上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

电子信息与电气工程学院
School of Electronic Information and Electrical Engineering

“智能网络计算与信息安全”暑期学校

移动端智能计算

吴 帆

计算机科学与工程系

上海交通大学

1 研究背景

移动端智能设备的源起

第一台移动机器人
Shakey the robot

第一架现代军事无人机
Tadiran Mastiff

第一辆自动驾驶
汽车Navlab

CPU: 16MHz
RAM: 1MB

第一部智能手机
IBM Simon

第一块Linux
智能手表

1966~1970 — 1973 — 1984 — 1994 — 2000

移动端智能设备的发展

CPU: 1 × 800MHz
RAM: 512MB

苹果iPhone 4

谷歌Firefly

CPU: 1 × 3.13GHz
+ 3 × 2.54GHz
+ 4 × 2.05GHz
RAM: 12GB
GPU、NPU

华为P50 Pro

2010 — 2021

大疆无人机

小米手环

Fitbit Ionic手表

巡检机器人

移动端设备种类和数量快速增长，计算、存储等资源日益丰富

端智能的应用

端智能技术

交互式
商品推荐

人脸识别

语音识别

体征观测、
疾病诊断

智能决策

应用场景

智能零售

智能家居

智能交通

智能医疗

智能战场

为什么要发展端智能

云智能约束

不能传

2018.5
欧盟《通用数据保护条例》
2021.6
《数据安全法》

- 数据隐私
- 数据安全

传不了

- 数据规模大
- 特征丰富

算不动

- 终端数量大
- 突发峰值高

端智能优势

性能好

- 实时性高
- 服务模型个性精准

成本低

- 利用终端资源
- 离线自服务

易扩展

- 快速部署
- 大规模扩展

端智能在国家层面的战略作用

美国特朗普政府2018年推出的《出口管制改革法案》将端智能涉及的人工智能和机器学习技术、智能集群技术、情境感知数据分析技术列入了管制清单[1]

[1] https://www.federalregister.gov/documents/2018/11/19/2018-25221/review-of-controls-for-certain-emerging-technologies

端智能在产业界中的重要价值

MIT Technology Review

10 Breakthrough Technologies 2020

Tiny AI

We can now run powerful AI algorithms on our phones.

AI has a problem: in the quest to build more powerful algorithms, researchers are using ever greater amounts of data and computing power, and relying on centralized cloud services. This not only generates alarming amounts of carbon emissions but also limits the speed and privacy of AI applications.

But a countertrend of tiny AI is changing that. Tech giants and academic researchers are working on new algorithms to shrink existing deep-learning models without losing their capabilities. Meanwhile, an emerging generation of specialized AI chips promises to pack more computational power into tighter physical spaces, and train and run AI on far less energy.

These advances are just starting to become available to consumers. Last May, Google announced that it can now run Google Assistant on users' phones without sending requests to a remote server. As of iOS 13, Apple runs Siri's speech recognition capabilities and its QuickType keyboard locally on the iPhone. IBM and Amazon now also offer developer platforms for making and deploying tiny AI.

Tiny AI

Why it matters
Our devices no longer need to talk to the cloud for us to benefit from the latest AI-driven features.

Key players
Google, IBM, Apple, Amazon

Availability
Now

[2]

阿里巴巴淘系技术 原创
2019/11/22 12:03

莫凌、桑杨、明依 作者
阿里巴巴新零售淘系技术部 出品

成交总额

端智能揭秘 | 促使双十一GMV大幅提升，手淘用了什么秘密武器？

导读：信息流作为手淘的一大流量入口，对手淘的浏览效率转化和流量分发起到至关重要的作用。在探索如何给用户推荐其喜欢的商品这条路上，我们首次将端计算大规模应用在手淘客户端，通过端侧丰富的用户特征数据和触发点，利用机器学习和深度神经网络，在端侧持续感知用户意图，抓住用户转瞬即逝的兴趣点，并给予用户及时的结果反馈。通过大半年的不断改进，手淘信息流端上智能推荐在9月中旬全量，并在双十一当天对信息流的点击量和GMV都带来了大幅的提升。下文将给大家分享我们在探索过程中发现的问题，对其的思考和解决方案。

[3]

[2] https://www.technologyreview.com/10-breakthrough-technologies/2020/
[3] https://www.jiqizhixin.com/articles/2019-11-22-4

2 挑战问题及解决方案

移动端智能的挑战问题

设备资源受限且差异化大

如何设计模型结构（单终端）和协同框架（多终端）解决以下问题

1: 模型参数量大 vs. 设备资源有限

2: 设备资源差异化大

数据异质性高

如何设计分布式优化算法消除以下偏差

1: 本地最优模型的聚合结果偏离全局最优

2: 全局模型偏向高可用终端的数据特征

终端用户可靠性低

1: 恶意终端的攻击目标与方式？

2: 如何评估终端对于全局模型的影响？

移动端开发环境欠缺

1: 开发与部署的工程量大、效率低

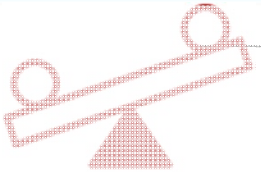
移动端智能的挑战问题

设备资源受限且差异化大

如何设计模型结构（单终端）和协同框架（多终端）解决以下问题

1: 模型参数量大 vs. 设备资源有限

2: 设备资源差异化大

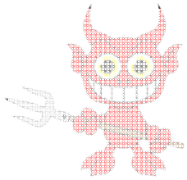


数据异质性高

如何设计分布式优化算法消除以下偏差

1: 本地最优模型的聚合结果偏离全局最优

2: 全局模型偏向高可用终端的数据特征



终端用户可靠性低

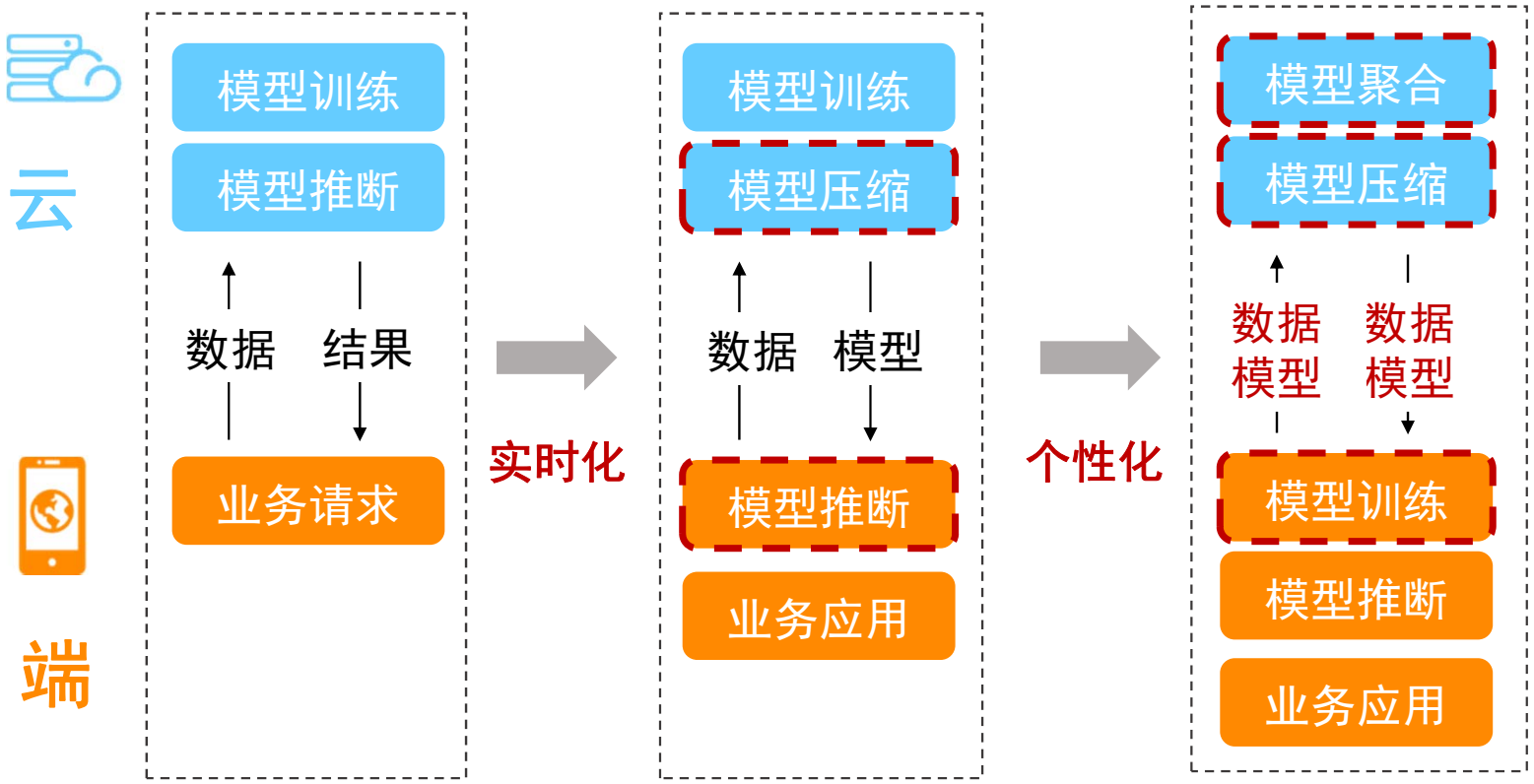
1: 恶意终端的攻击目标与方式？

2: 如何评估终端对于全局模型的影响？

移动端开发环境欠缺

1: 开发与部署的工程量大、效率低

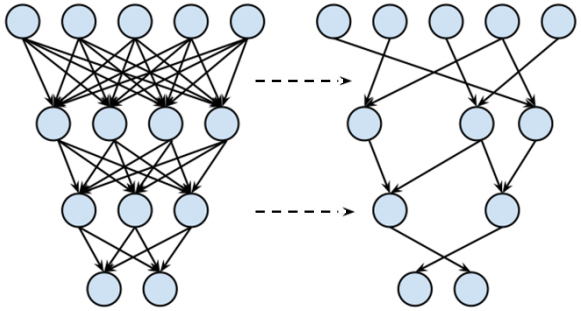
端智能架构发展趋势



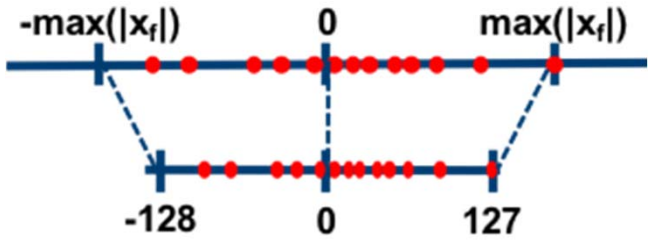
1、云上训练+推断　2、云上训练+端上推断　3、多终端协同训练

单终端 • 模型压缩与嵌套

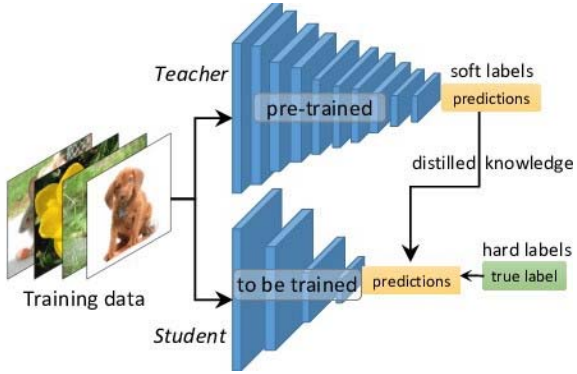
模型压缩



剪枝
[NIPS'15, ICLR'17,
CVPR'17, NIPS'19]

