Device-Cloud Collaborative Learning for Recommendation

Jiangchao Yao† Feng Wang† Kunyang Jia†
Bo Han‡ Jingren Zhou† Hongxia Yang†

†DAMO Academy, Alibaba Group ‡Hongkong Baptist University

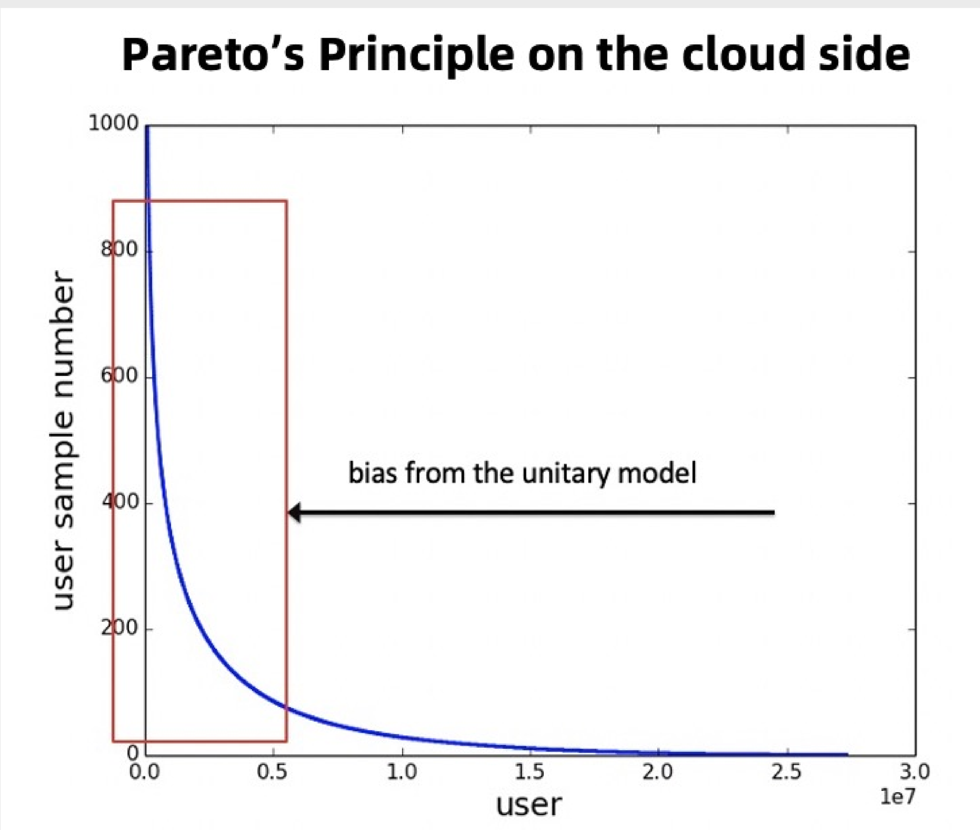
Overview

TL;DR: Nascent applications for mobile computing and the Internet of Things (IoTs) are driving computing toward dispersion. Mobile recommender systems with the on-device engines like TFLite and CoreML, hence attract more and more attention. Early works could be classified into two categories.

- **On-device Inference**: Ranking models are first trained offline, and then compressed or split for online serving.
- **On-device Learning for Encryption**: train model pieces on device and aggregate into centralized models in cloud.
- We explore the third direction about **collaborative AI** between device and cloud for recommendation, where a general framework **DCCL** is proposed.
  - ◇ Conduct the on-device model personalization to take the minority of patterns in Non-IID data into account.
  - ◇ Calibrate the centralized backbone model with thousands of on-device experts by means of distillation.

