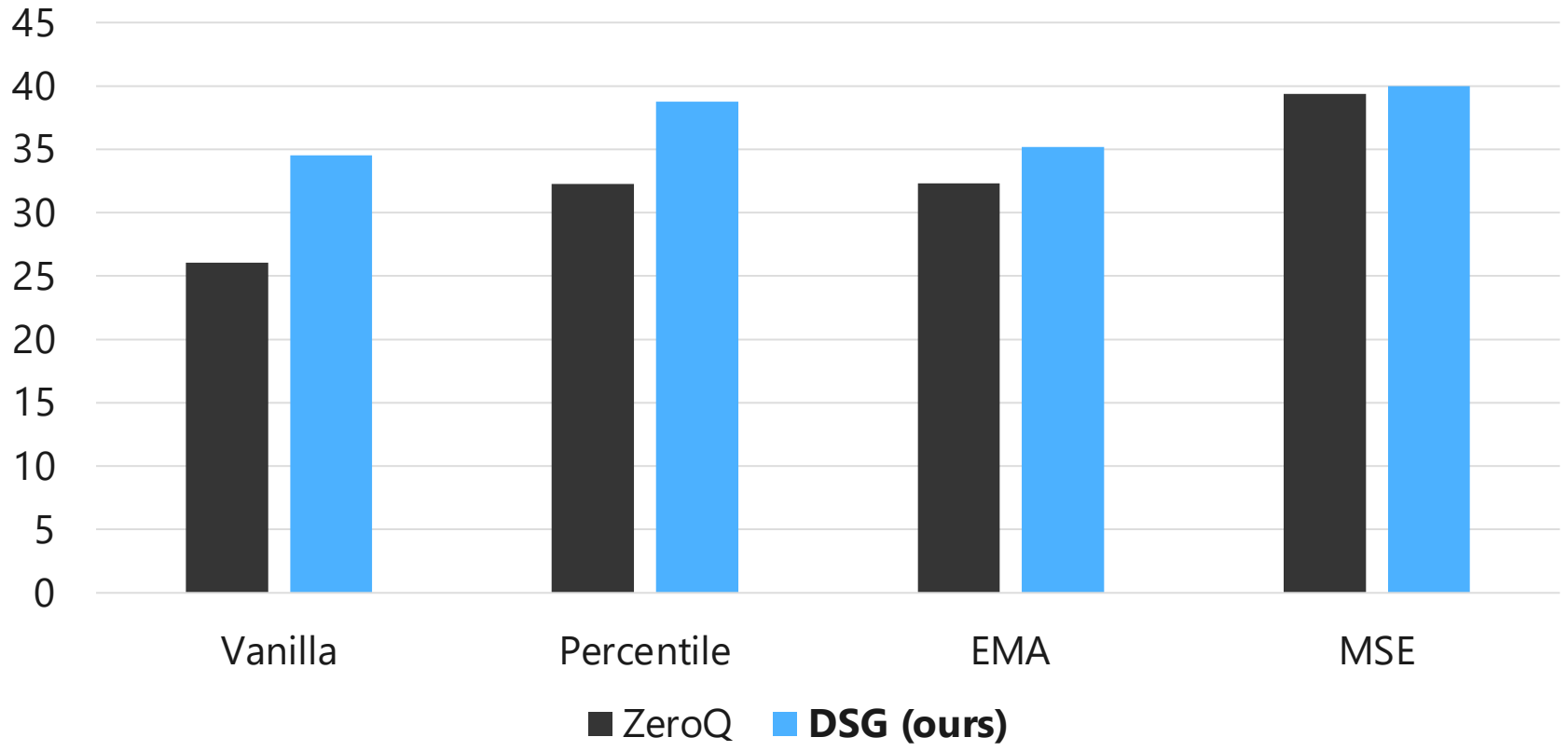
(a) ResNet-18

Method	No D	No FT	W-bit	A-bit	Top-1
Baseline	–	–	32	32	71.47
Real Data	✗	✓	4	4	31.86
ZeroQ	✓	✓	4	4	26.04
DSG (Ours)	✓	✓	4	4	**34.53**
Real Data	✗	✓	6	6	70.62
Integer-Only	✗	✗	6	6	67.30
DFQ	✓	✓	6	6	66.30
ZeroQ	✓	✓	6	6	69.74
DSG (Ours)	✓	✓	6	6	**70.46**
Real Data	✗	✓	8	8	71.44
RVQuant	✗	✗	8	8	70.01
DFQ	✓	✓	8	8	69.70
DFC	✓	✗	8	8	69.57
ZeroQ	✓	✗	8	8	71.43
DSG (Ours)	✓	✓	8	8	**71.49**

8% improvement (4W/4A)

Diversifying Sample Generation: experiment

- Outperform SOTA data-free quantization method over **various calibration methods** (quantizer)



Diversifying Sample Generation: experiment

Method	Label	Image Prior	W-bit	A-bit	Top-1
Real Data	No	No	3	32	64.16
ZeroQ	No	No	3	32	49.86
DSG (ours)	**No**	**No**	**3**	**32**	**56.09**
DSG (ours)	**Yes**	**No**	**3**	**32**	**58.27**
DSG (ours)	**Yes**	**Yes**	**3**	**32**	**61.32**

Diversifying Sample Generation: conclusion

1. We first revisit the data generation process in data-free quantization and **demonstrate that homogenization** exists at both **distribution and sample level**;

2. We present a **novel sample generation method** for accurate data-free quantization, dubbed as **DSG** scheme, to mitigate the homogenization issue;

3. Extensive experiments demonstrate the **leading accuracy and versatility** of the DSG scheme, especially **on ultra-low bit-width**.

Microsoft

Thank you!