THE WEB CONFERENCE ACM

Research Track | Systems and Infrastructure for Web, Mobile, and WoT

Enabling Real-Time Inference in Online Continual Learning via Device-Cloud Collaboration

Haibo Liu, Chen Gong, Zhenzhe Zheng, Shengzhong Liu, Fan Wu

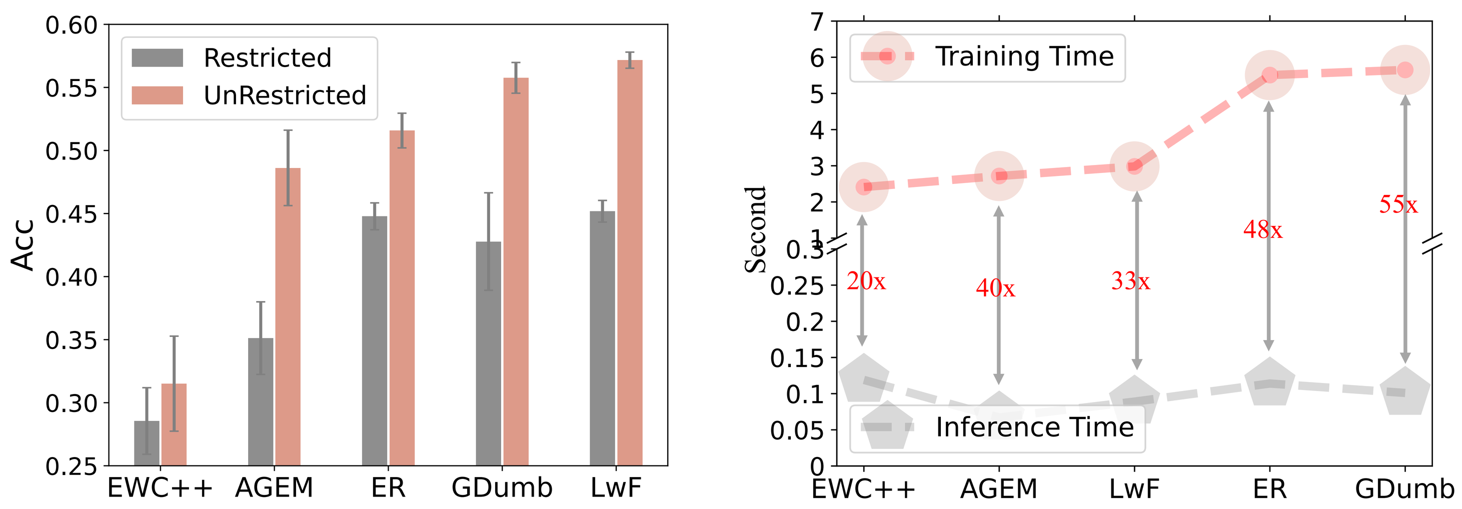
Shanghai Jiao Tong University

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Introduction

- Nowadays, massive data are continuously collected from ubiquitous end devices, and required immediate process to support real-time data analysis applications. Online continual learning (CL) is becoming a mainstream paradigm to learn incrementally from task streams without forgetting previously learned knowledge.
- However, the current online CL primarily focuses on learning performance, such as avoiding catastrophic forgetting, neglecting the critical demands of system performance, such as real-time inference. As a result, the performance of real-time inference in online CL degrades significantly due to frequent data distribution variations and time-consuming model adaptation.