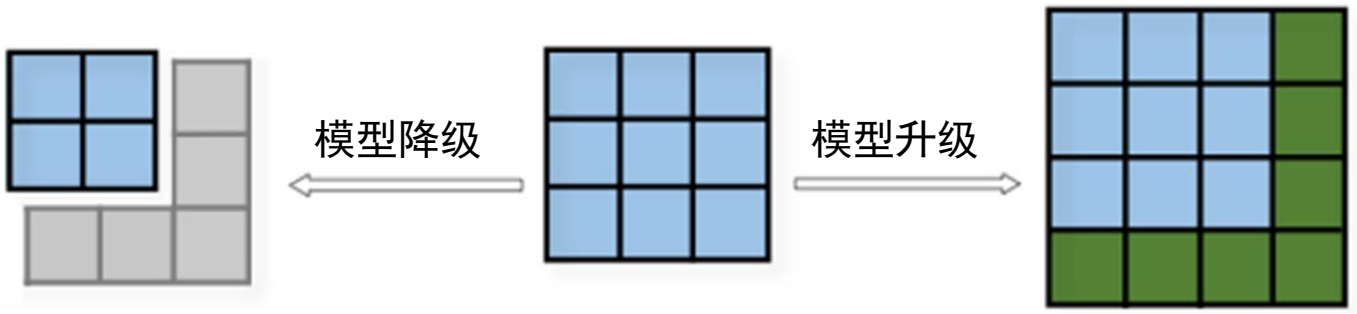


量化
[ICLR'15, ICLR'17,
CVPR'18, CVPR'19]

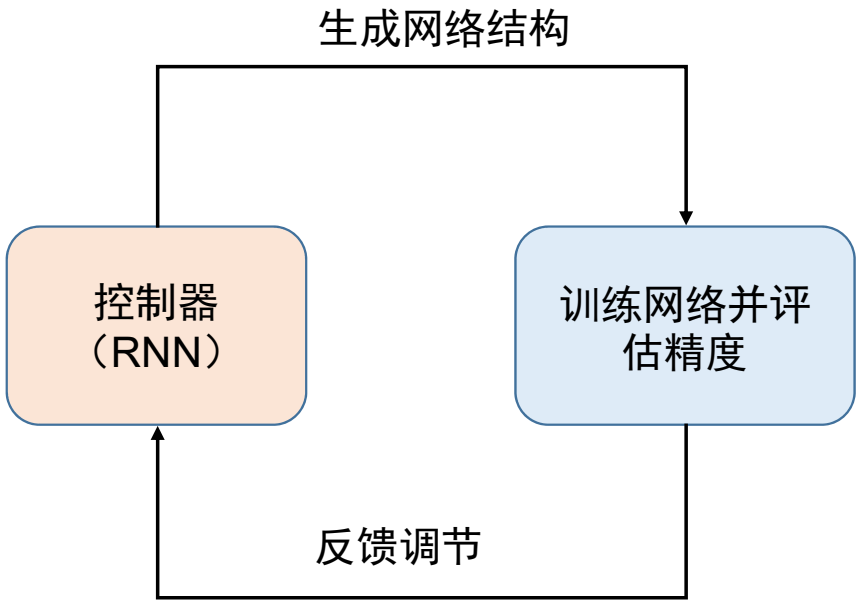


知识蒸馏
[NIPS'14, S&P'16,
NIPS'17, CVPR'19, AAAI'20]

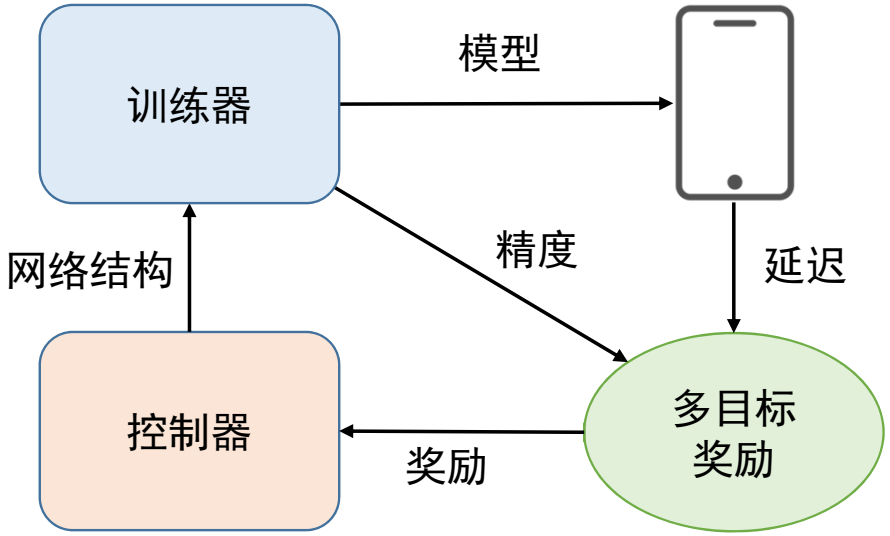
模型嵌套
[MobiCom'18]



单终端 • 模型结构搜索



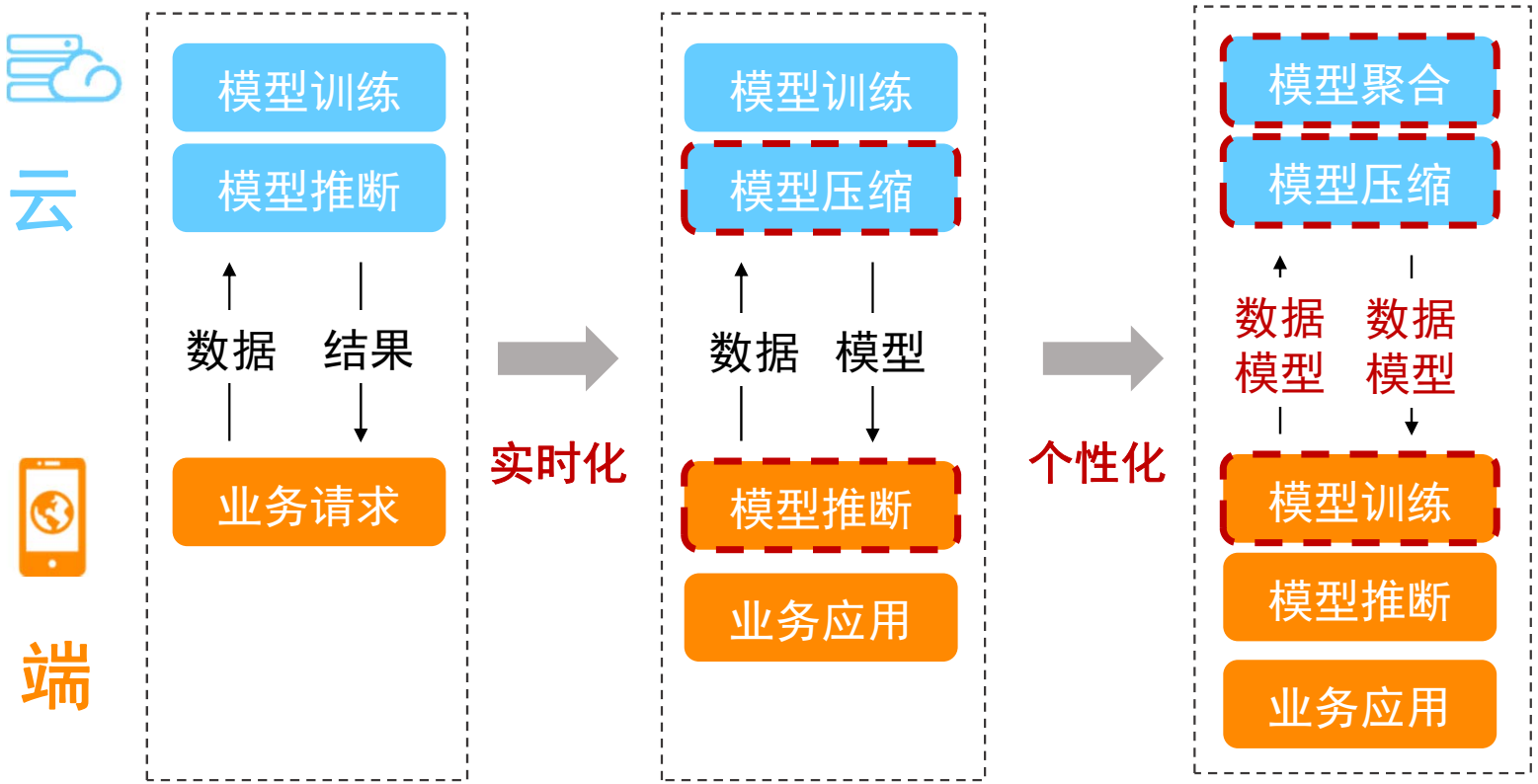
首次提出神经网络结构搜索
[ICLR'17]



针对设备硬件优化
MnasNet [CVPR '19]

