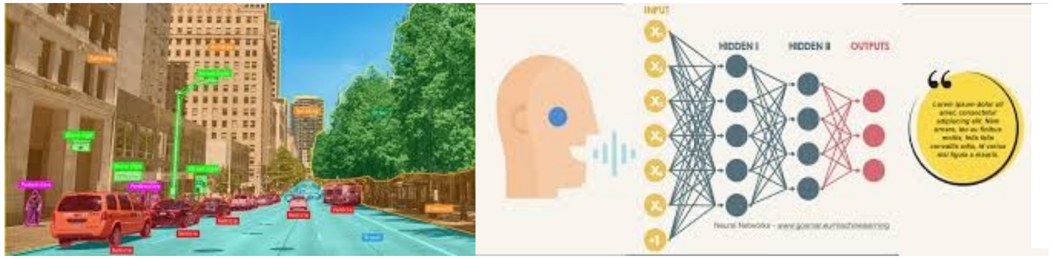
IOWA: An I/O-Aware Adaptive Sampling Framework for Deep Learning

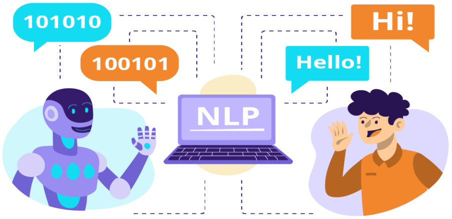
Shuang Hu, Weijian Chen, Yanlong Yin, Shuibing He

浙江大学
Zhejiang University

Deep Neural Network (DNN)

- Deep neural networks (DNNs) have achieved great success.





- The training time and cost are increasing rapidly.

Estimated training cost of select AI models, 2016–23

Source: Epoch, 2023 | Chart: 2024 AI Index report

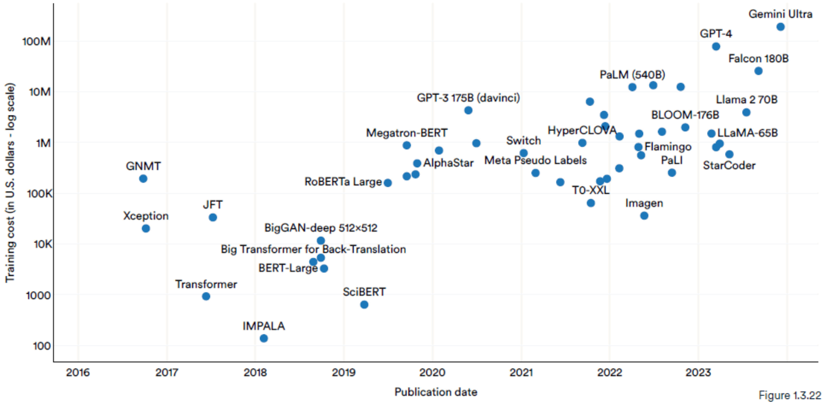
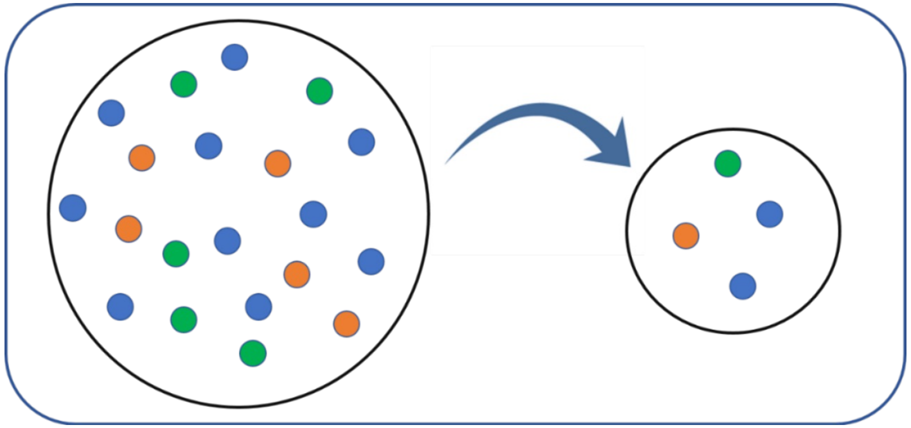


Figure 1.3.22

Importance Sampling

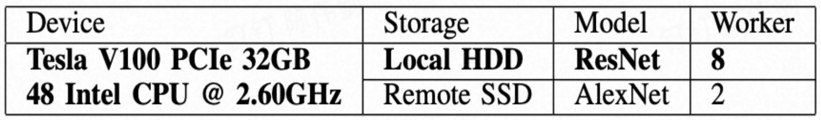
- **Traditional Training:** each data instance is treated equally.
- **Importance Sampling-Based Training:** some data are less informative, storing and training these instances has negligible benefit to improve the model accuracy.