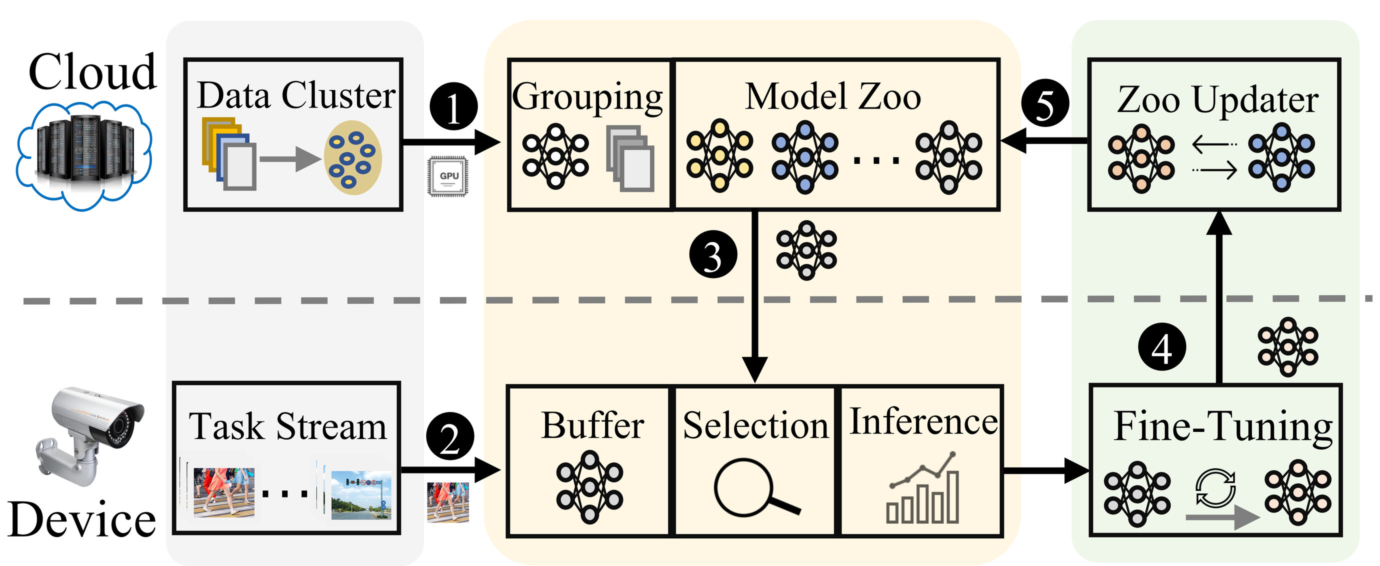
Background and Motivation



Model	Size	Training Time	Comm Latency
CNN	0.304MB	1.401s	0.0026s
LeNet5	2.181MB	2.021s	0.0043s
SqueezeNet	2.869MB	3.850s	0.0114s
ShuffleNet V2	8.772MB	5.020s	0.0384s
MobileNet V2	13.501MB	5.331s	0.0819s
ResNet18	42.838MB	6.577s	0.1295s

- The model performance of on-device CL degrades significantly in resource-limited scenarios. Moreover, the time consumption of model adaptation is up to 55 times of that of model inference, which would result in a long-time model adaptation for the encountered new task.
- Comparing to previous efforts, we prefer to enable real-time inference on resource-constrained end devices by retrieving suitable models from the cloud with model transmission.

Device-Cloud Collaboration



(a) Initialization (b) ELITE (c) Enhancement

Device-cloud collaboration may involve the following three stages:

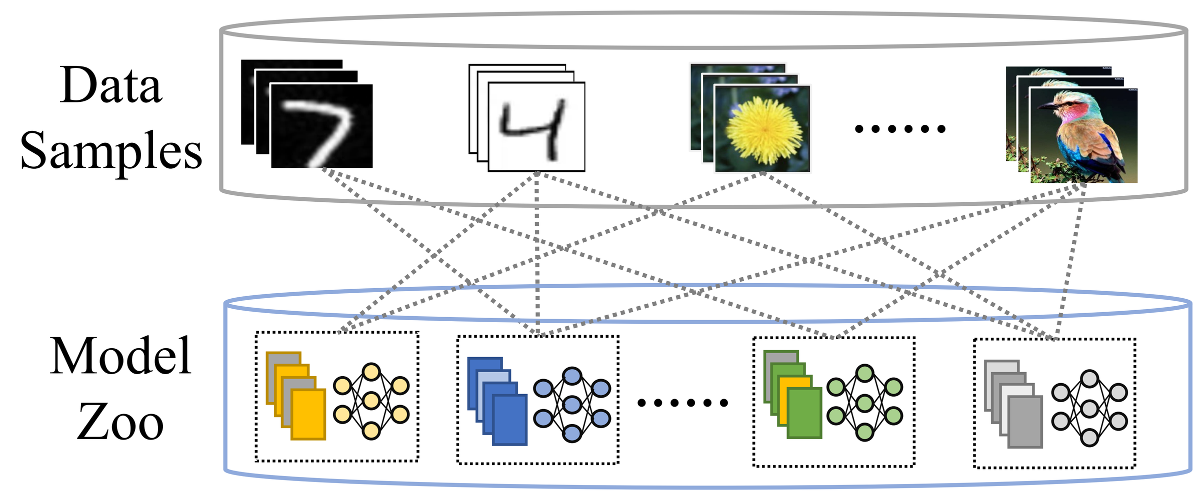
(a) Initialization: This stage serves as the preparation for model zoo generation and real-time inference. It involves the clustering of massive data for multi-task model training, coupled with the establishment of task streams;

(b) ELITE: This is our primary design to realize real-time inference with two components: the cloud-enabled model zoo and on-device real-time inference;