根据设备硬件与运行延迟自动适配模型结构

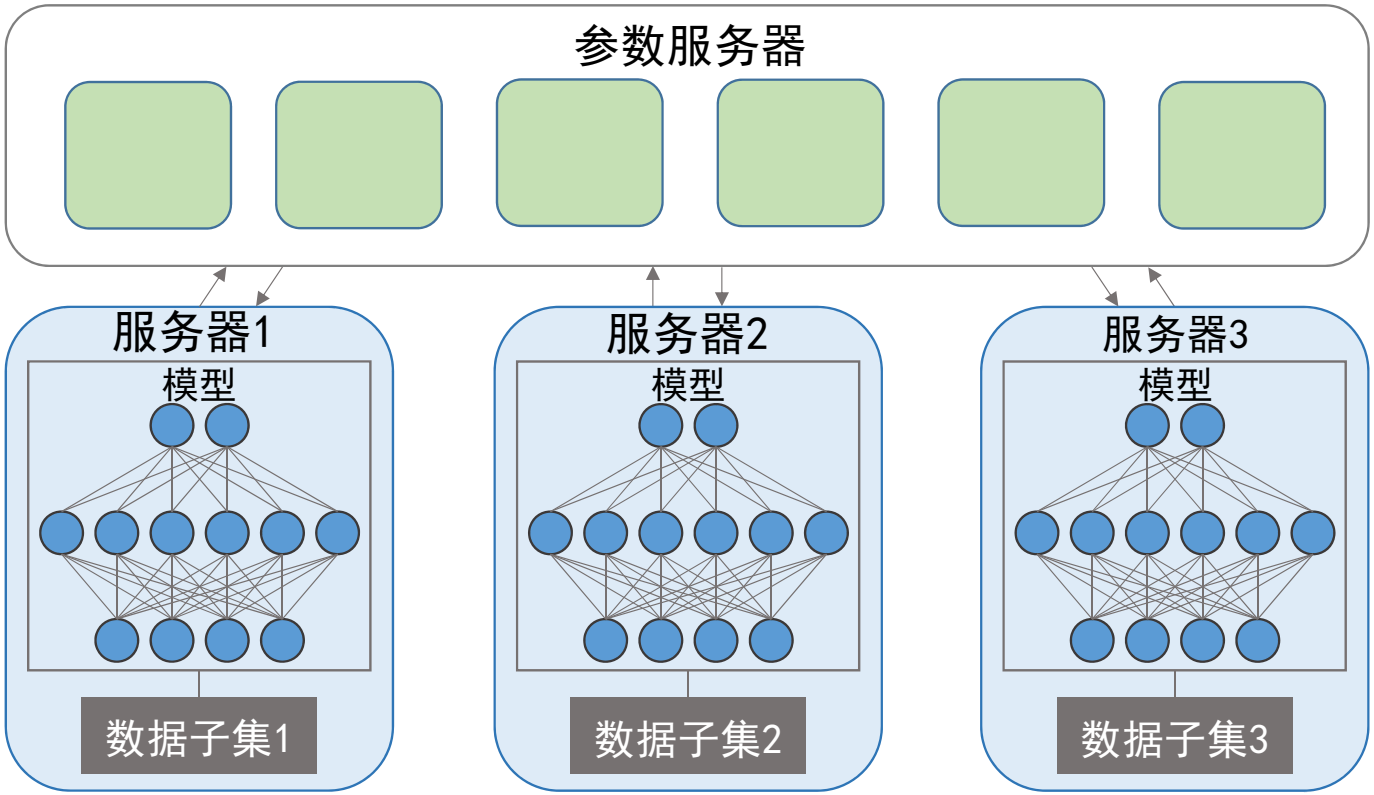
端智能架构发展趋势



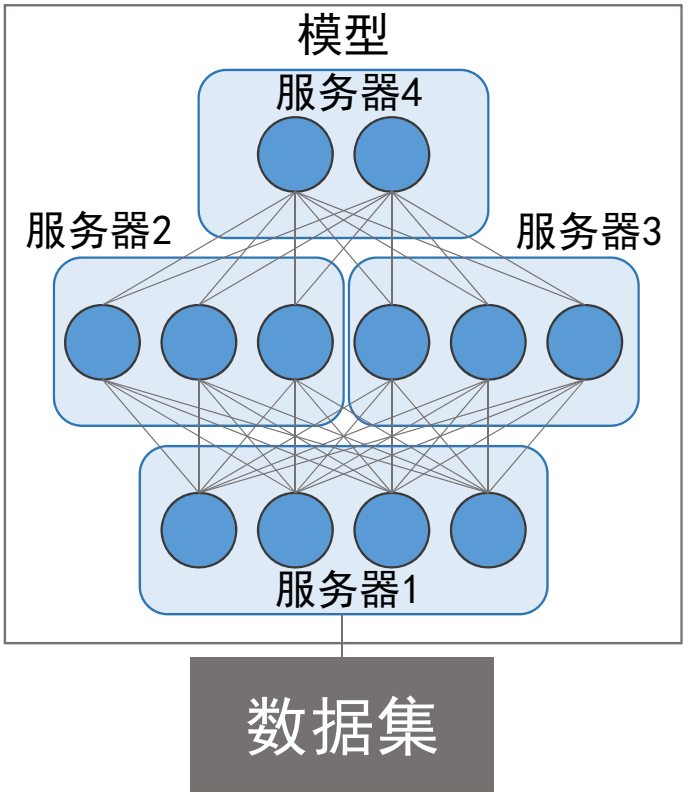
1、云上训练+推断　2、云上训练+端上推断　3、多终端协同训练

多终端协同 • 数据/模型并行

数据并行

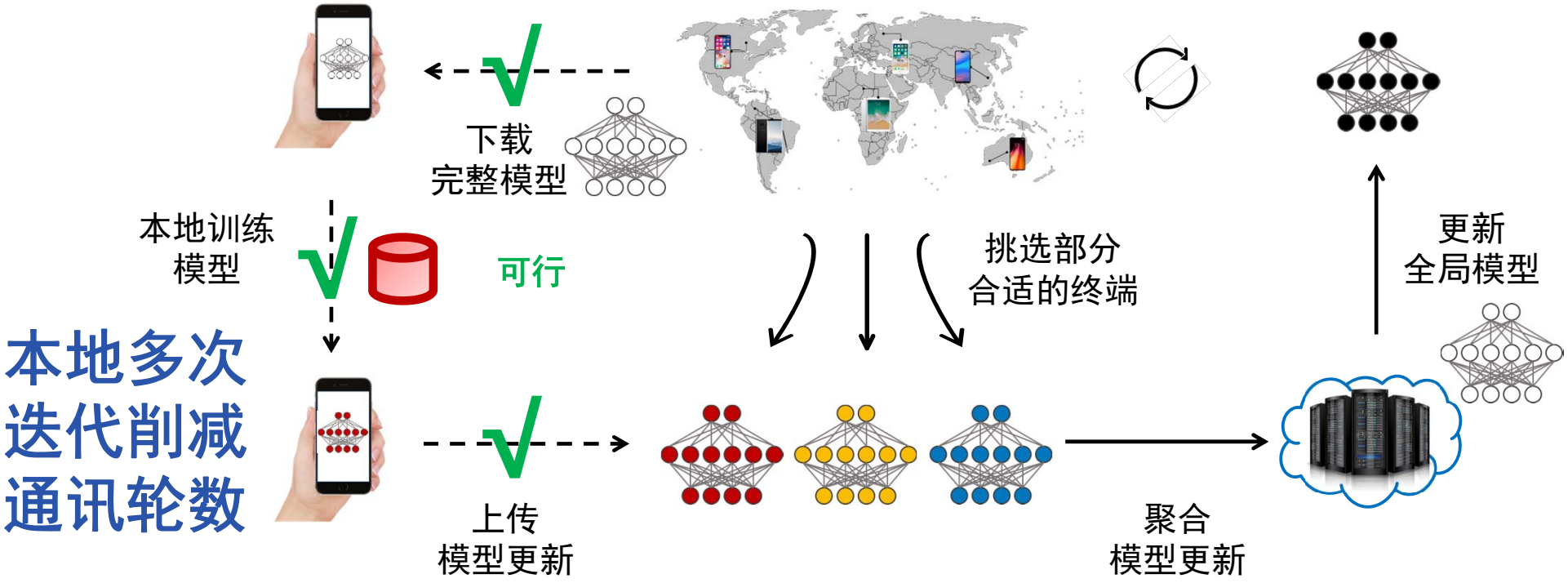


模型并行



传统分布式机器学习方案：随机拆分数据/模型

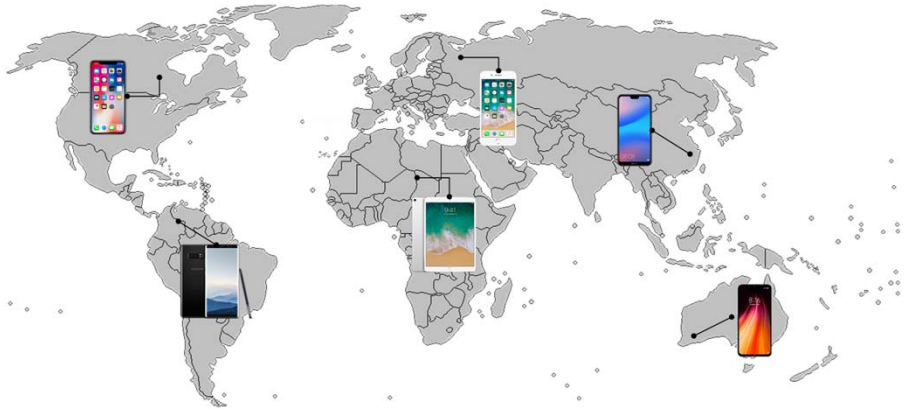
联合（联邦）学习框架 数据并行（自然切分）



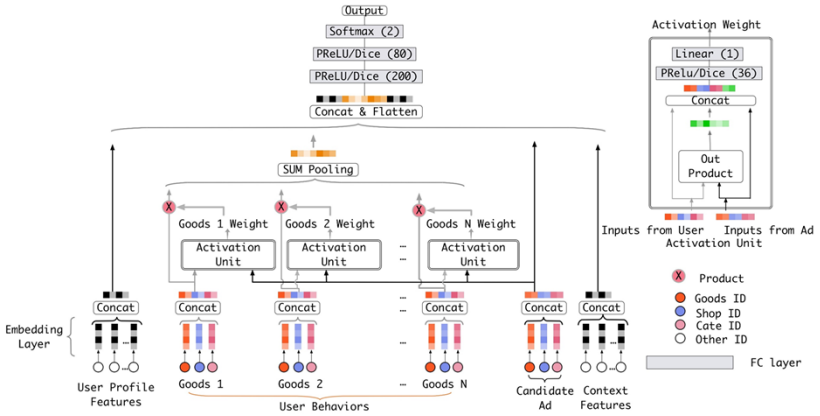
谷歌键盘Gboard的完整语言模型
（嵌入**10K个单词**）

≈ **1.4 MB**

移动端超大规模联合学习



协同训练



资源受限的移动端设备

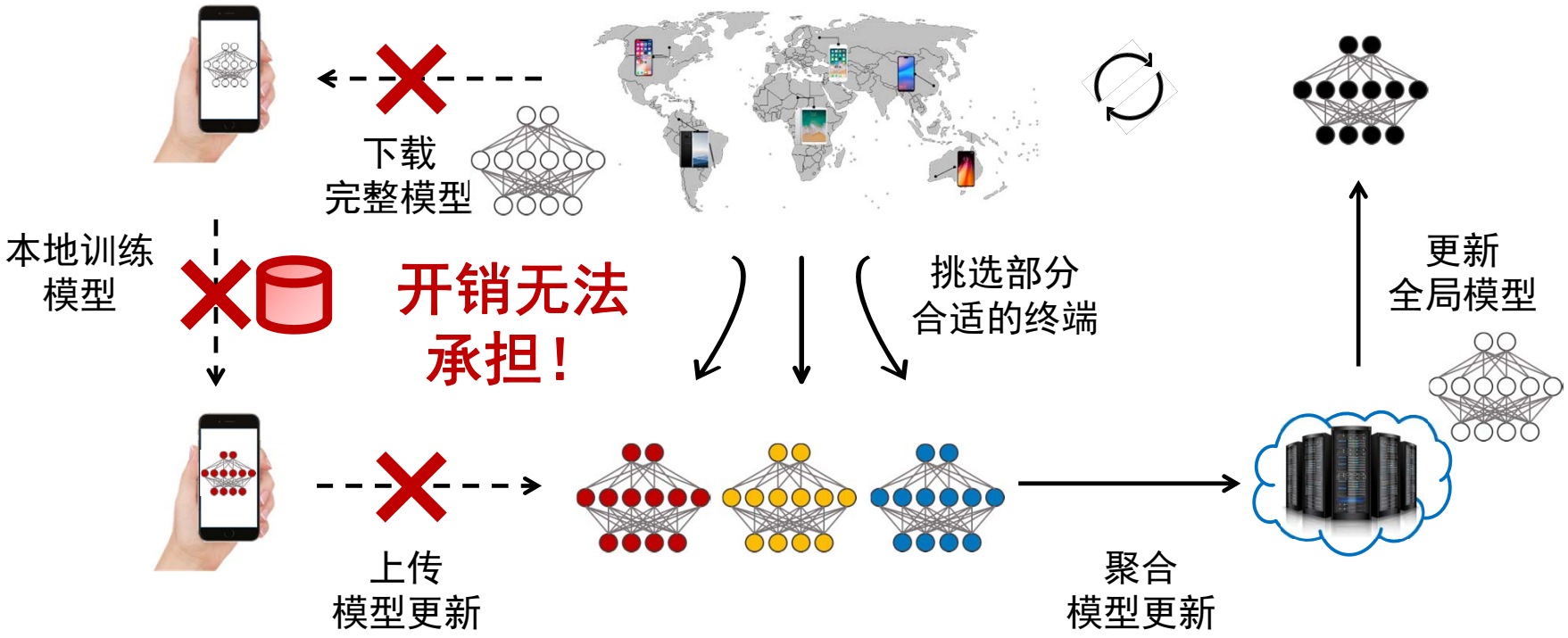
VS

产业级的深度学习模型

10亿移动终端用户、**20亿**候选商品、
200MB应用APP运行内存上限

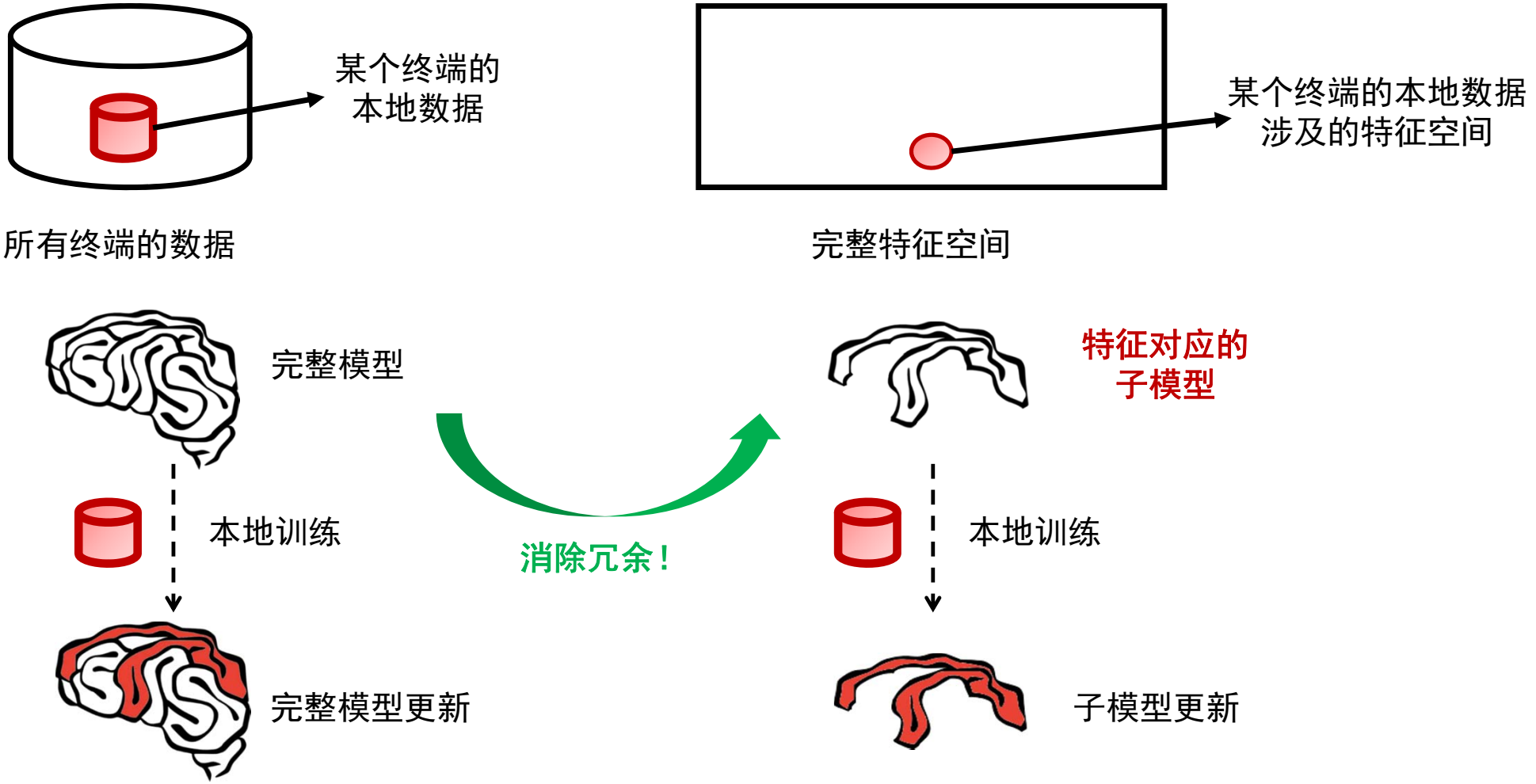