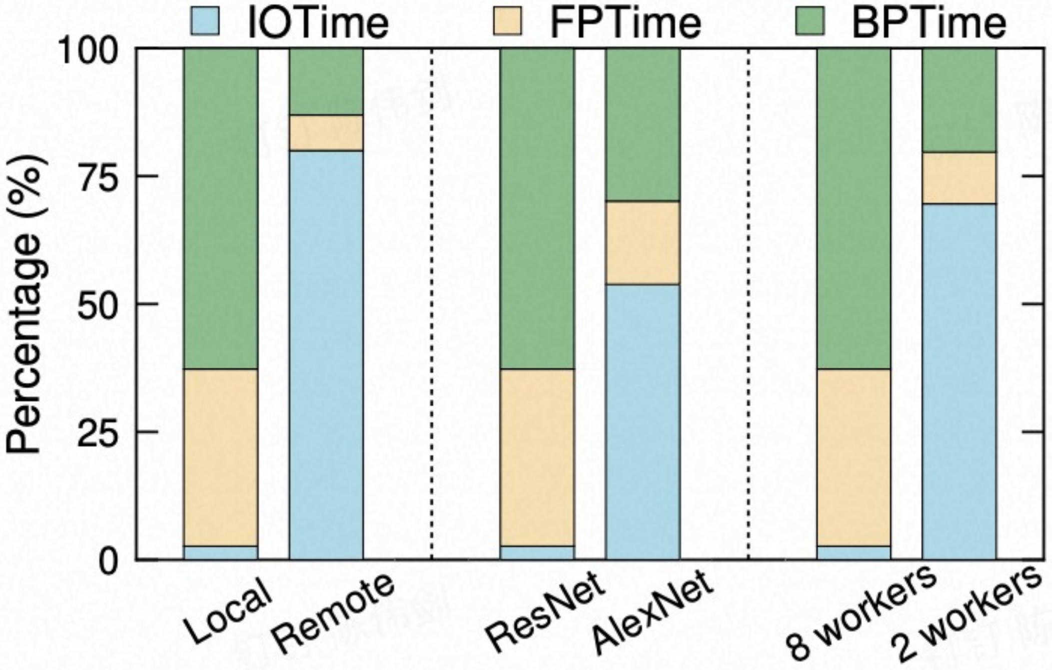
All training samples Informative samples

Motivation

- ➢Existing methods focus on reducing computational time, but computation is not always the bottleneck in training.



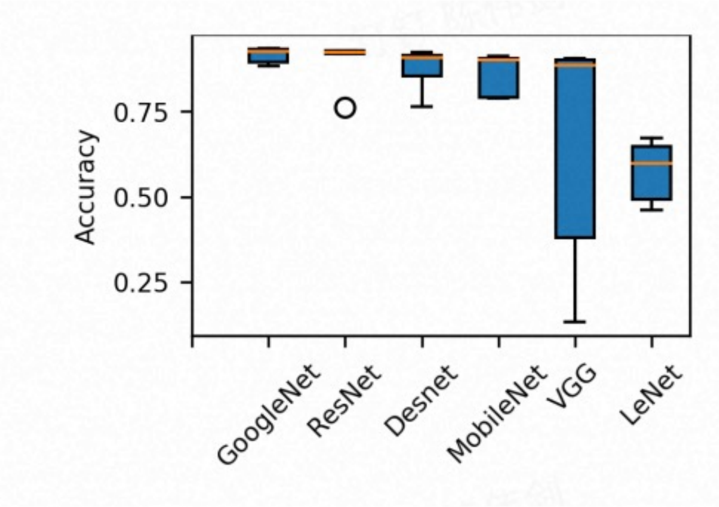
Device	Storage	Model	Worker
Tesla V100 PCIe 32GB	**Local HDD**	**ResNet**	**8**
48 Intel CPU @ 2.60GHz	Remote SSD	AlexNet	2



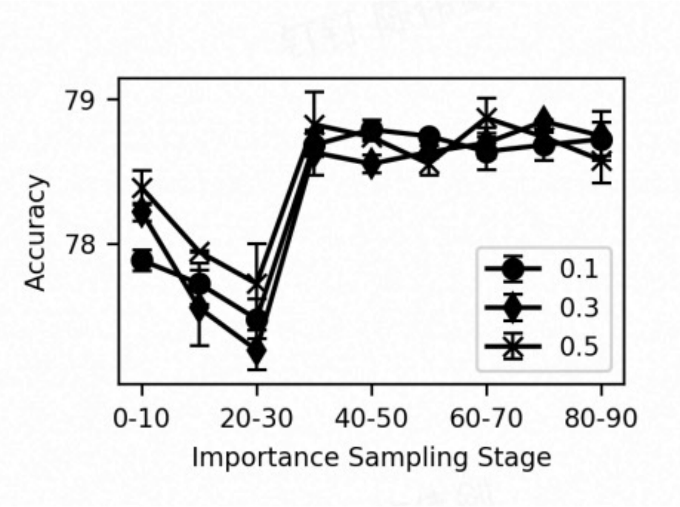
The training time breakdown on forward, backward, and I/O with different storage devices, models, and number of workers.

Motivation

- ➢Adaptive sampling rates should be applied when a model has higher discriminative ability.



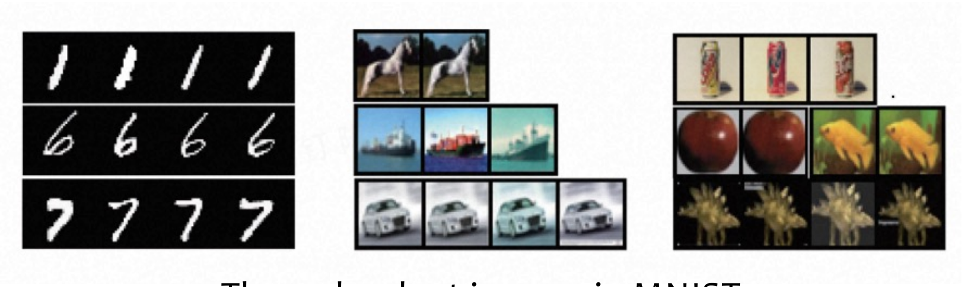
Six models on CIFAR10



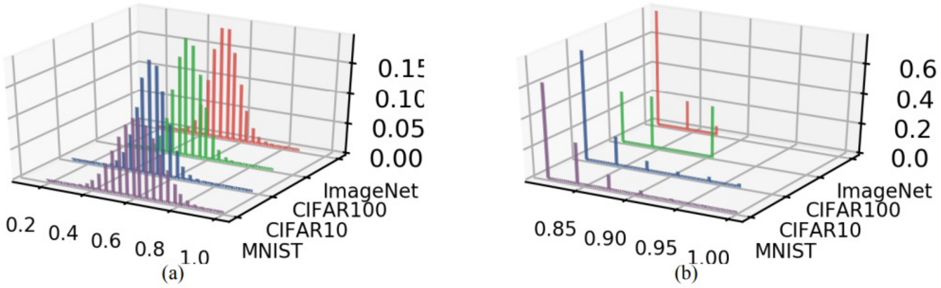
Various training stages with different sampling rates (ResNet18 on Food101)

Motivation

- ➢Redundant data instances can lead to unnecessary I/O access costs



The redundant images in MNIST, CIFAR10, and CIFAR100.