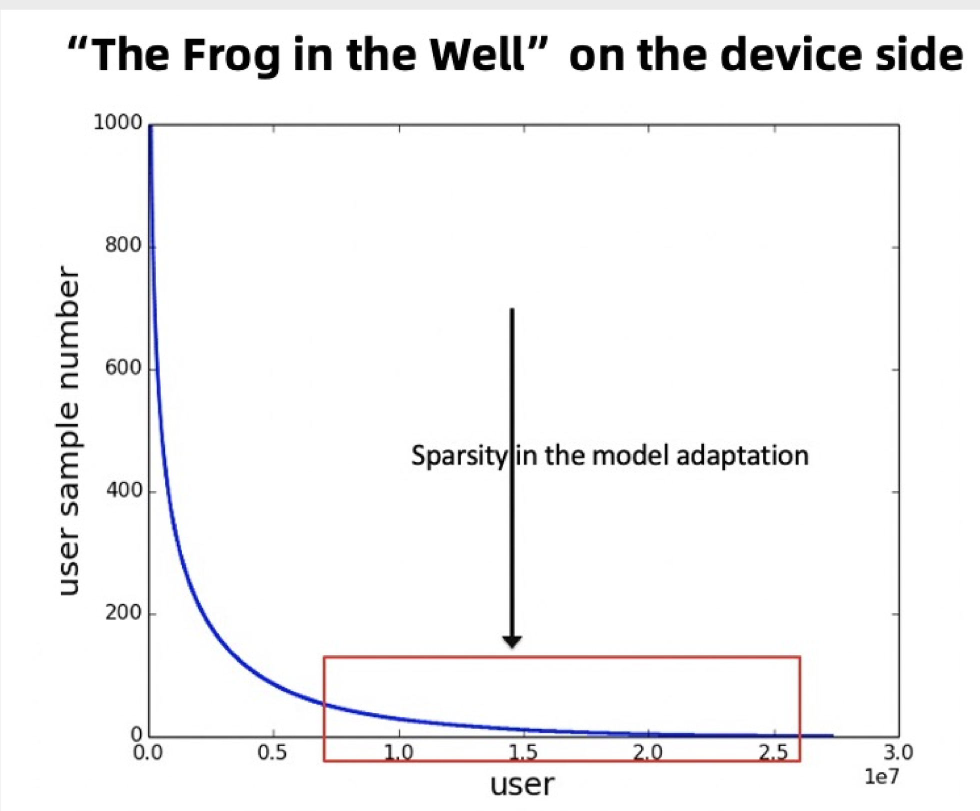
The extensive experiments on a range of datasets show the promise compared to the cloud recommenders.

Deficiency of Unitary Modeling



- Maximization of the global revenue via the unitary model might sacrifice the patterns of the minority.

Deficiency of Local Optimization

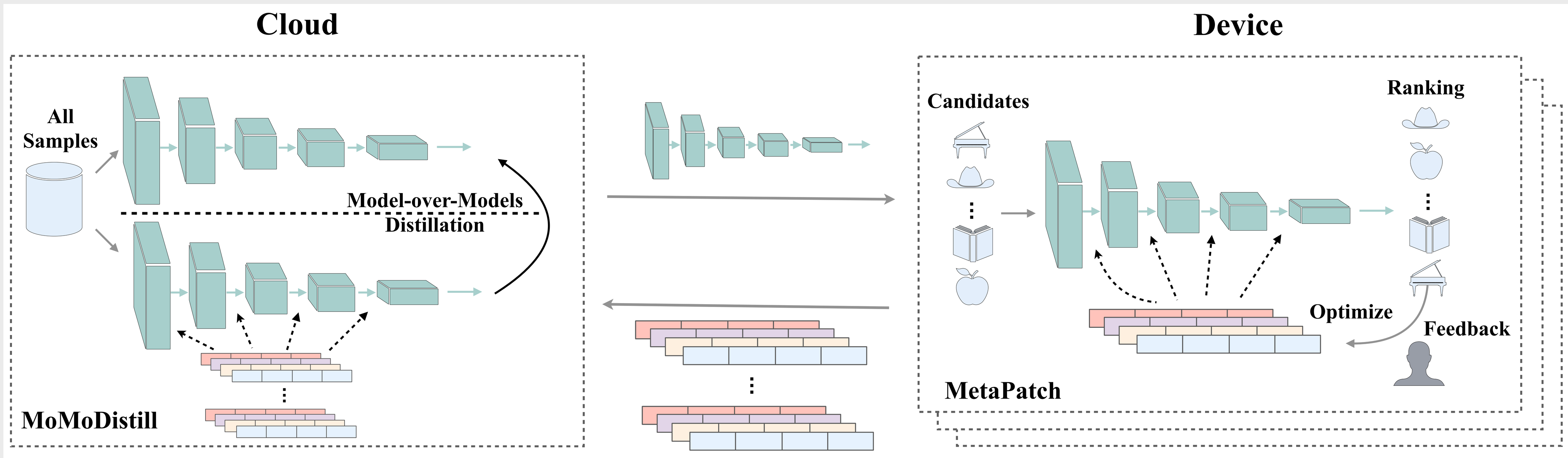


- Lengthy adaptation of the model to each device might fall into the sub-optimal due to the limited user samples.