Chaoyue Niu, Fan Wu, Lifeng Hua, Rongfei Jia, Chengfei Lv, Zhihua Wu, and Guihai Chen,
"Billion-Scale Federated Learning on Mobile Clients: A Submodel Design with Tunable Privacy," ACM **MobiCom** 2020.

联合学习框架 不可行

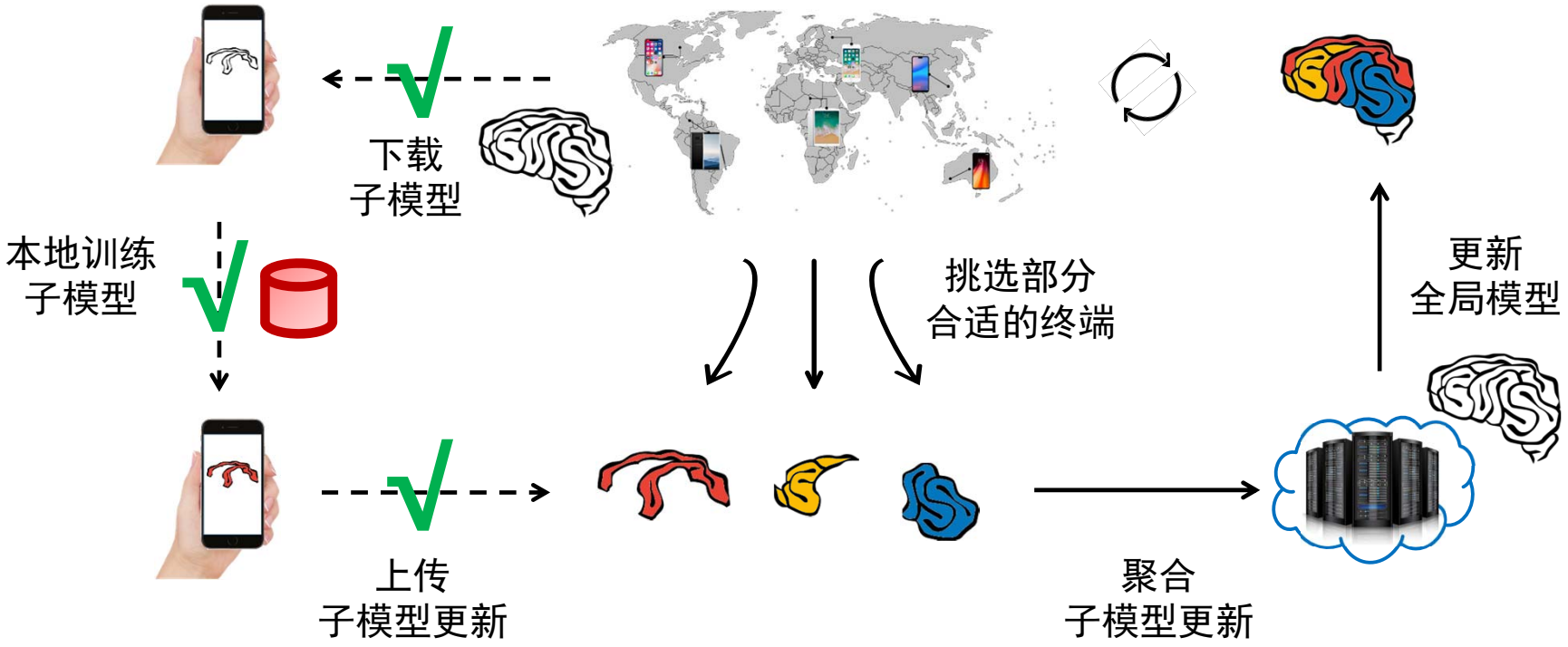


淘宝完整的深度兴趣网络推荐模型
（嵌入全部20亿个商品标识）
> 100 GB

子模型 • 基于特征的模型切分

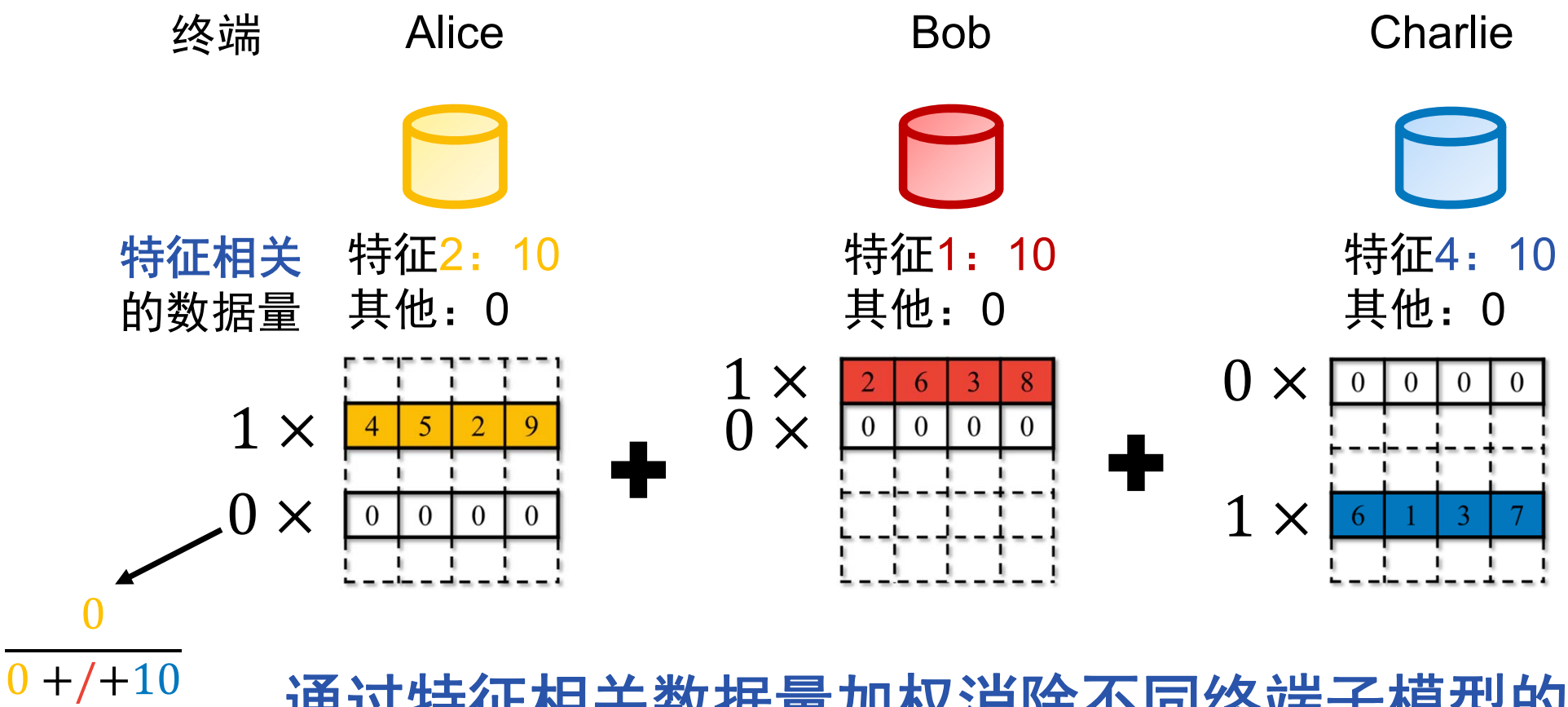


联合子模型学习 数据+模型并行



每个手机淘宝终端的深度兴趣网络子模型
（嵌入本地300个商品标识）
≈ 0.27 MB

联合子模型均值



通过特征相关数据量加权消除不同终端子模型的错位性造成的聚合误差

子模型位置泄露隐私

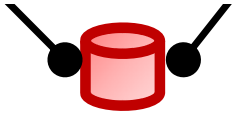
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