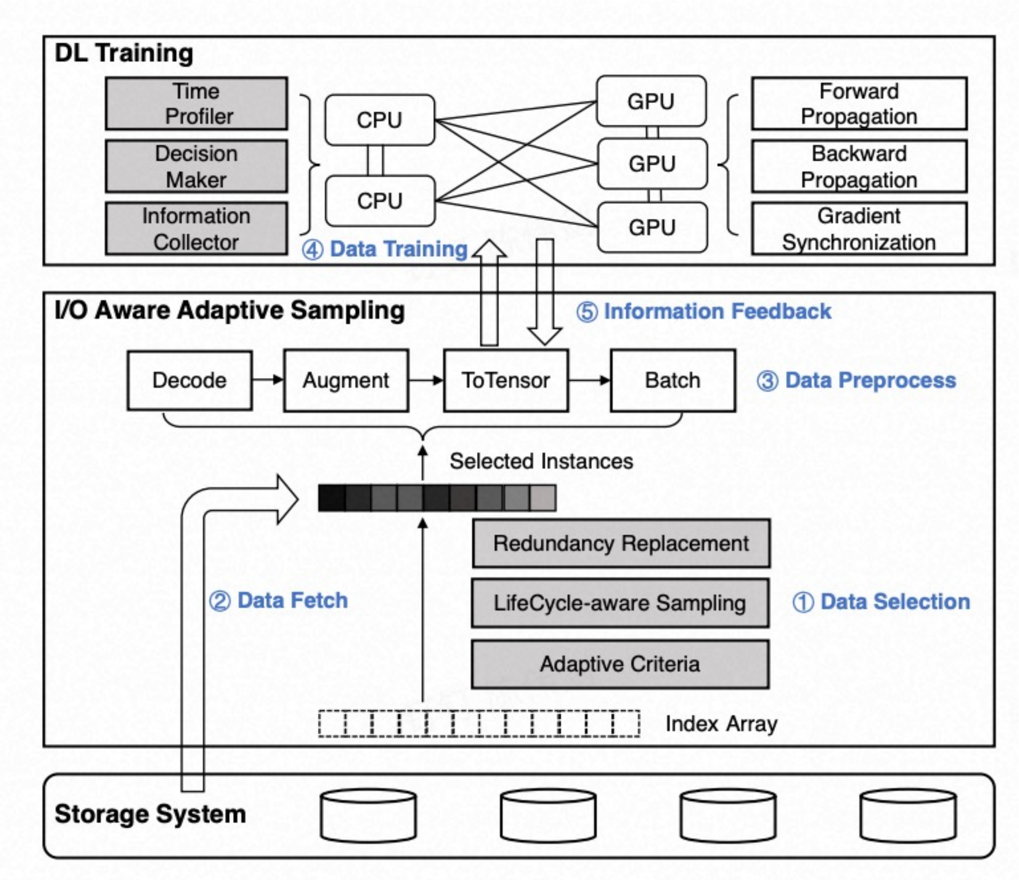


The distribution of similarity scores in MNIST, CIFAR10, CIFAR100, and ImageNet.

Outline

- ➢ Background & Motivation
- ➢ **Design of IOWA: An I/O-Aware Adaptive Sampling Framework**
- ➢ Evaluation
- ➢ Summary & Conclusion

Overview



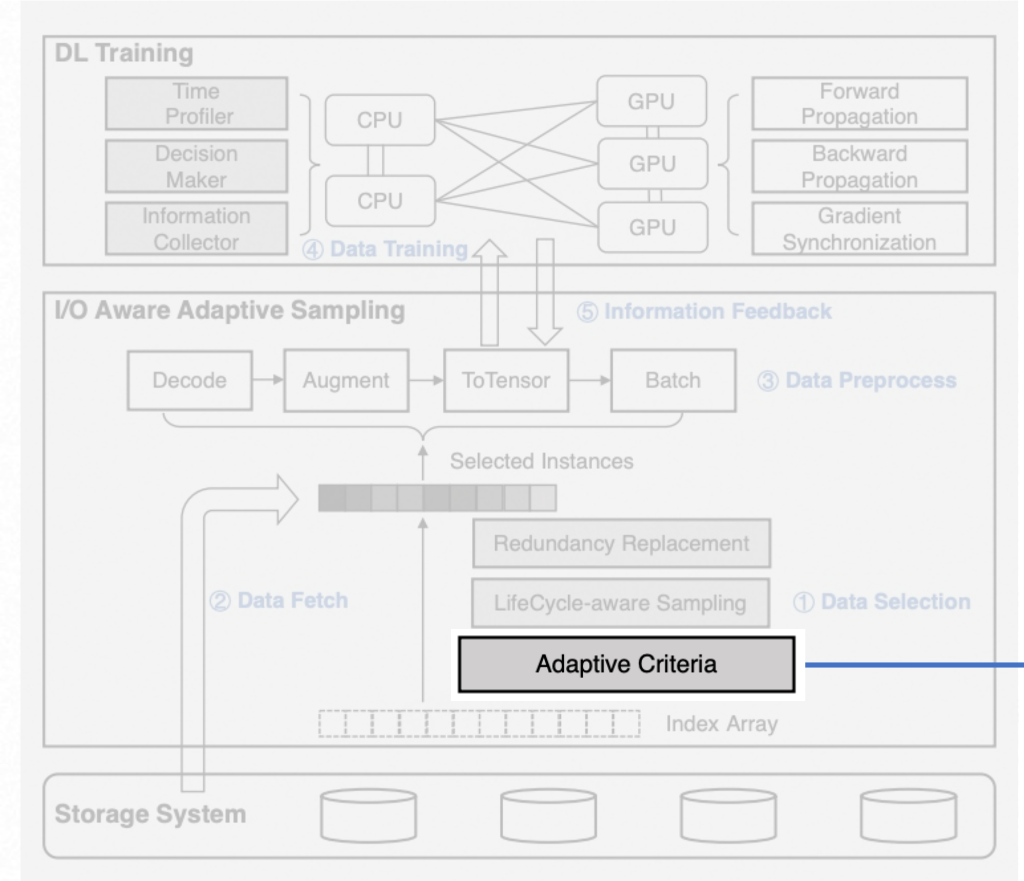
Three Important Parts:

