AiDD AI+研发数字峰会
AI+ Development Digital summit
第5届 5

大小模型端云协同智能

张圣宇 | 浙江大学软件学院

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing

K+峰会

K+峰会 · 敦煌站
K+ 思考周®研习社
时间：2025.08.29-30

K+峰会 · 上海站
K+ 金融专场
时间：2025.10.17-18

K+峰会 · 香港站
K+ 思考周®研习社
时间：2025.11.25-26

K+峰会详情

AiDD峰会

AiDD峰会 · 上海站
AI+研发数字峰会
时间：2025.05.17-18

AiDD峰会 · 北京站
AI+研发数字峰会
时间：2025.08.08-09

AiDD峰会 · 深圳站
AI+研发数字峰会
时间：2025.11.28-29

AiDD峰会详情

演讲嘉宾

张圣宇

浙江大学平台”百人计划”研究员、博士生导师

浙江大学启真优秀青年学者，研究方向包括大小模型端云协同计算，多媒体计算与推荐系统。在TPAMI、TKDE、KDD、CVPR等CCF A类期刊和会议上发表论文三十余篇。曾获2023年度计算机学会科技进步一等奖，中国人工智能学会CICAI最佳论文奖、2021年WAIC云帆奖-明日之星（全球15人）。承担国家自然科学基金青年基金、副省级市人才工程项目等项目。曾作为主席联合举办国际学术会议ACM MM Asia 大小模型协同进化论坛，全国大模型与决策智能大会“大小模型协同理论与技术”论坛。

目录
CONTENTS

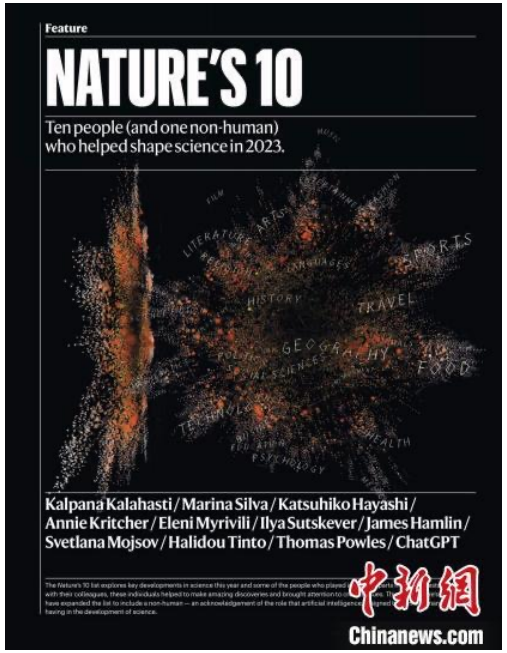
PART 01

大小模型端云协同智能的背景

ChatGPT： 人工智能的IPHONE时刻?

- 2007年1月9日，乔布斯发布第一代iPhone苹果手机，把iPod、电话、移动互联网设备等进行有机整合，推动了移动互联网进入了黄金发展年代。

- 今天大模型给人类社会诸多生产、生活模式带来一次大变革。2023年2月，英伟达创始人兼CEO黄仁勋提出随着ChatGPT为代表的大模型出现，我们已经进入“人工智能的iPhone时刻（iPhone moment of AI）”，这一观点受到美国《财富》杂志、华尔街时报等媒体的广泛认可并转载。



- 《自然》杂志列出2023年度十大人物(Nature’s10)，除了按惯例从全球的重大科学事件中评选出十位人物外，还有一位非人类——人工智能(AI)工具ChatGPT也“抢镜”上榜。

GPT4o

GPT-4o（"o"代表"全能"）是朝着更自然的人类与计算机交互迈出的一步——接受任何文本、音频、图像和视频的组合作为输入，并生成任何文本、音频和图像的组合作为输出。它对音频输入的响应时间最短仅为232毫秒，平均为320毫秒，这与人类在对话中的响应时间相近（在新窗口中打开）。

GPT-4o：成为人的“眼睛”