Bob真实的子模型位置，也称为“真实索引集”

攻击者（例如不可信的协调服务器）知道Bob的真实索引集，泄露其数据隐私！

0	0	0	0

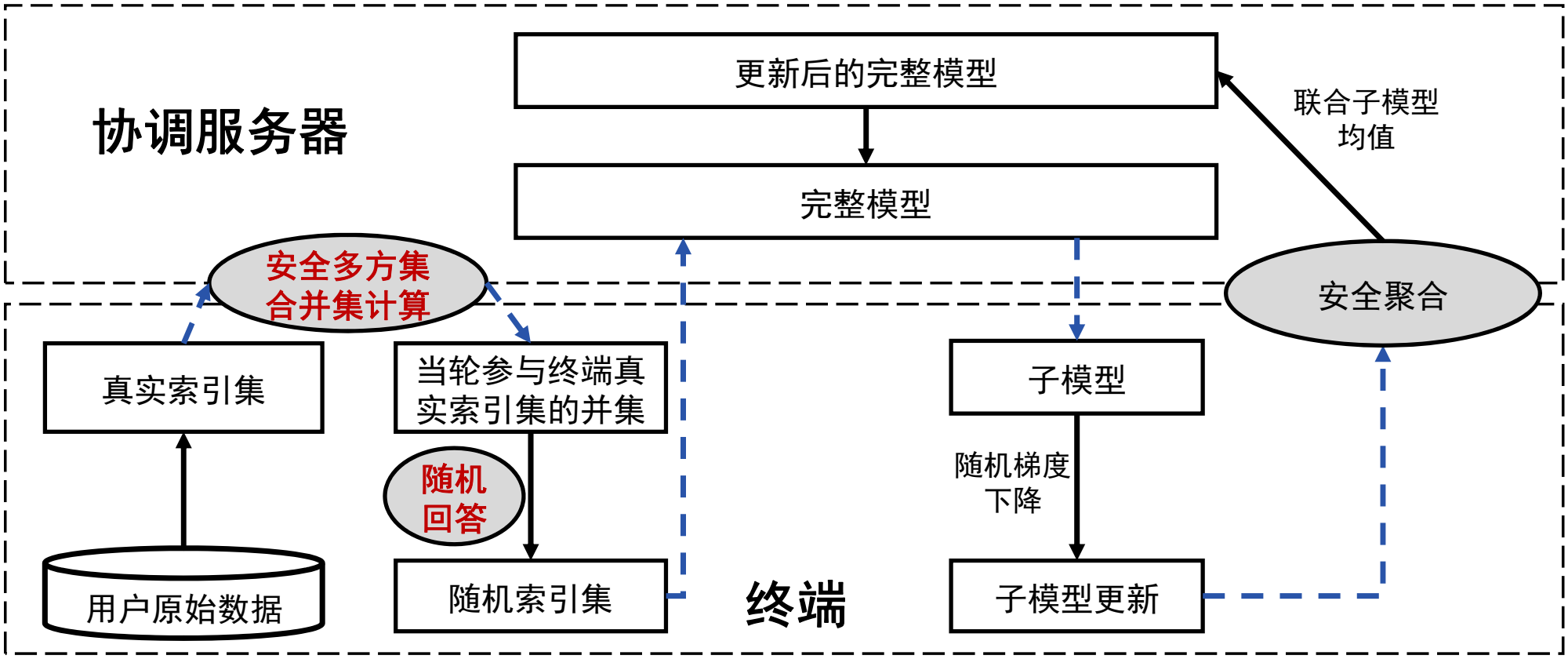
例如，某个淘宝终端本地数据涉及的商品标识集合作为其真实索引集



本地训练

2	6	3	8

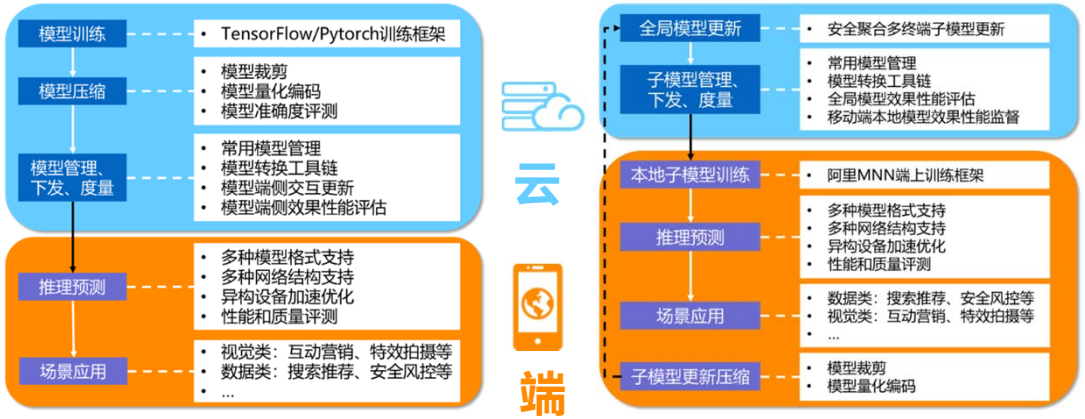
子模型隐私保护机制



赋予终端对其真实子模型位置可调控的抵赖性

联合子模型学习 实测结果

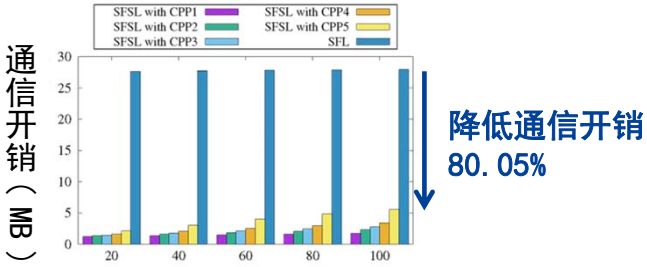
智能刷新、端上重排、意图卡片



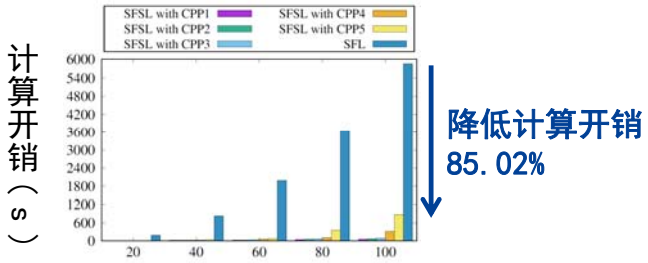
阿里传统“云管端”方案　　新“端-云协同”方案

模型精准化

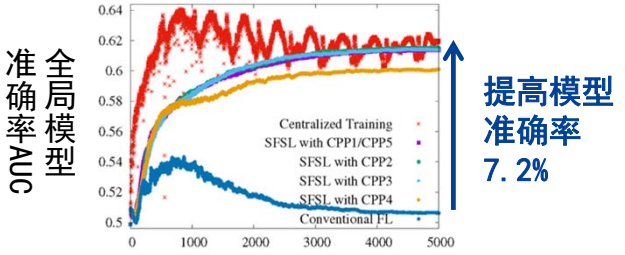
交互实时性



每轮挑选的终端数量



每轮挑选的终端数量



轮数

一种可行的亿级维度深度学习模型端上联合训练方法

联合子模型学习 同行评价

Phillip B. Gibbons
美国卡内基梅隆大学 教授
ACM/IEEE Fellow
PLDI、ISCA最具影响力论文
NSDI 2006、ICDE 2004最佳论文
IPSN 2006程序委员会主席
Yossi Levy Prize

24位顶尖大学教授（含3位ACM/IEEE双Fellow）在共同撰写的综述中表示我们的联合子模型学习是**“富有前景”**的研究方向。

promising

Is it possible to achieve communication-efficient sub-model federated learning while also keeping the client's sub-model choice private? One promising approach is to use PIR for private sub-model download, while aggregating model updates using a variant of secure aggregation optimized for sparse vectors [94, 216, 310].

Advances and Open Problems in Federated Learning

4.4.4 Preserving Privacy While Training Sub-Models

Many scenarios arise in which each client may have local data that is only relevant to a relatively small portion of the full model being trained. For example, models that operate over large inventories, including natural language models (operating over an inventory of words) or content ranking models (operating over an inventory of content), frequently use an embedding lookup table as the first layer of the neural network. Often, clients only interact with a tiny fraction of the inventory items, and under many training strategies, the only embedding vectors for which a client's data supports updates are those corresponding to the items with which the client interacted.

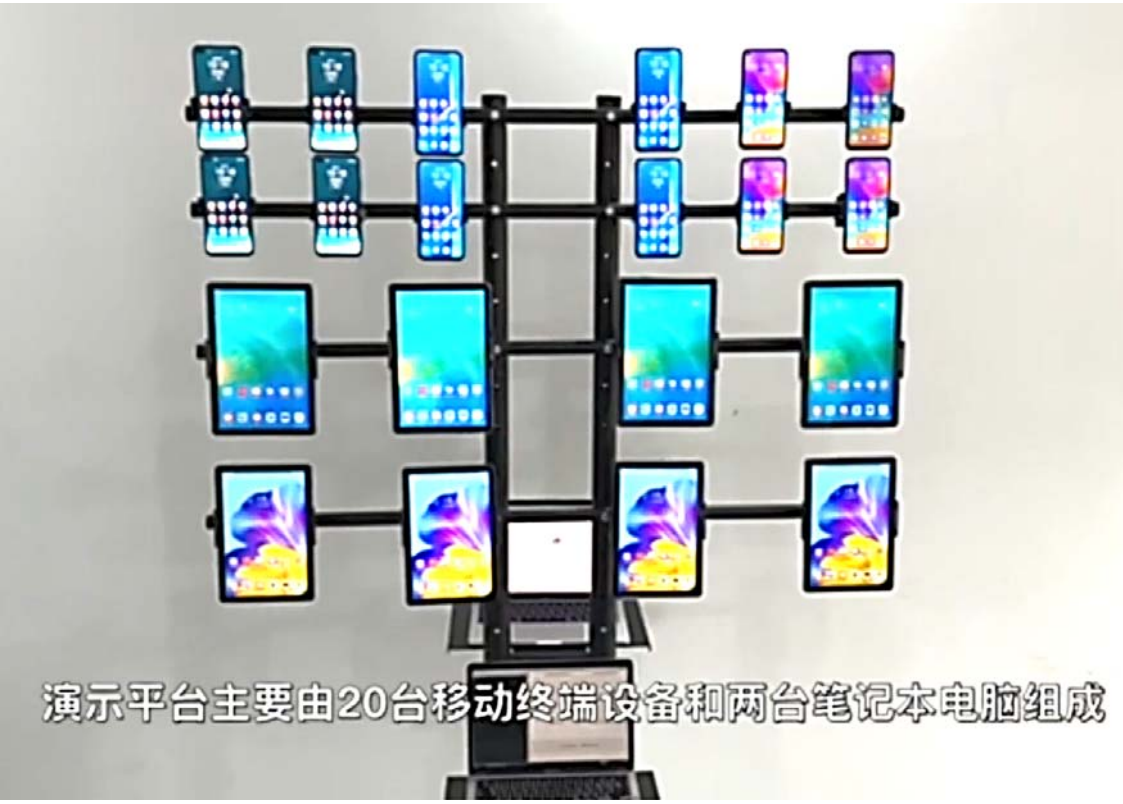
As another example, multi-task learning strategies can be effective approaches to personalization, but may give rise to compound models wherein any particular client only uses the submodel that is associated with that client's cluster of users, as described in Section 3.3.2.