1. Adaptive Criteria

2. LifeCycle-aware Sampling

3. Redundancy Replacement

Overview



Adaptive Criteria: an I/O-Aware Multi-Criteria Importance Metric to evaluate the importance of training data instances, taking I/O-intensive scenarios into consideration.

I/O-Aware Multi-Criteria Importance Metric

- **CPU-bound Task**

$$CrossEntropyScore = -P_\theta(y_{target}|x)log_2 P_\theta(y_{target}|x)$$

The smaller the value, the higher the probability of predicting the correct label, indicating less importance.

- **I/O-bound Task, add**

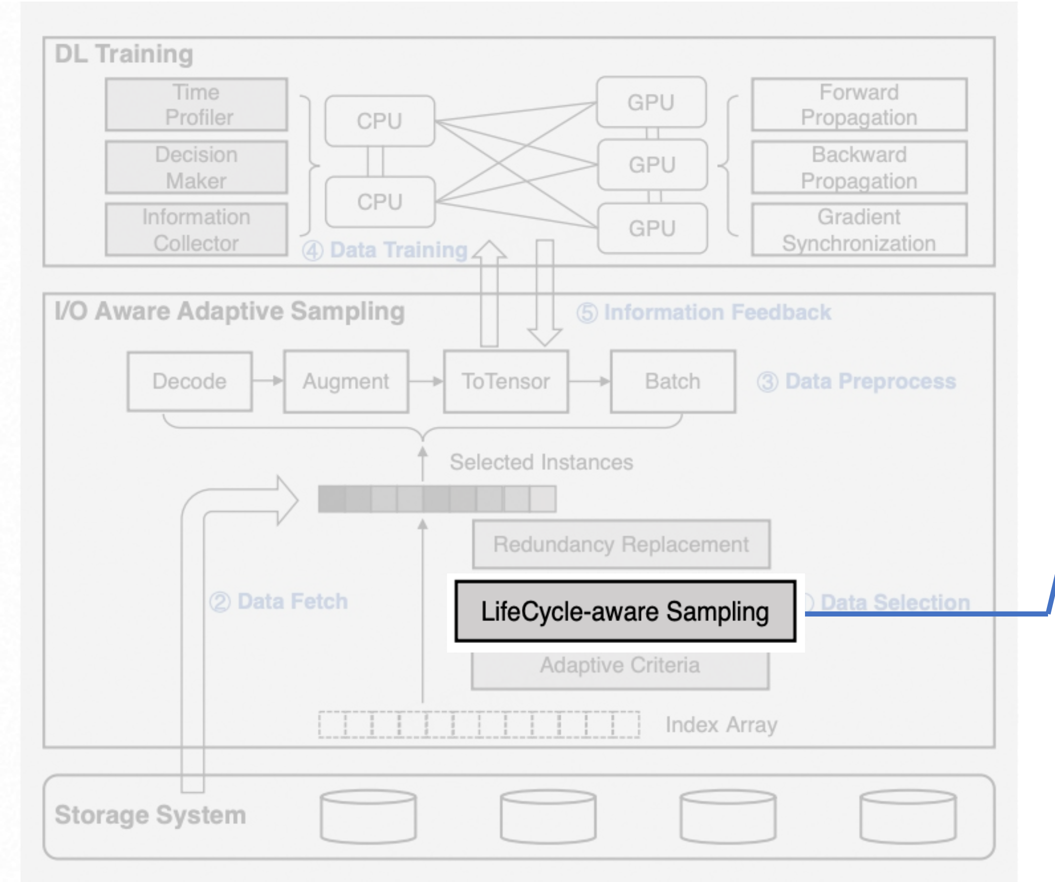
$$Margin\ Score = 1 - (P_\theta(y_{target}|x) - P_\theta(y_{max}|x))$$

The smaller the value, the lower the probability of classifying it as an incorrect label, indicating less importance.

$$I/O\ Score = 1/(\alpha \times Time_{load} + (1-\alpha) \times Size)$$

The smaller the value, the worse the I/O bottleneck.