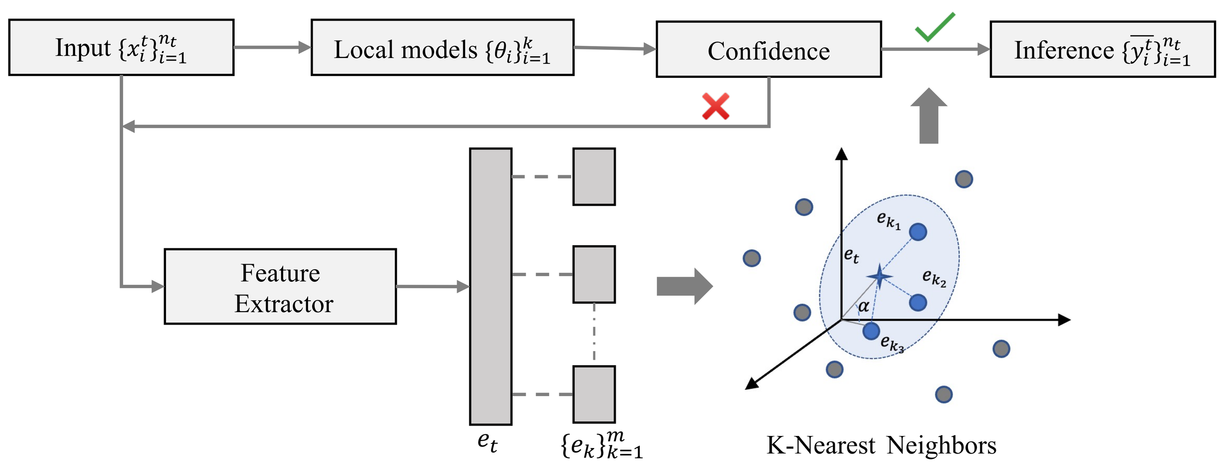
(c) Enhancement: To prevent the performance degradation of ELITE, we propose the latency-aware model fine-tuning on end devices, and dynamic model zoo updating in the cloud to adapt to new tasks.

Design of ELITE

- Without the information of task streams on end devices, we utilize multi-task training with abundant cloud-side data resources to pre-train various models for different tasks, and optimize task-model allocation to maximize the diversity of tasks that the pretrained multi-task model involve with;