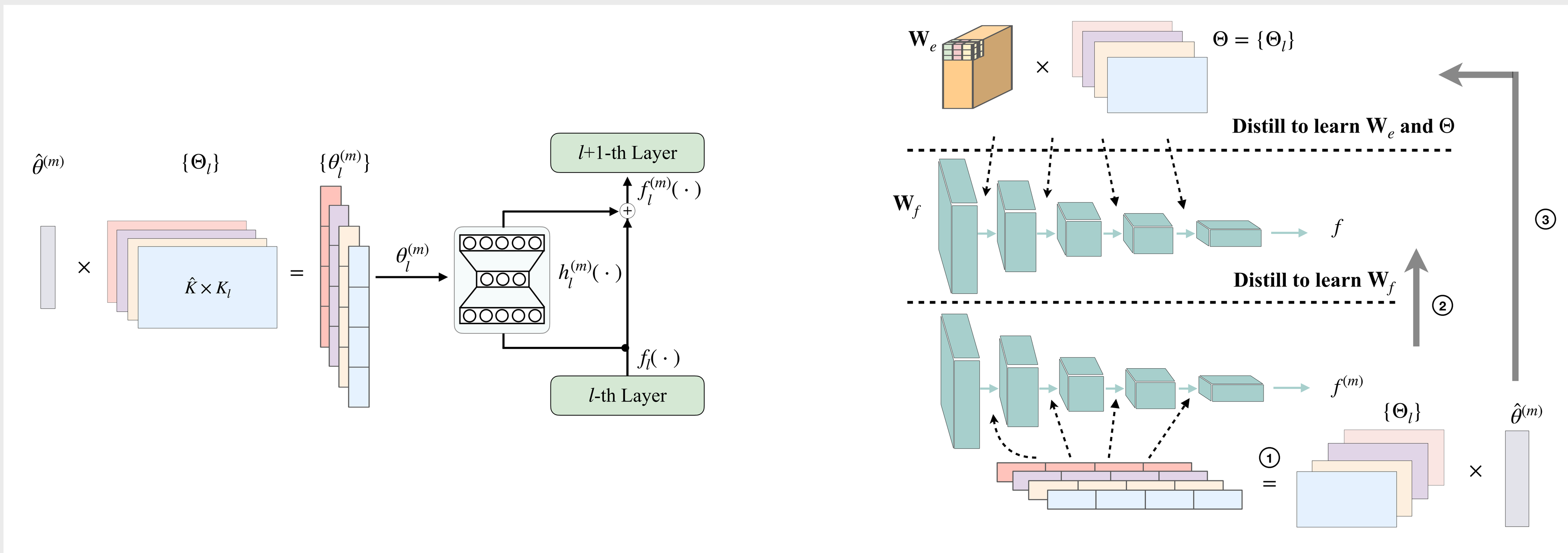
What the Collaboration Brings?

- Device provides us new opportunities to adapt models.
- Cloud provides the knowledge to maintain generalization.