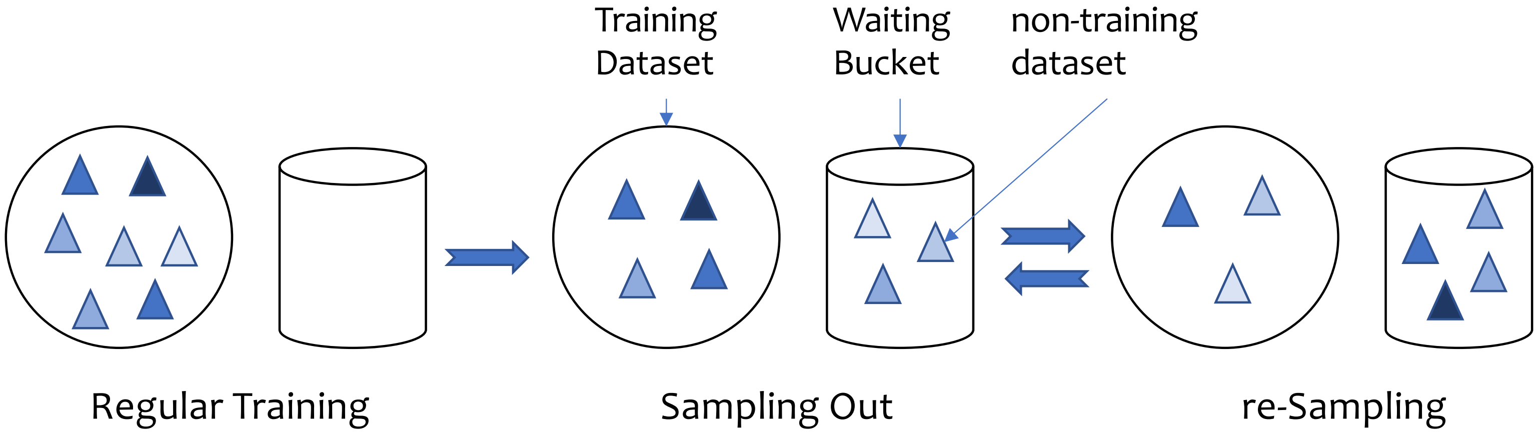
Overview



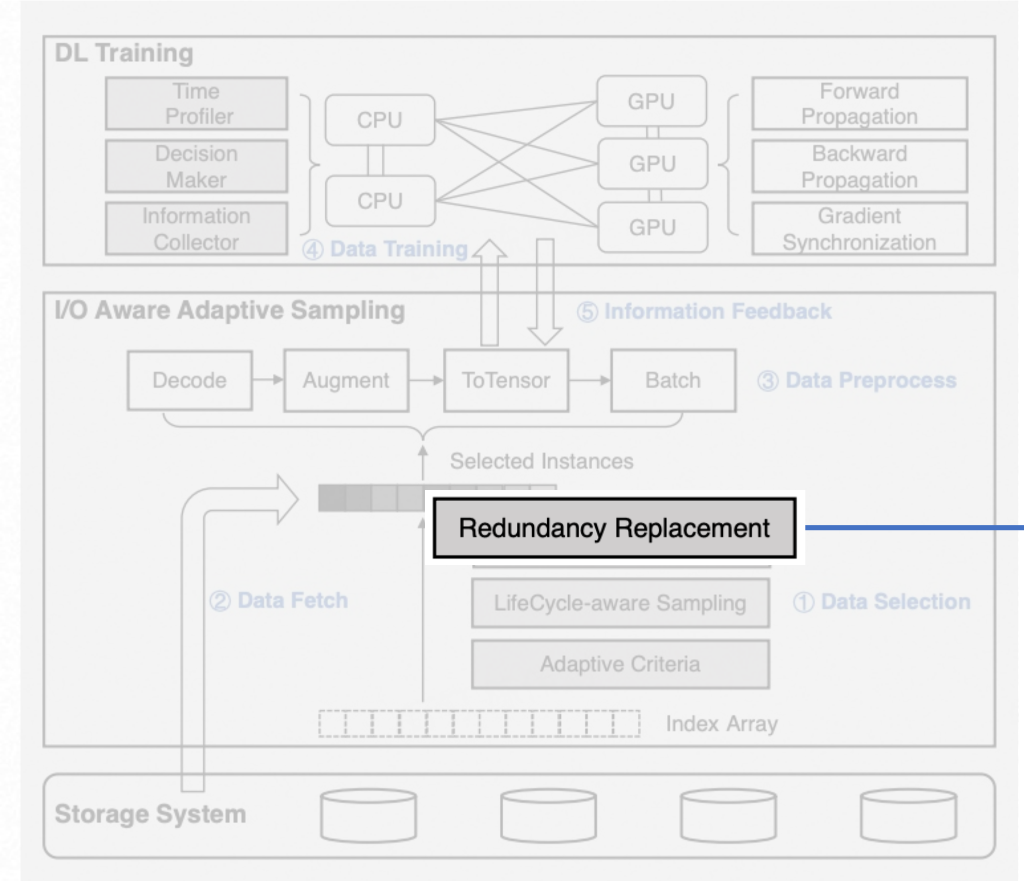
LifeCycle-aware Sampling: a Learning-Ability-Conscious Sampling Strategy, called Bucket Sampling, selects data instances according to model training.

Learning-Ability-Conscious Sampling Strategies

1. **Regular Training:** profile the bottleneck and get importance of all samples.
2. **Sampling Out:** dynamically adjust the training datasets by gradually moving data from the in-process to the waiting bucket.
3. **re-Sampling:** randomly re-sample data from the waiting bucket when training accuracy stagnates for *k* epochs.



