KDD2025 August 3-7, 2025

SHANGHAI JIAO TONG UNIVERSITY

A Two-Stage Data Selection Framework for Data-Efficient Model Training on Edge Devices

Chen Gong, Rui Xing, Zhenzhe Zheng, Fan Wu

Introduction

Goal: Accelerate on-device model training by prioritizing limited hardware resources for important streaming data.

Challenge:

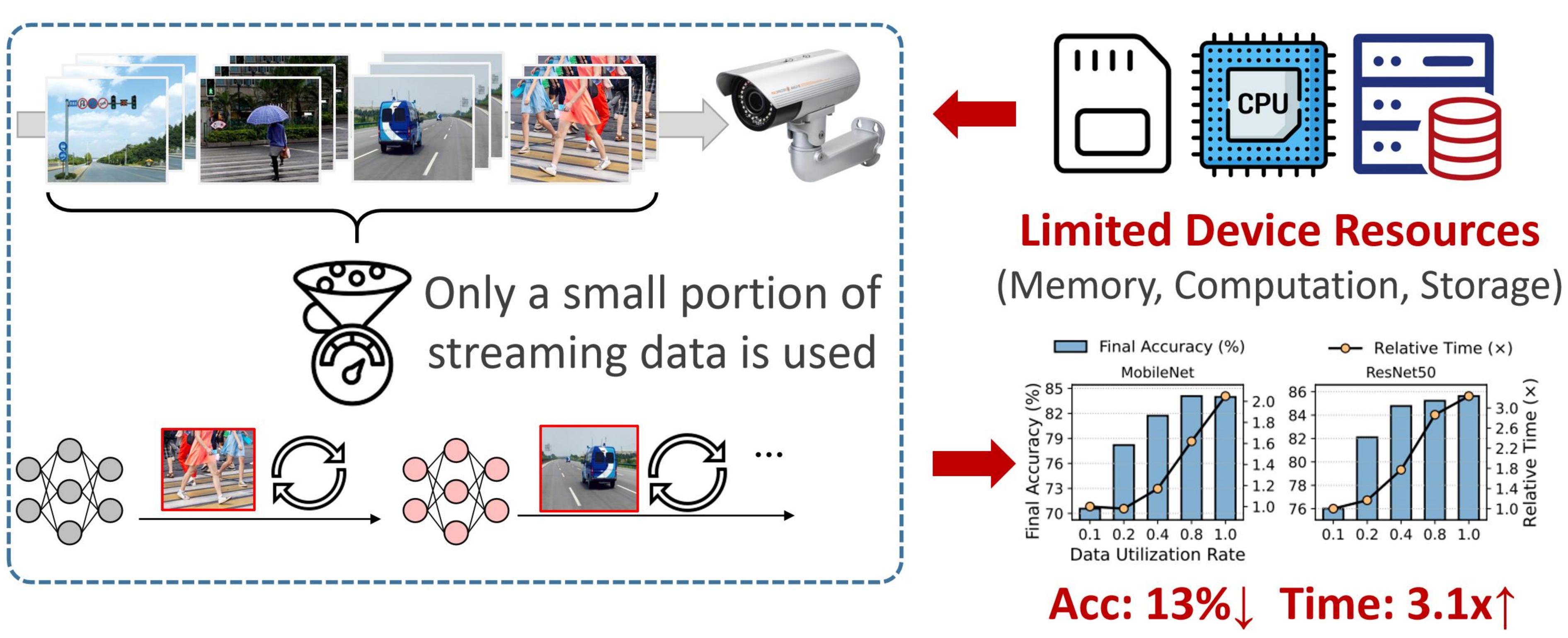
- **Effectiveness:** Provide both theoretical and empirical performance guarantees for data selection.
- **Efficiency:** Achieve low latency and resource contention.
- **Trade-off:** Higher effectiveness demands more accurate but costly evaluations on more data → lower efficiency.