Device-Cloud Collaborative Learning



Two Components: MetaPatch and MoMoDistill



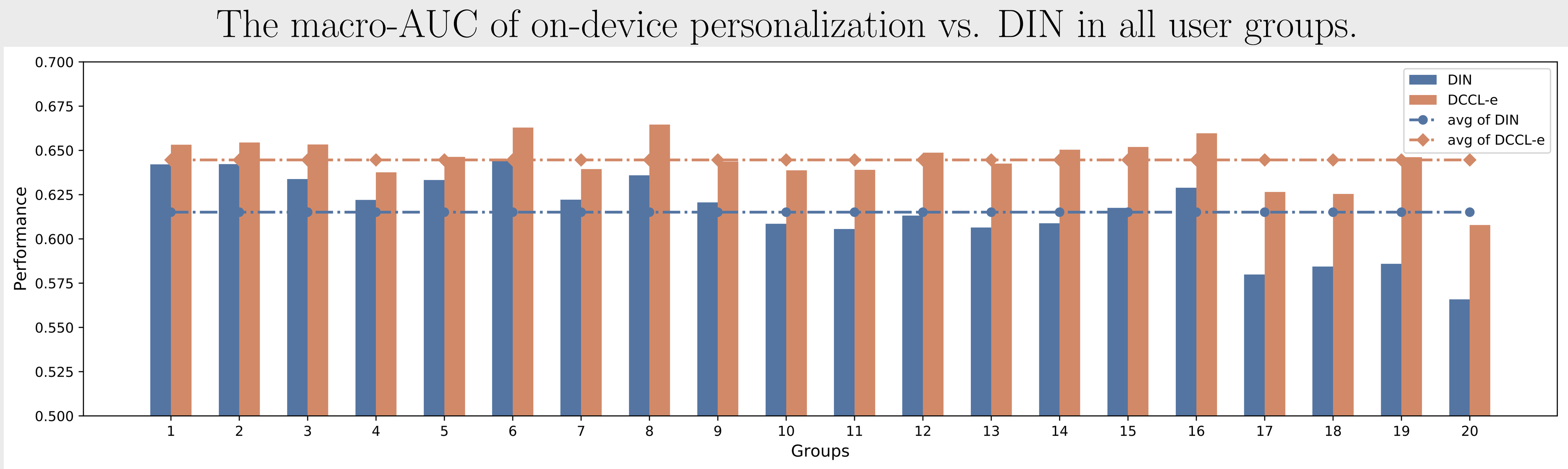
(a) MetaPatch (b) MoMoDistill

Compare with Cloud-based Models

Datasets	Metric	MF	FM	NeuMF	DeepFM	SASRec	DIN	DCCL	Improv.
Amazon	HitRate@1	23.69	21.53	26.10	25.43	26.53	26.56	**26.94**	1.43%
	HitRate@5	35.74	36.74	42.98	42.48	44.22	44.00	**44.79**	1.29%
	HitRate@10	44.38	47.90	52.32	53.51	54.94	55.43	**56.59**	2.09%
	NDCG@5	29.83	29.17	34.74	34.12	35.60	35.48	**36.95**	3.79%
	NDCG@10	32.61	32.77	37.76	37.67	39.07	39.22	**40.45**	3.14%
MovieLens-1M	HitRate@1	14.60	14.90	16.45	15.41	34.85	37.45	**38.69**	3.31%
	HitRate@5	44.85	47.13	46.24	47.35	69.17	70.71	**71.97**	1.78%
	HitRate@10	63.54	64.40	65.36	65.46	80.69	81.25	**82.23**	1.21%
	NDCG@5	29.87	30.27	31.71	31.82	53.18	55.22	**56.43**	2.19%
	NDCG@10	35.89	36.49	37.90	37.68	56.94	58.65	**59.77**	1.91%
Taobao	HitRate@1	24.88	25.29	29.11	33.28	35.19	52.17	**55.71**	6.79%
	HitRate@5	50.83	51.18	55.42	57.26	60.13	68.12	**70.31**	3.21%
	HitRate@10	62.28	63.96	65.78	66.09	69.30	74.80	**76.70**	2.54%
	NDCG@5	38.46	38.80	43.03	46.09	48.52	60.65	**63.42**	4.57%
	NDCG@10	42.17	42.93	46.40	48.95	51.50	62.81	**65.49**	4.27%

Ablation Study

The macro-AUC of on-device personalization vs. DIN in all user groups.



Measure	One-round DCCL vs. DIN.			One-round DCCL with different patch positions.		
	DIN	DCCL-e	DCCL-m	1st Junction	2nd Junction	3rd Junction
HitRate@1	52.17	55.03	**55.71**	53.26	**54.10**	52.36
HitRate@5	68.12	70.03	**70.31**	68.89	**69.36**	68.14
HitRate@10	74.80	76.46	**76.70**	75.54	**75.86**	74.85
NDCG@5	60.65	62.99	**63.42**	61.56	**62.19**	60.74
NDCG@10	62.81	65.07	**65.49**	63.71	**64.29**	62.92