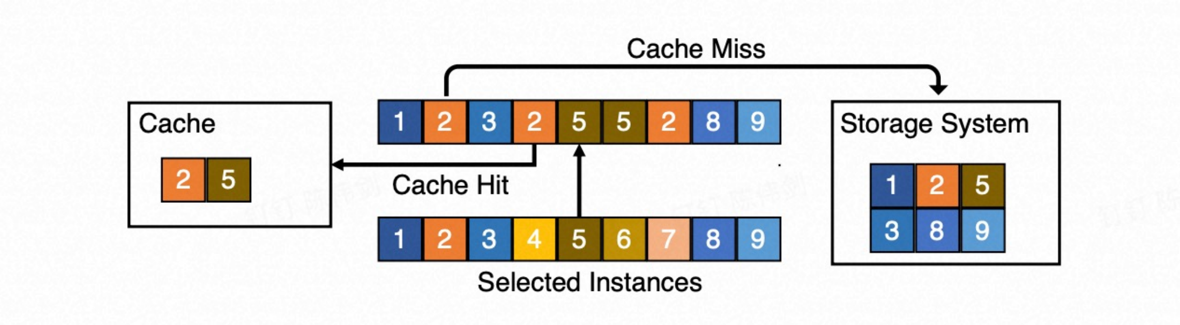
Overview



Redundancy Replacement: further reduce the training data size by using copies of data that are already loaded in memory or discarding similar copies.

Redundancy Removed Importance Sampling

- CPU bound task: discard samples to avoid redundant computation
- I/O bound task: use similar copies when cache miss to avoid I/O stall



Example: Instances {2, 4, 7} and {5, 6} are two groups with higher similar scores. In case we need to read data instances #4 and #7, we can replace them with instance #2 which is cached in memory rather than loading them from remote storage system.

Experimental Setup

- Platform
  - **single-machine tests:** a server equipped with 8-core Intel Xeon CPUs, 128GB of RAM, and an NVIDIA Tesla V100 GPU with 32GB of GPU memory
  - **data-parallel distributed training:** a platform with 64-core AMD CPUs, 256GB of RAM, and 4 A100-SXM GPUs, each with 40GB of GPU memory
  - **storage:** 2TB SSD

- Workloads and Datasets :
  - For the image classification task, we use four datasets: MNIST, CIFAR10, CIFAR100, and ImageNet, representing small, medium, medium, and large datasets respectively.

Experimental Setup

- ➢Baseline
  - **Origin:** uniform sampling
  - **Online:** samples data instances according to a loss-rank based probability.
  - **Active Bias:** re-weights data instances according to prediction variance
  - **Biggest Loser:** back-propagate instances with high loss values
  - **Select Via Proxy:** target models were trained on the coreset selected by a pre-trains lightweight model
- ➢Ours
  - **BS-norecall**: BS without re-Sampling stage
  - **BS**
    - **BS-random:** BS with re-Sampling stage
    - **BS-random-RRemove**: remove similar instance
    - **BS-random-RRepeat:** re-train the similar instance

Speedup with a Single GPU

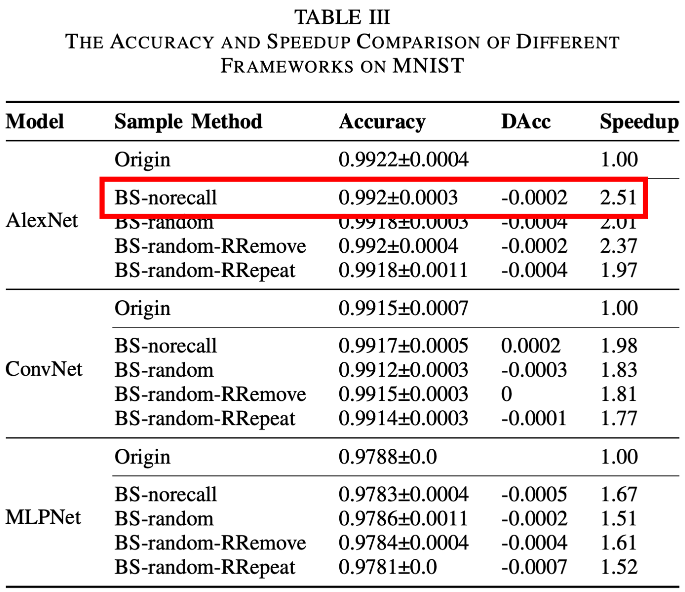


TABLE III
The Accuracy and Speedup Comparison of Different Frameworks on MNIST

Model	Sample Method	Accuracy	DAcc	Speedup
AlexNet	Origin	0.9922±0.0004		1.00
	BS-norecall	0.992±0.0003	-0.0002	2.51
	BS-random	0.9918±0.0003	-0.0004	2.01
	BS-random-RRemove	0.992±0.0004	-0.0002	2.37
	BS-random-RRepeat	0.9918±0.0011	-0.0004	1.97
ConvNet	Origin	0.9915±0.0007		1.00
	BS-norecall	0.9917±0.0005	0.0002	1.98
	BS-random	0.9912±0.0003	-0.0003	1.83
	BS-random-RRemove	0.9915±0.0003	0	1.81
	BS-random-RRepeat	0.9914±0.0003	-0.0001	1.77
MLPNet	Origin	0.9788±0.0		1.00
	BS-norecall	0.9783±0.0004	-0.0005	1.67
	BS-random	0.9786±0.0011	-0.0002	1.51
	BS-random-RRemove	0.9784±0.0004	-0.0004	1.61
	BS-random-RRepeat	0.9781±0.0	-0.0007	1.52