Solutions:

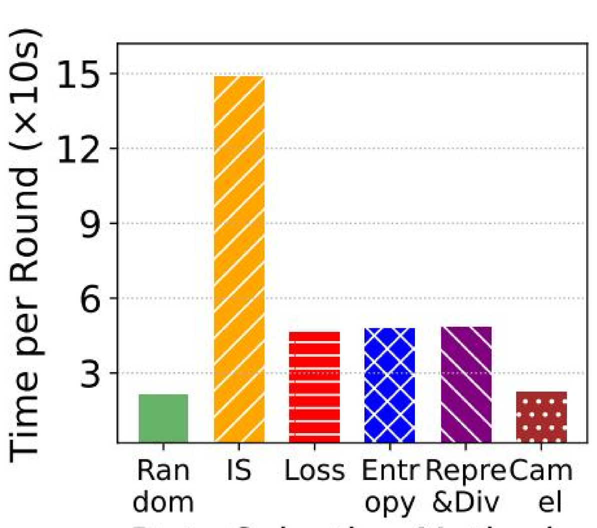
- **Effectiveness:** A theoretically optimal data selection algorithm to maximize training performance.
- **Time-Efficiency:** A coarse-grained filter to estimate each streaming data's importance in real time.
- **Resource-Efficiency:** A pipeline design to offload data selection to idle hardware resources.

Motivation

Data Bottleneck: On-device streaming data is significantly underutilized due to limited hardware resources.

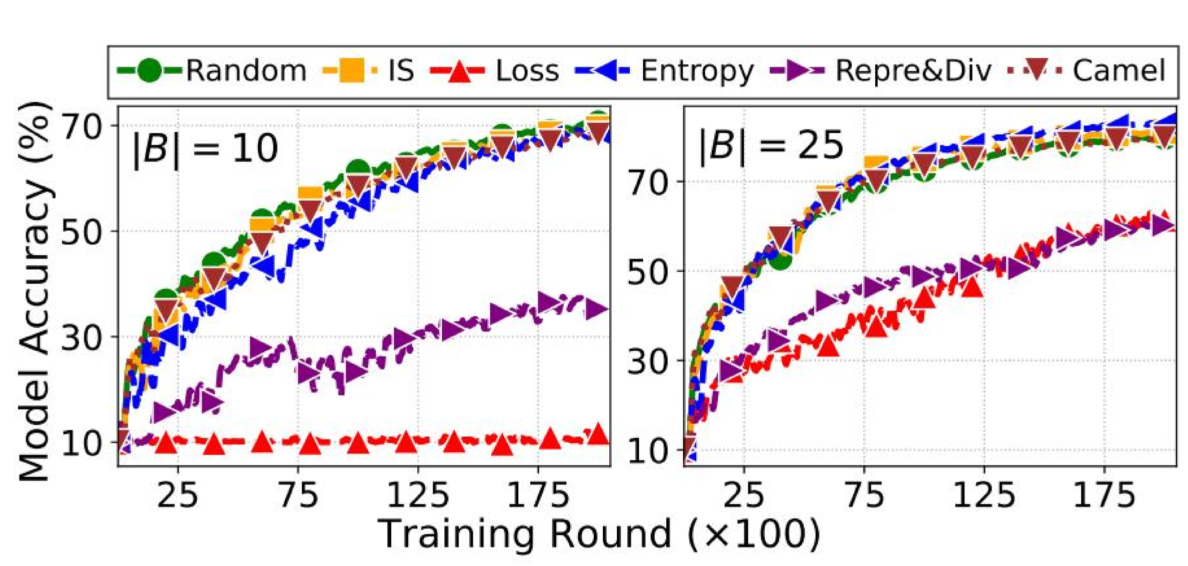


Limitations of Existing Works: Prior cloud-side data selection methods are not well-suited for on-device settings.



(a) Per-round training time.