端云协同计算动机：传统云计算在算法层面围绕特定场景进行端云协同

用户

芯片硬件

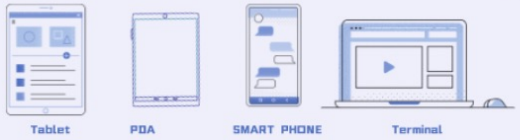
AMD intel NVIDIA arm

编程框架

TensorFlow 飞桨 PyTorch 昇思 MindSpore

碎片化、异构化、算力受限

终端设备



任务场景

视觉类 语言类 语音类 触达类

算法层面协同

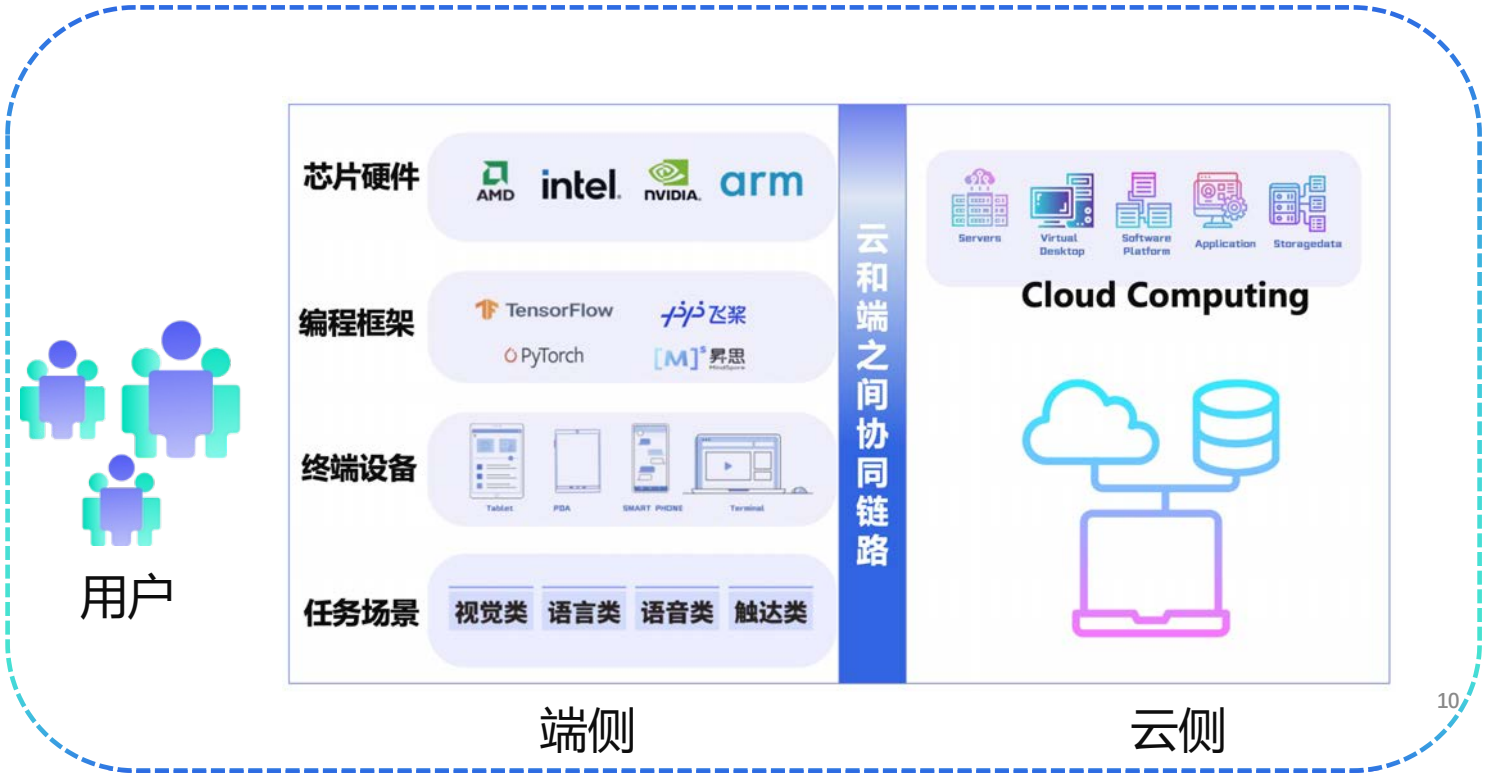
Cloud Computing

延时高、成本大、负载多、部署难

具有靠近用户和数据的天然优势
端侧

具有算力和资源丰富天然优势
云侧

端云协同计算动机：打通端侧推理和云侧赋能之间的协同链路



端侧
推理模型（个性化）