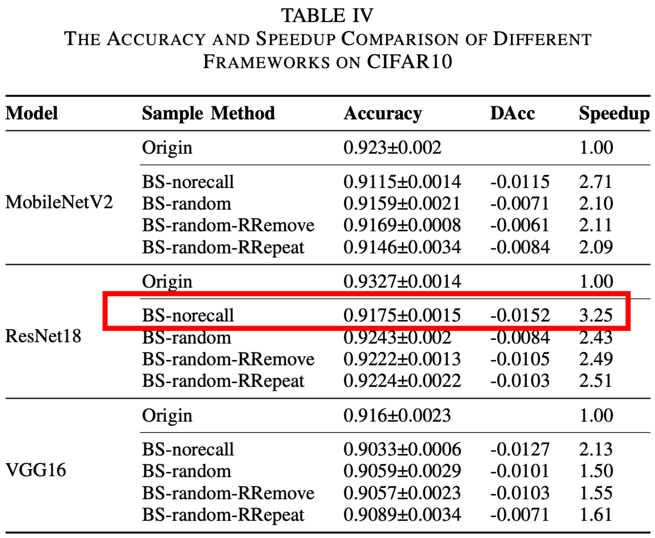


TABLE IV
The Accuracy and Speedup Comparison of Different Frameworks on CIFAR10

Model	Sample Method	Accuracy	DAcc	Speedup
MobileNetV2	Origin	0.923±0.002		1.00
	BS-norecall	0.9115±0.0014	-0.0115	2.71
	BS-random	0.9159±0.0021	-0.0071	2.10
	BS-random-RRemove	0.9169±0.0008	-0.0061	2.11
	BS-random-RRepeat	0.9146±0.0034	-0.0084	2.09
ResNet18	Origin	0.9327±0.0014		1.00
	BS-norecall	0.9175±0.0015	-0.0152	3.25
	BS-random	0.9243±0.002	-0.0084	2.43
	BS-random-RRemove	0.9222±0.0013	-0.0105	2.49
	BS-random-RRepeat	0.9224±0.0022	-0.0103	2.51
VGG16	Origin	0.916±0.0023		1.00
	BS-norecall	0.9033±0.0006	-0.0127	2.13
	BS-random	0.9059±0.0029	-0.0101	1.50
	BS-random-RRemove	0.9057±0.0023	-0.0103	1.55
	BS-random-RRepeat	0.9089±0.0034	-0.0071	1.61

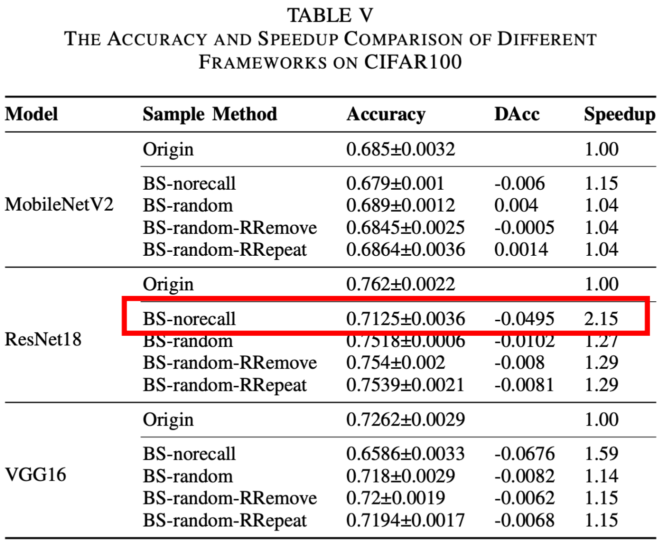
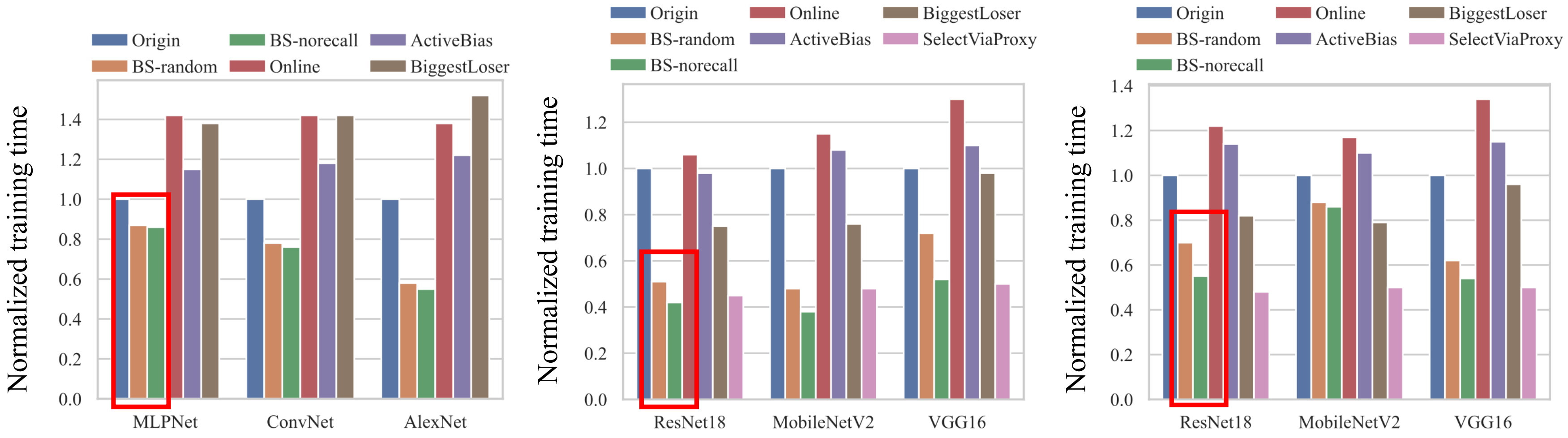


TABLE V
The Accuracy and Speedup Comparison of Different Frameworks on CIFAR100

Model	Sample Method	Accuracy	DAcc	Speedup
MobileNetV2	Origin	0.685±0.0032		1.00
	BS-norecall	0.679±0.001	-0.006	1.15
	BS-random	0.689±0.0012	0.004	1.04
	BS-random-RRemove	0.6845±0.0025	-0.0005	1.04
	BS-random-RRepeat	0.6864±0.0036	0.0014	1.04
ResNet18	Origin	0.762±0.0022		1.00
	BS-norecall	0.7125±0.0036	-0.0495	2.15
	BS-random	0.7518±0.0006	-0.0102	1.27
	BS-random-RRemove	0.754±0.002	-0.008	1.29
	BS-random-RRepeat	0.7539±0.0021	-0.0081	1.29
VGG16	Origin	0.7262±0.0029		1.00
	BS-norecall	0.6586±0.0033	-0.0676	1.59
	BS-random	0.718±0.0029	-0.0082	1.14
	BS-random-RRemove	0.72±0.0019	-0.0062	1.15
	BS-random-RRepeat	0.7194±0.0017	-0.0068	1.15

Compared to the origin, BS can achieve up to a 3× speedup.