High Evaluation Latency



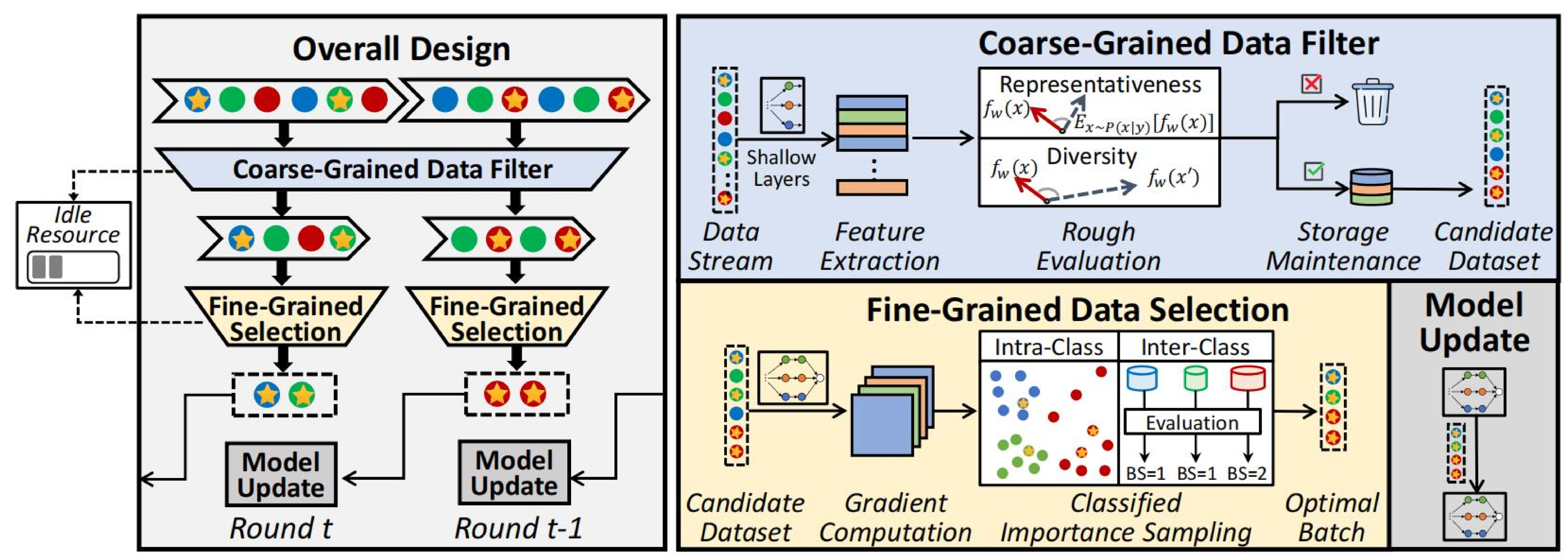
(b) Training processes with batch sizes 10 and 25.

Low Performance Gains w.r.t. random sampling

Design

Overview: During model taining, Titan uses idle resources to

- filter a small candidate dataset with high *representativeness* and *diversity* **(coarse-grained filter)**,
- optimally determine *how many* and *which* samples to select for each data class **(fine-grained selection)**,
- concurrently train model using previously selected data on the main hardware resource **(pipeline design)**.



Theoretical Optimality:

- **Theorem 1:** Model training performance is inversely correlated with the gradient variance of selected data batch.
- **Theorem 2:** To minimize gradient variance, the optimal selection size $|B_y|$ for each class y and probability $P_y(x)$ for each sample x are:

$$|B_y| \propto |S_y| \Big[\mathbb{V}_{(x,y)\sim P_{t,y}}[\nabla l(w_t, x, y)] - \mathbb{V}_{(x,y)\sim P_{t,y}}[||\nabla l(w_t, x, y)||_2] \Big]^{\frac{1}{2}}$$