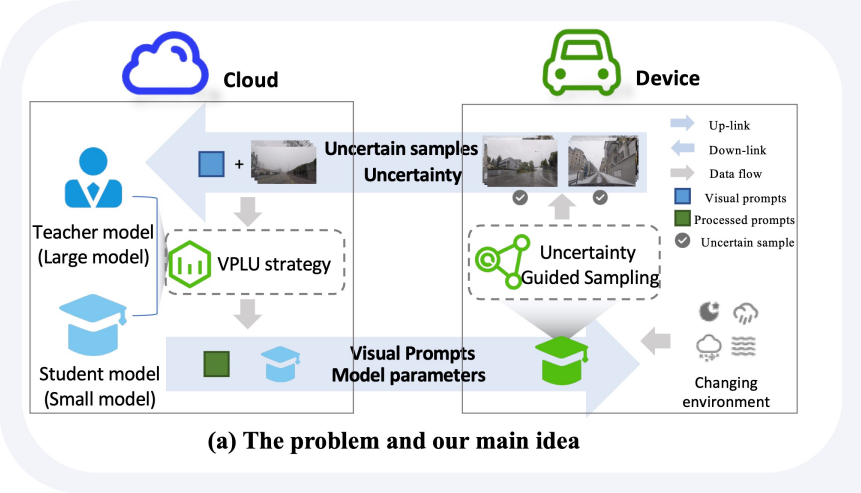
云侧
模型训练（泛化能力）

在端侧和云侧之间为数据、特征、模型和中间结果架构协同链路

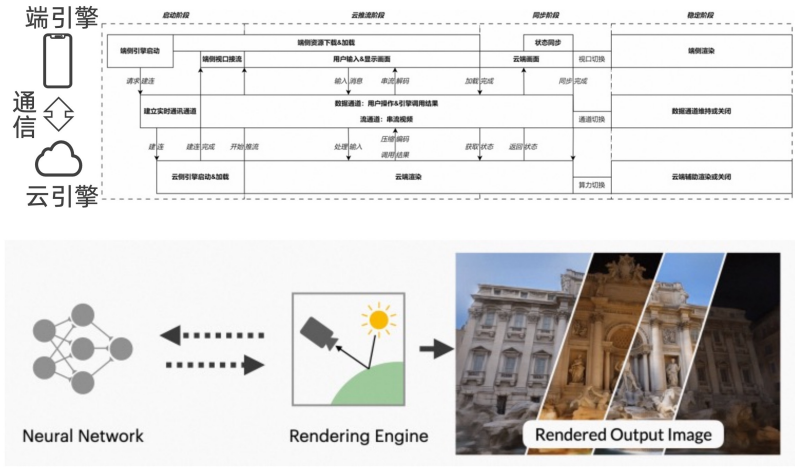
端云异构模型协同计算

- 端云协同计算通过卸载部分学习任务至端侧，让端和云协同完成任务，从而发挥终端靠近用户和数据源的天然优势，降低服务延时至毫秒级，增强模型个性化精准推理能力，缓解云服务器中心负载压力，同时支持用户原始数据在设备本地处理
- 有效克服主流云学习范式在实时性、个性化、负载成本、隐私安全等方面的不足

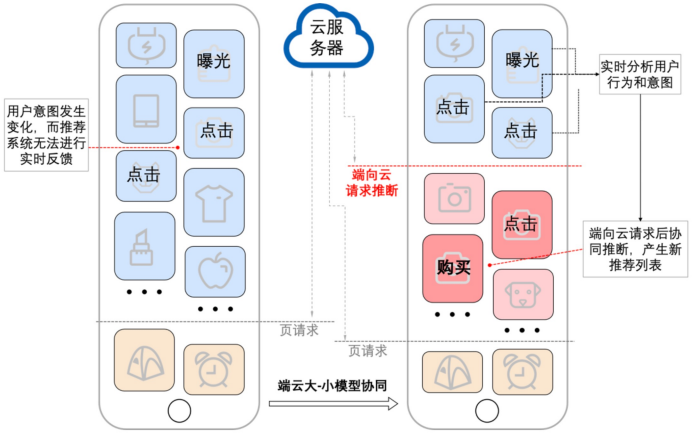
前沿应用



自动驾驶 (Gan et al.)