Speedup with a Single GPU



MNIST

CIFAR10

CIFAR100

The speed up of different algorithms

Compared to other importance sampling systems, BS achieves the shortest training time.

Speedup with Distributed GPUs

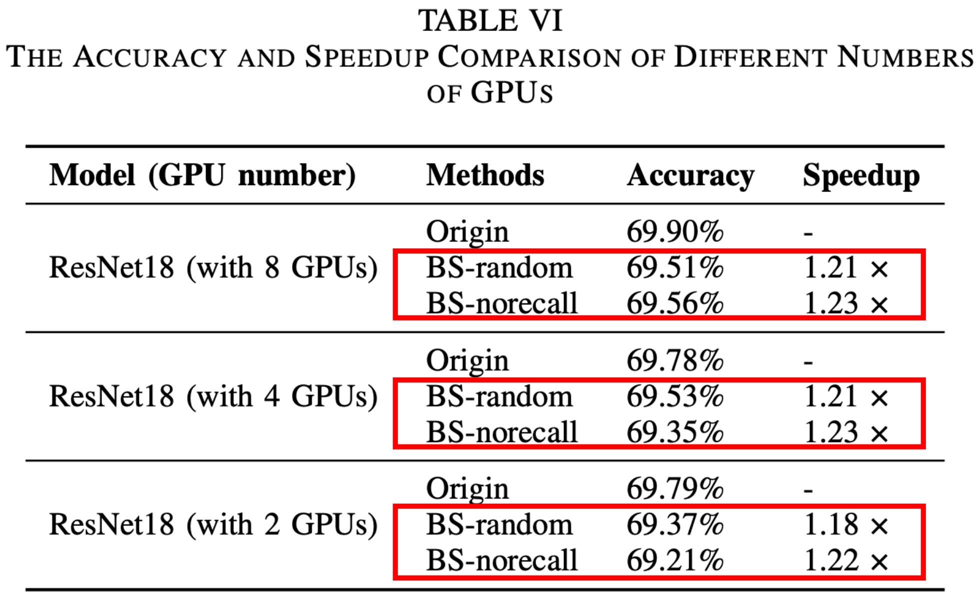
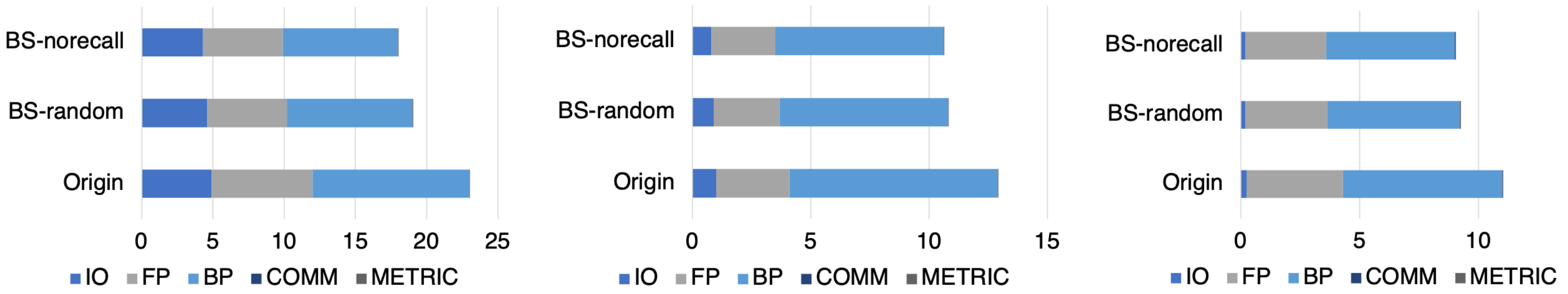


TABLE VI
THE ACCURACY AND SPEEDUP COMPARISON OF DIFFERENT NUMBERS OF GPUS

Model (GPU number)	Methods	Accuracy	Speedup
ResNet18 (with 8 GPUs)	Origin	69.90%	-
	BS-random	69.51%	1.21 ×
	BS-norecall	69.56%	1.23 ×
ResNet18 (with 4 GPUs)	Origin	69.78%	-
	BS-random	69.53%	1.21 ×
	BS-norecall	69.35%	1.23 ×
ResNet18 (with 2 GPUs)	Origin	69.79%	-
	BS-random	69.37%	1.18 ×
	BS-norecall	69.21%	1.22 ×

BS achieves a 1.2x speedup for distributed DNN training.

Overhead



(a) With 2 GPUs (b) With 4 GPUs (c) With 8 GPUs

The time breakdown (in minutes) of training ResNet on ImageNet.

The additional computation time (METRIC) introduced by the BS is negligible compared to the I/O and computation time it reduces.

Thanks!

IOWA: An I/O-Aware Adaptive Sampling Framework for Deep Learning

浙江大学
Zhejiang University

A short BIO of the speaker

Yanlong Yin received the PhD degree in Computer Science from Illinois Institute of Technology, in 2014. He is now an adjunct professor with Zhejiang University, China. Before that, he was an associate researcher with Zhejiang Lab and Beijing Open Source Chip Institute. His research interests include intelligent computing, parallel computing, and parallel storage systems.