If communication efficiency is not a concern, then sub-model training looks just like standard federated learning: clients would download the full model when they participate, make use of the sub-model relevant to them, then submit a model update spanning the entire set of model parameters (i.e. with zeroes everywhere except in the entries corresponding to the relevant sub-model). However, when deploying federated learning, communication efficiency is often a significant concern, leading to the question of whether we can achieve communication-efficient sub-model training.

Open problems in this area include characterizing the sparsity regimes associated with sub-model training problems of practical interest and developing of sparse secure aggregation techniques that are communication efficient in these sparsity regimes. It is also an open question whether private information retrieval (PIR) and secure aggregation might be co-optimized to achieve better communication efficiency than simply having each technology operate independently (e.g. by sharing some costs between the implementations of the two functionalities.)

Some forms of local and distributed differential privacy also pose challenges here, in that noise is often added to all elements of the vector, even those that are zero; as a result, adding this noise on each client would transform an otherwise sparse model update (i.e. non-zero only on the submodel) into a dense privatized model update (non-zero almost everywhere with high probability). It is an open question whether this tension can be resolved, i.e. whether there is a meaningful instantiation of distributed differential privacy that also maintains the sparsity of the model updates.

端智能联合学习平台


演示平台主要由20台移动终端设备和两台笔记本电脑组成


训练完成后，再将本地模型更新上传至参数服务器进行聚合，
聚合完成并更新全局模型后开始下一轮训练

移动端智能的挑战问题

设备资源受限且差异化大

如何设计模型结构（单终端）和协同框架（多终端）解决以下问题

1: 模型参数量大 vs. 设备资源有限

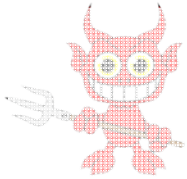
2: 设备资源差异化大

数据异质性高

如何设计分布式优化算法消除以下偏差

1: 本地最优模型的聚合结果偏离全局最优

2: 全局模型偏向高可用终端的数据特征



终端用户可靠性低

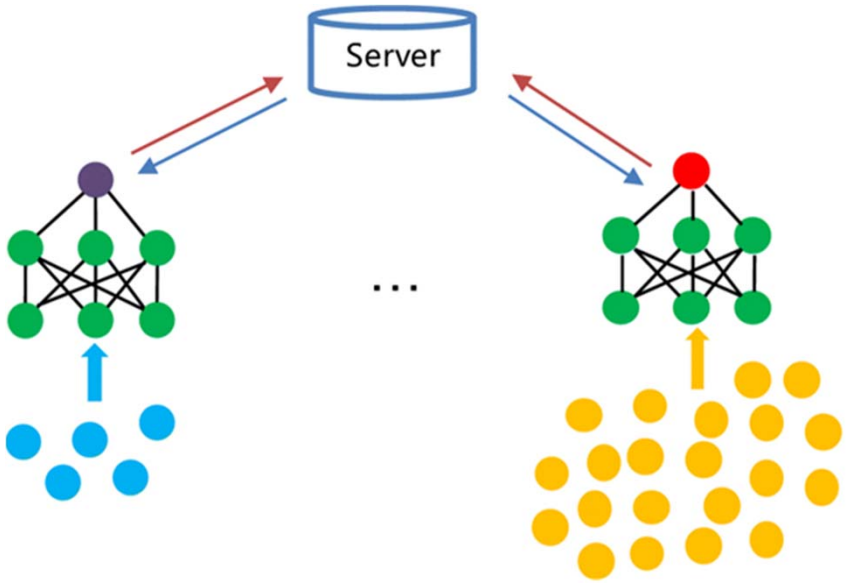
1: 恶意终端的攻击目标与方式？

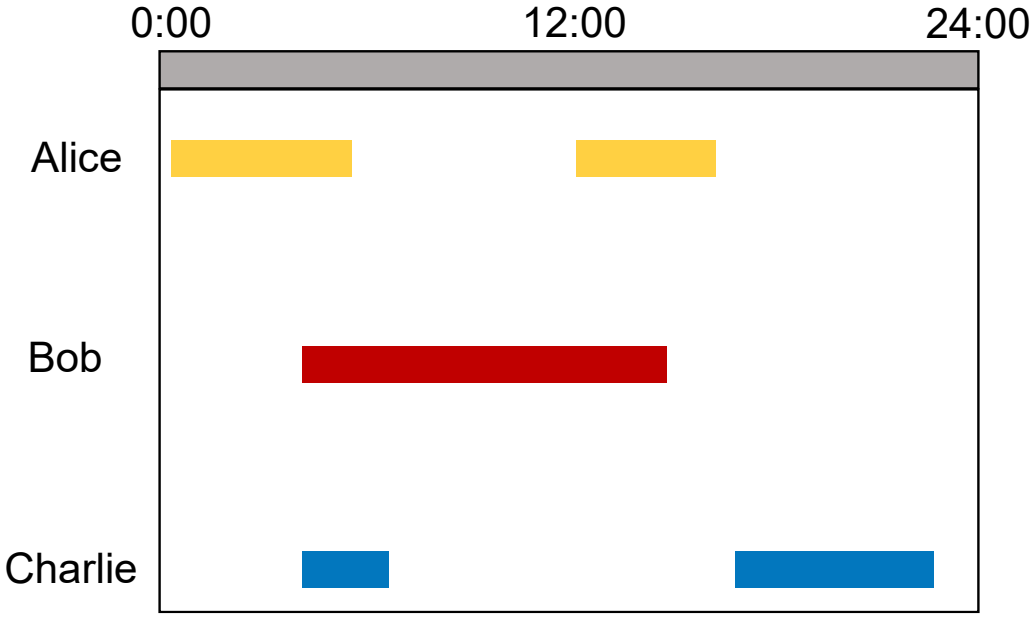
2: 如何评估终端对于全局模型的影响？

移动端开发环境欠缺

1: 开发与部署的工程量大、效率低

数据异质性

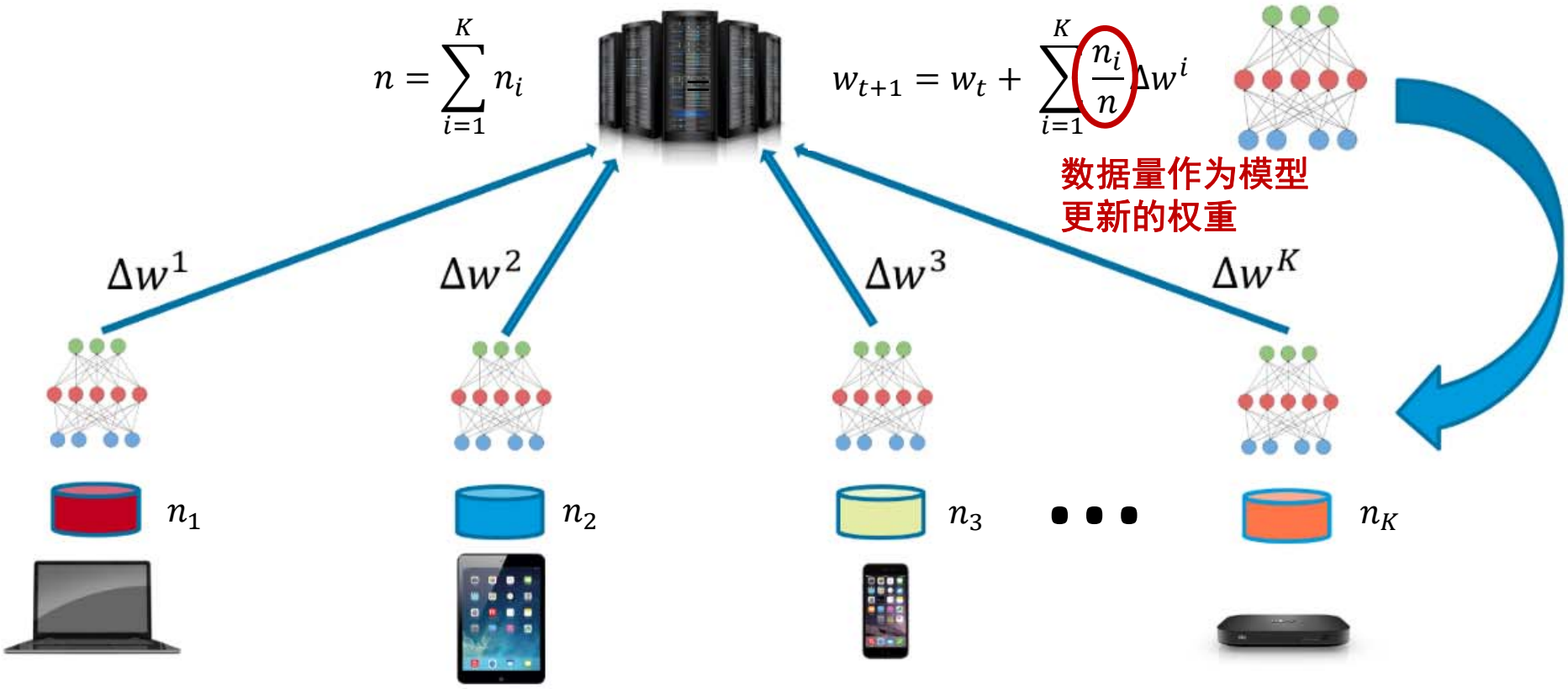




不同终端数据量不均衡、
数据分布不一致

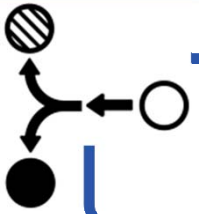
不同终端数据的可用性
不一致

联合均值面临数据量不均衡问题

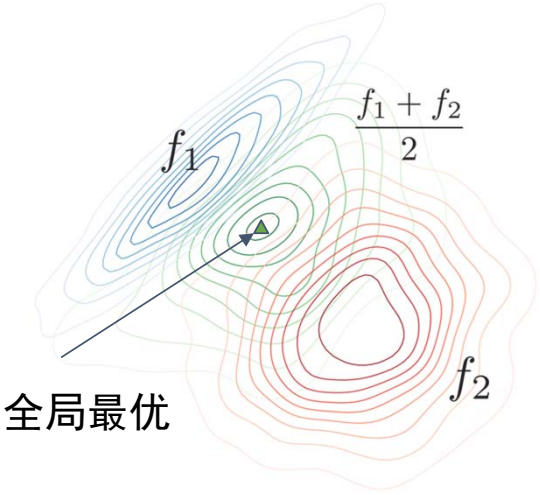


谷歌联合学习的底层优化算法：联合均值 [AISTATS'17]