$$P_y(x) \propto I_t(x, y) \triangleq \|\nabla l(w_t, x, y)\|_2$$

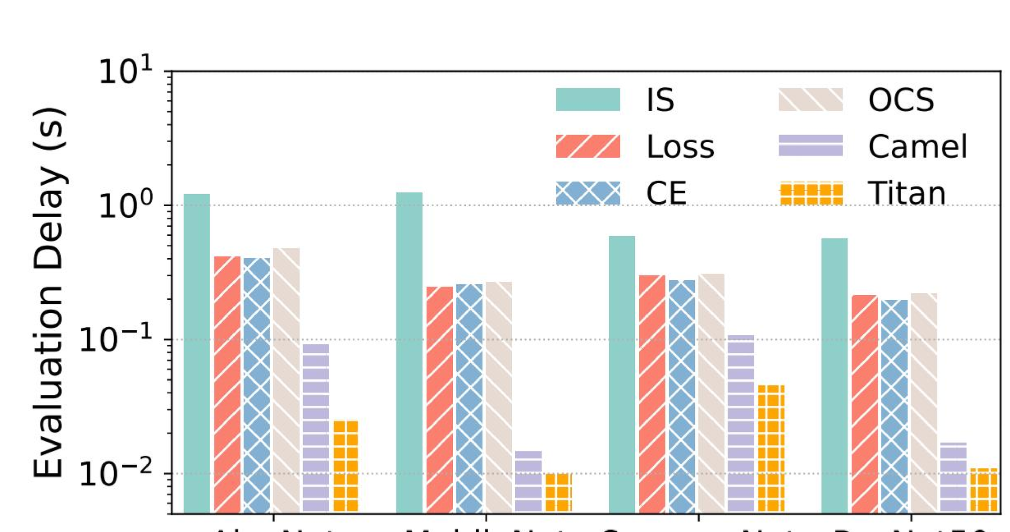
Evaluation

3 Tasks and Data Modalities:
Image Classification (IC), Audio Recognition (AR), Human Activity Recognition (HAR)

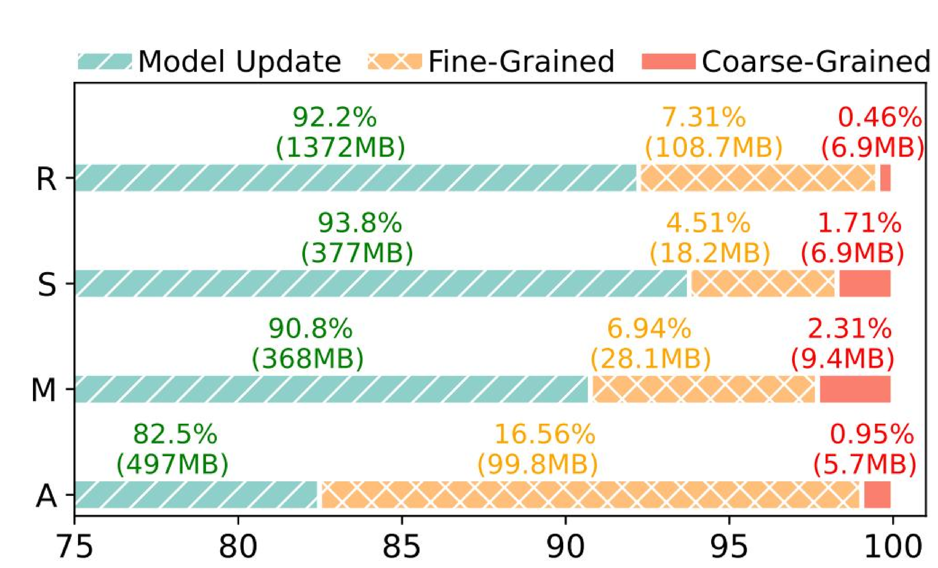
Task	Model	Normalized Time-to-Accuracy (×) RS	IS	LL	HL	CE	OCS	Camel	Titan
IC	AlexNet	1.00	3.25	3.98	3.98	3.59	4.06	2.07	0.70
	MobileNet	1.00	3.22	3.45	3.45	3.41	3.67	1.15	0.57
	SqueezeNet	1.00	3.96	3.97	3.97	3.04	4.06	2.07	0.69
	ResNet50	1.00	2.32	3.14	3.14	2.20	2.18	1.11	0.66
AR	ResNet34	1.00	2.04	3.14	3.14	2.96	3.19	0.81	0.77
HAR	MLP	1.00	3.56	6.30	6.47	5.28	14.4	12.5	0.71

Training Speedup:

- IC: 30%-43%
- AR: 23%
- HAR: 29%



Evaluating Steaming Data in ***ms*-level Latency**



Marginal Memory Footprint (≤120MB)