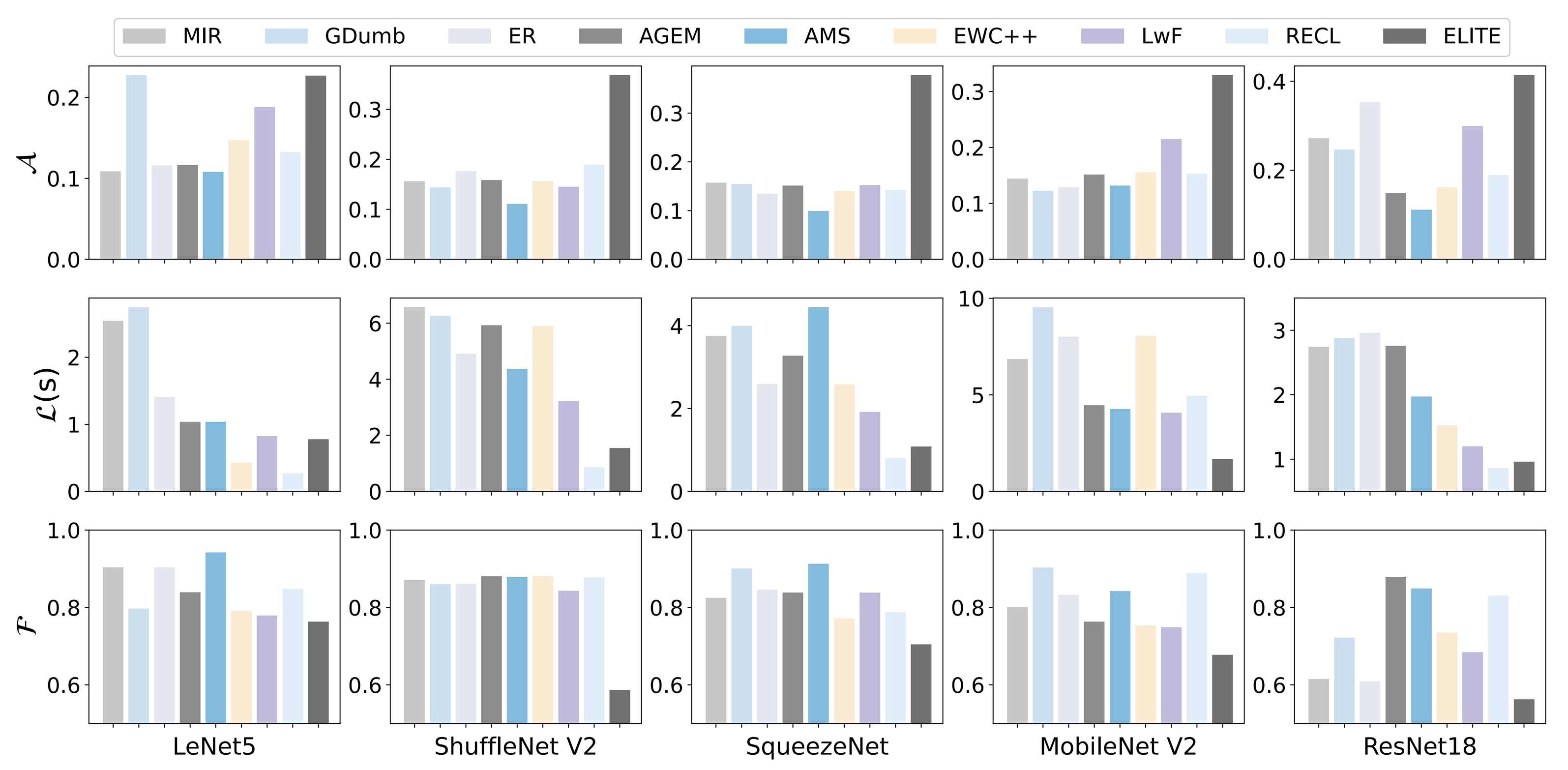


- To realize on-device model selection, we extract features of data samples as task embeddings, and select the most k suitable multi-task models by calculating domain similarity. After obtaining the k most suitable, ELITE selects the model with highest confidence to realize model inference.



Experimental Results

		EWC++	MIR	LwF	AMS	RECL	ELITE
CIFAR10	$\mathcal{A}$	0.176 ± 0.063	0.268 ± 0.093	0.275 ± 0.028	0.125 ± 0.021	0.172 ± 0.002	0.413 ± 0.039
	$\mathcal{L}(s)$	2.011 ± 0.488	3.031 ± 0.518	1.496 ± 0.292	1.971 ± 0.034	1.441 ± 0.097	1.127 ± 0.201
	$\mathcal{F}$	0.844 ± 0.063	0.736 ± 0.093	0.581± 0.030	0.881 ± 0.021	0.791 ± 0.001	0.581 ±0.039
CIFAR100	$\mathcal{A}$	0.176 ± 0.023	0.174 ± 0.029	0.153 ± 0.008	0.059 ± 0.011	0.164 ± 0.011	0.397 ± 0.014
	$\mathcal{L}(s)$	3.334 ± 0.566	6.884 ± 0.108	4.078 ± 0.097	4.241 ± 0.343	1.884 ± 0.199	1.341 ± 0.117
	$\mathcal{F}$	0.796 ±0.028	0.821 ±0.027	0.848 ±0.016	0.921 ±0.015	0.834 ±0.013	0.587 ±0.056
Tiny-ImageNet	$\mathcal{A}$	0.182 ± 0.053	0.176 ± 0.034	0.207 ± 0.027	0.117 ± 0.051	0.196 ± 0.038	0.275 ± 0.028
	$\mathcal{L}(s)$	1.718 ± 0.225	3.538 ± 0.632	1.589 ± 0.213	3.064 ± 0.567	1.945 ± 0.474	1.034 ± 0.059
	$\mathcal{F}$	0.890± 0.042	0.881± 0.027	0.811 ±0.012	0.958 ±0.032	0.827± 0.029	0.728 ±0.018
HDMB51	$\mathcal{A}$	0.157 ± 0.152	0.220 ± 0.136	0.346 ± 0.129	0.136 ± 0.143	0.543 ± 0.130	0.654 ± 0.043
	$\mathcal{L}(s)$	3.006 ± 0.777	4.117 ± 0.761	2.482 ± 0.476	6.269 ± 2.253	1.123 ± 0.263	1.032 ± 0.067
	$\mathcal{F}$	0.952± 0.016	0.771± 0.071	0.675 ±0.019	0.954 ±0.008	0.563 ±0.047	0.328± 0.013
UCF101	$\mathcal{A}$	0.129 ± 0.153	0.392 ± 0.138	0.252 ± 0.135	0.136 ± 0.143	0.412 ± 0.106	0.652 ± 0.075
	$\mathcal{L}(s)$	2.846 ± 0.431	4.509 ± 1.022	2.519 ± 0.232	6.269 ± 2.253	1.139 ± 0.258	1.033 ± 0.078
	$\mathcal{F}$	0.923 ±0.034	0.483± 0.081	0.831 ±0.020	0.954 ±0.008	0.565± 0.038	0.376± 0.066



Conclusion

- In this work, we focused on the real-time inference on resource-constraint end devices in online CL, and proposed a new device-cloud collaborative CL framework, namely ELITE, for time-varying task streams.
- To realize real-time model inference, ELITE formed model zoo in the cloud server, and proposed task-oriented on-device model selection on end devices.
- Extensive evaluations demonstrate that ELITE improves 16.3% inference performance and reduces up to 1.98x response latency compared to the-state-of-art solutions.