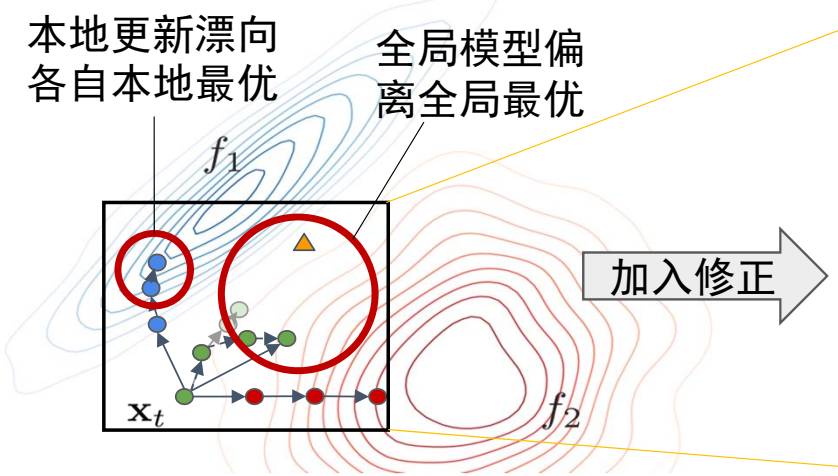
随机受控均值克服非独立同分布



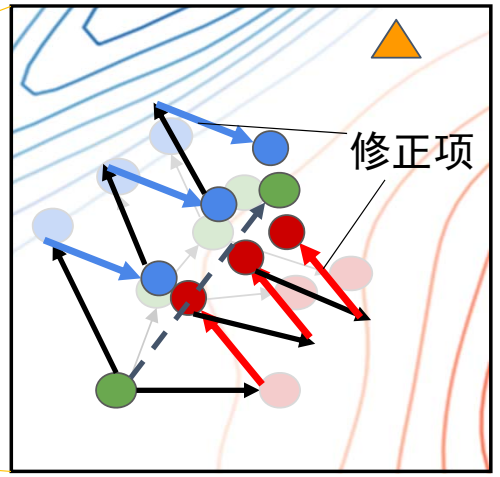
如何避免本地最优的平均偏离全局最优？



目标



实际联合学习的轨迹

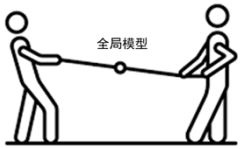


修正后的训练轨迹

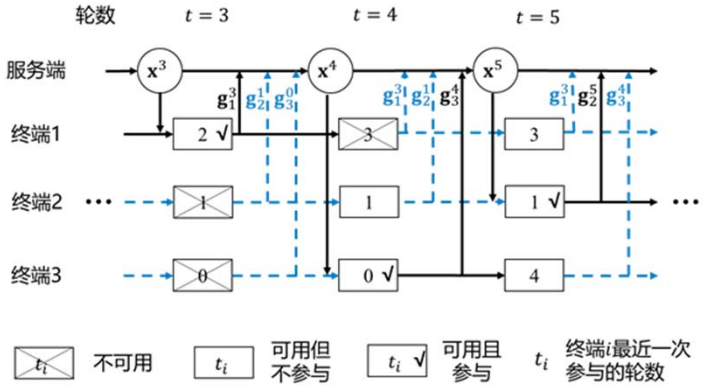
随机受控均值算法 [ICML'19]： $y_i \leftarrow y_i - \eta(g_i(y_i) + \underbrace{c - c_i}_{\text{修正项}})$，$c_i$为终端$i$**上一轮**迭代梯度的均值，$c$为所有$c_i$的平均

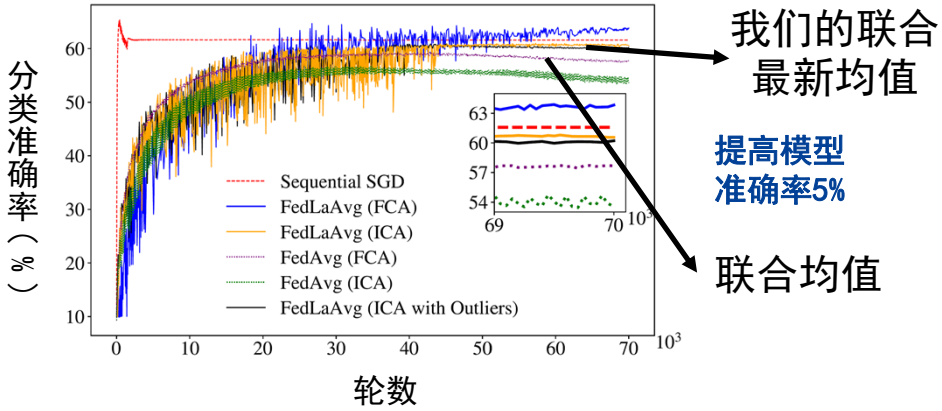
用上轮梯度估计全局方向做本地矫正

联合最新均值解决可用性偏差



如何避免**全局模型偏向高可用终端**？





联合最新均值的的核心思路：

- 每轮**选取最长时间未参与的**部分终端
- **复用**未参与终端**最近一次**提交的**梯度**
- 模拟**异步**梯度聚合消除动态可用偏差

理论上证明了算法的收敛性，
实验验证了模型准确率的提升

Yikai Yan, Chaoyue Niu, Fan Wu, et al., "Distributed Non-Convex Optimization with Sublinear Speedup under Intermittent Client Availability," arXiv: 2002.07399, 2020.

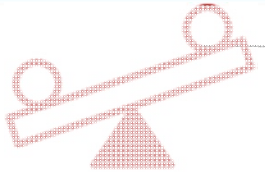
移动端智能的挑战问题

设备资源受限且差异化大

如何设计模型结构（单终端）和协同框架（多终端）解决以下问题

1: 模型参数量大 vs. 设备资源有限

2: 设备资源差异化大



数据异质性高

如何设计分布式优化算法消除以下偏差

1: 本地最优模型的聚合结果偏离全局最优

2: 全局模型偏向高可用终端的数据特征

终端用户可靠性低

1: 恶意终端的攻击目标与方式？

2: 如何评估终端对于全局模型的影响？

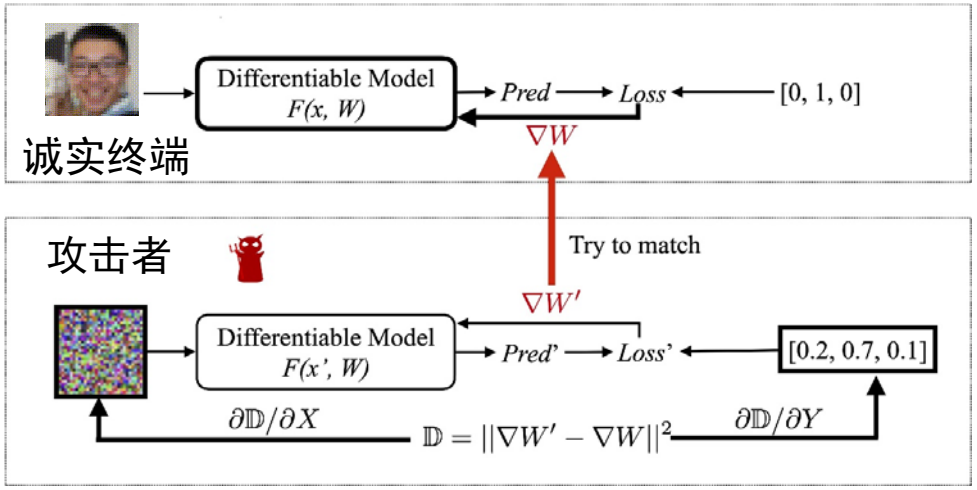
移动端开发环境欠缺

1: 开发与部署的工程量大、效率低

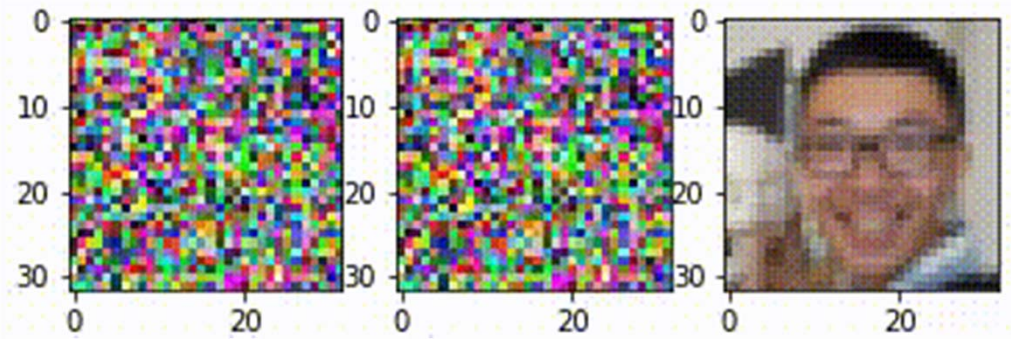
恶意终端 • 模型反演攻击

（聚合的）
模型（更新）

原始数据
（内容、特征、成员信息）



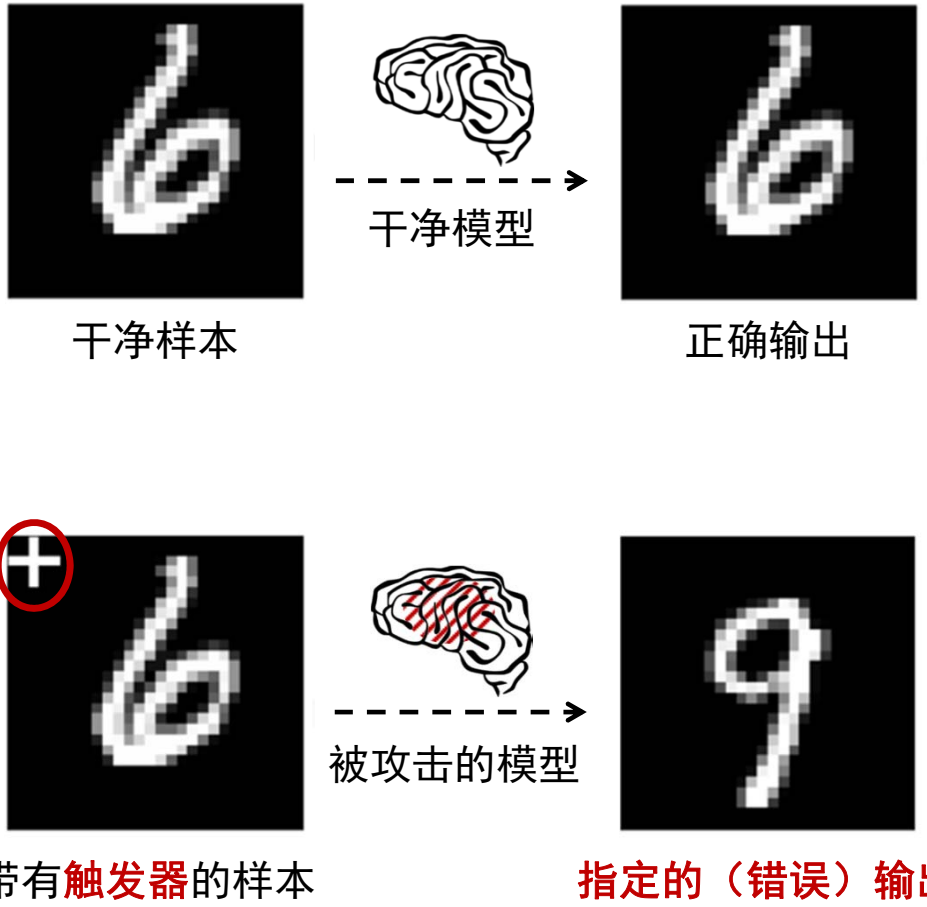
随机初始化的图片　　恢复的图片　　真实的图片

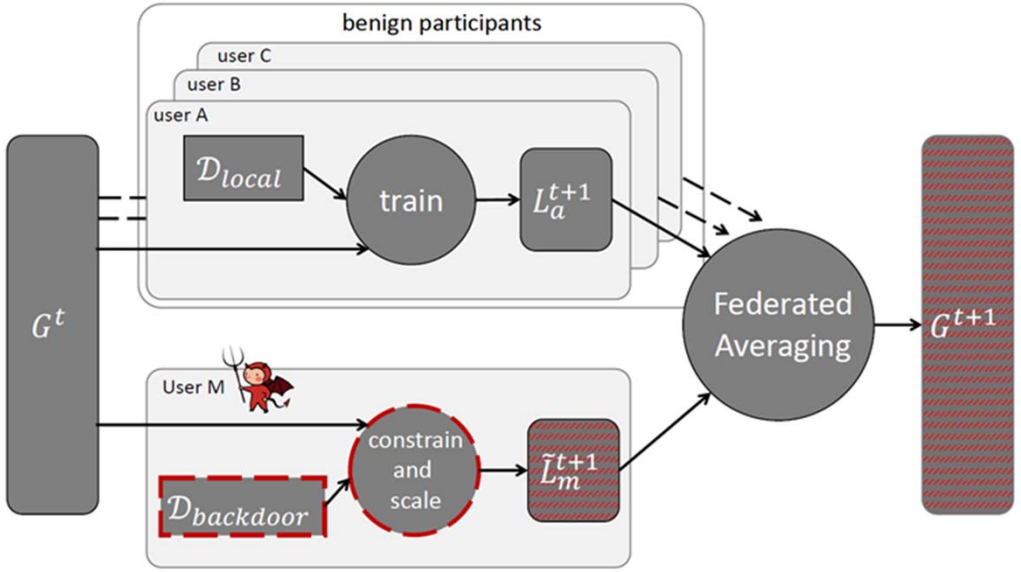


模型反演攻击的主要方式：

- 生成对抗网络 [CCS'17]
- 梯度匹配 [NeurIPS'19]
- 基于多任务学习的主动攻击 [S&P'19]
- 利用梯度多维方向 [NeurIPS'20]

恶意终端 • 后门攻击



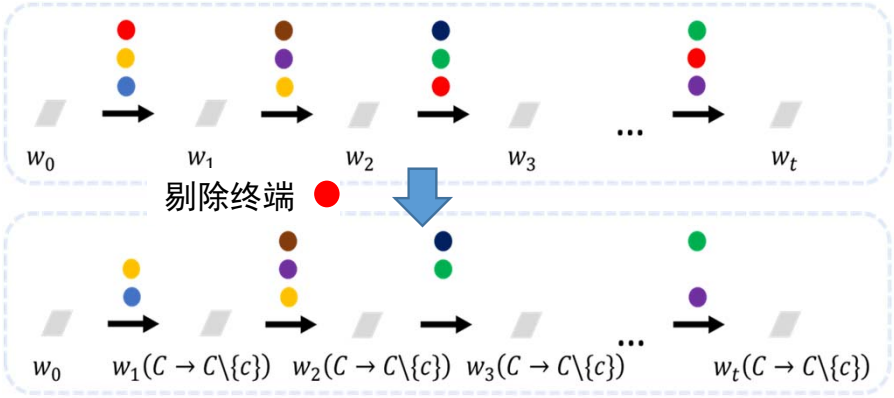


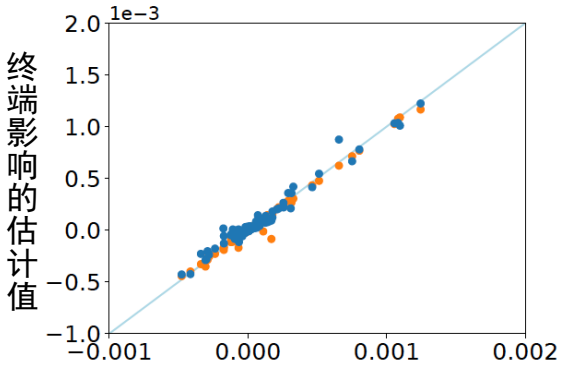
后门攻击的主要方式 [CCS'19, NeurIPS'19, AISTATS'20, ICLR'20]：

- 添加带触发器的恶意样本
- 修改损失函数
- 放大模型更新的参数值

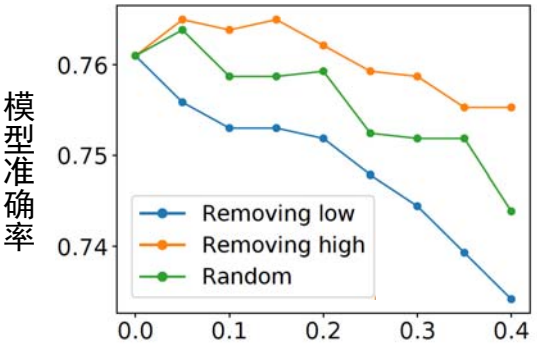
终端的影响度量

如何度量终端对**联合学习模型**的**影响**？





终端对全局模型影响的实际值



动态剔除终端的比例

算法设计的核心思路：

- 采用**leave-one-out**定义终端影响，并设计基于**一阶近似**和**Fisher information**的高效估计算法
- 提出**层次化的模型参数数值检查与截断**方法，降低面向**非凸**优化目标的估计误差

动态剔除产生负面影响的终端可以提升全局模型的准确率！

Yihao Xue, Fan Wu, et al., "Toward Understanding the Influence of Individual Clients in Federated Learning," AAAI 2021.

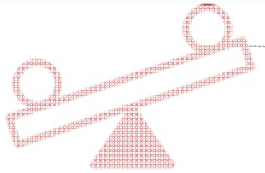
移动端智能的挑战问题

设备资源受限且差异化大

如何设计模型结构（单终端）和协同框架（多终端）解决以下问题

1: 模型参数量大 vs. 设备资源有限

2: 设备资源差异化大

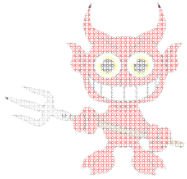


数据异质性高

如何设计分布式优化算法消除以下偏差

1: 本地最优模型的聚合结果偏离全局最优

2: 全局模型偏向高可用终端的数据特征



终端用户可靠性低

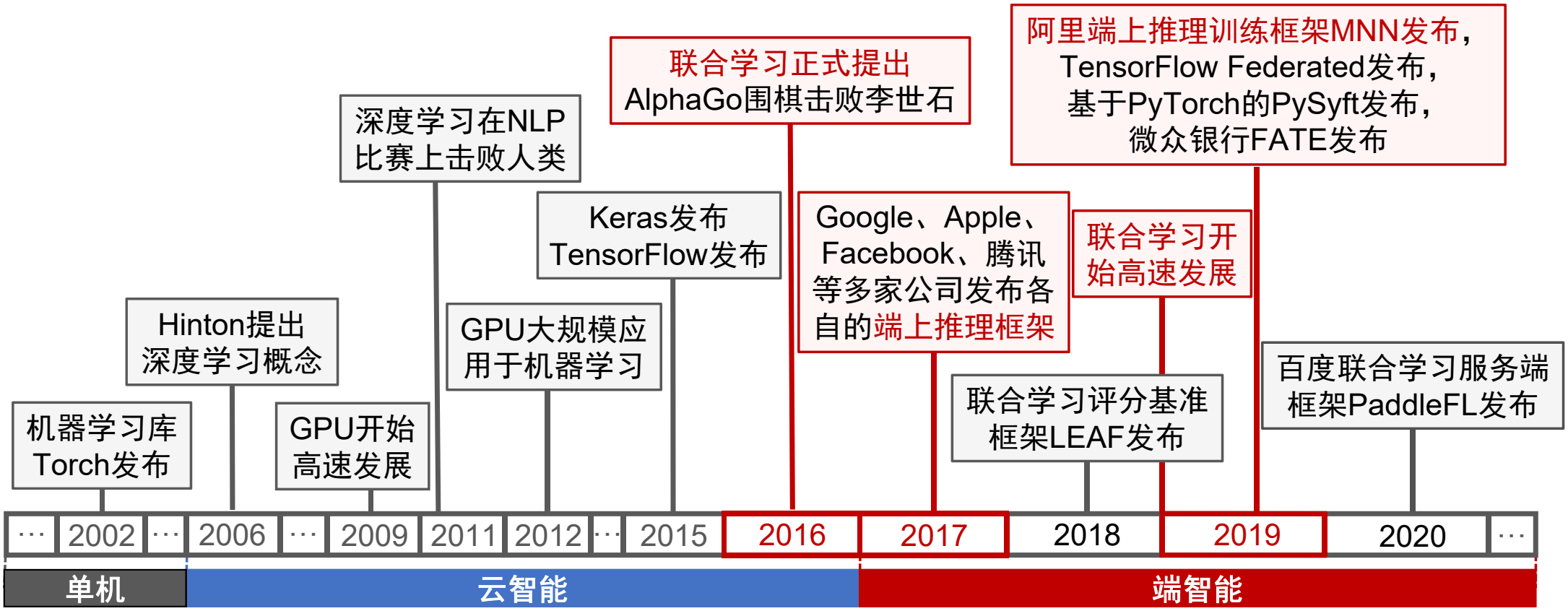
1: 恶意终端的攻击目标与方式？

2: 如何评估终端对于全局模型的影响？

移动端开发环境欠缺

1: 开发与部署的工程量大、效率低

人工智能产业生态的发展



开源代码框架

公司	云上集中式	云上分布式	端上推理	联合（联邦）学习
Google	TensorFlow	TensorFlow	TensorFlow Lite	TensorFlow Federated （实验模拟）
Apple	Create ML	×	Core ML	×
Facebook	PyTorch	PyTorch	PyTorch Mobile	PySyft （实验模拟）
腾讯	×	×	NCNN	FATE （不支持端设备）
百度	Paddle	Paddle	Paddle Lite	PaddleFL （实验模拟）
阿里巴巴	MNN	TensorFlow	MNN	MNN-FSL （待开源）

欠缺端-云融合联合学习开发环境！

3 未来研究方向

未来研究方向

- **高效框架**：同步框架优化、异步框架设计、端-云协同框架探索
- **精准算法**：端上个性化学习算法设计与评价指标、面向多模态异质数据的算法设计
- **可信终端**：攻击行为检测与防御、信誉管理体系建立（奖惩、追溯等）
- **开源框架**：端-云协同联合学习代码框架开源、数据集公开、测试基准

感谢聆听！敬请指正！

更多资源请访问 http://202.120.38.209:31180/

或扫描下方的二维码