3D渲染 (Lv et al.)



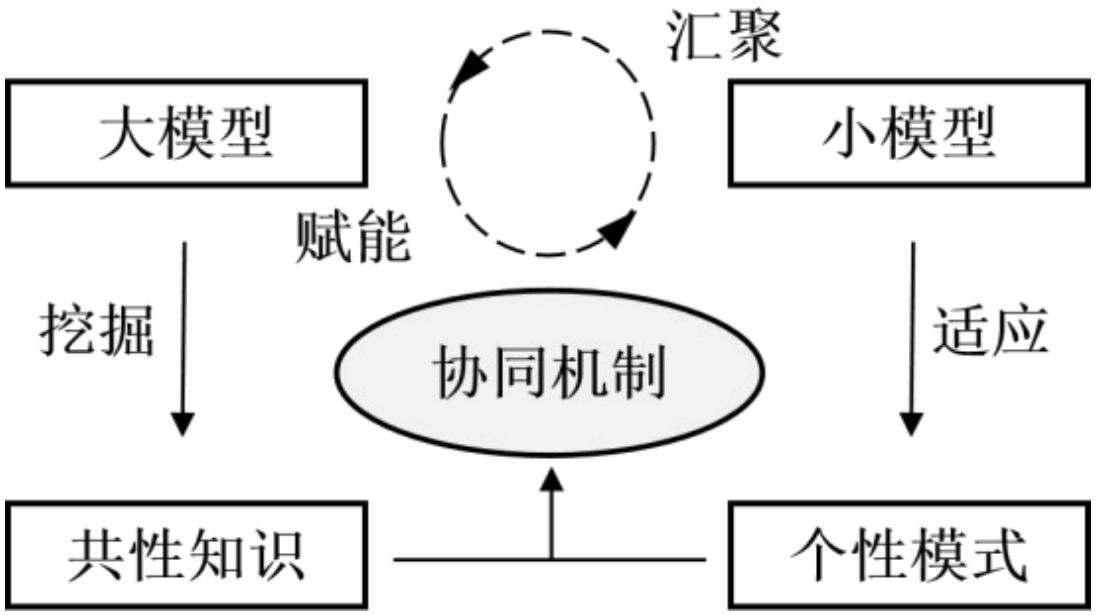
推荐系统 (Qian et al.)

Yulu Gan, Mingjie Pan, Rongyu Zhang, et al.: Cloud-Device Collaborative Adaptation to Continual Changing Environments in the Real-World. CVPR 2023: 12157-12166
Chengfei Lv, Chaoyue Niu, Renjie Gu, et al.: Walle: An End-to-End, General-Purpose, and Large-Scale Production System for Device-Cloud Collaborative Machine Learning. OSDI 2022: 249-265
Xufeng Qian, Yue Xu, Fuyu Lv, Shengyu Zhang, et al.: Intelligent Request Strategy Design in Recommender System. KDD 2022: 3772-3782

大小模型端云协同

核心概念

- 端云协同（Device-Cloud Collaboration）：指边缘设备（如智能手机、IoT设备）模型和云侧服务器模型协同进化推断。
- 云侧大模型（Large Model）：通用认知计算，拥有强大的计算能力、海量的数据、充分的知识库。
- 终端小模型（Small Model）：实时感知、实时响应，运行轻量级任务，响应速度快。



端云异构模型协同计算

研究Taxonomy

云侧中心化协同　　端侧中心化协同　　端云双向协同

- 云侧中心化协同：以云侧为中心进行模型汇聚，端侧仅提供分布式训练数据、计算中间结果和轻量计算资源，如联邦学习
- 端侧中心化协同：以端侧为中心进行模型个性化，云侧仅仅提供模型校正的数据和巨大算力时
- 端云协同：云侧有泛化模型、端侧有个性化模型，且两个模型相互协作学习和推理。


Edge-Cloud Polarization and Collaboration: A Comprehensive Survey

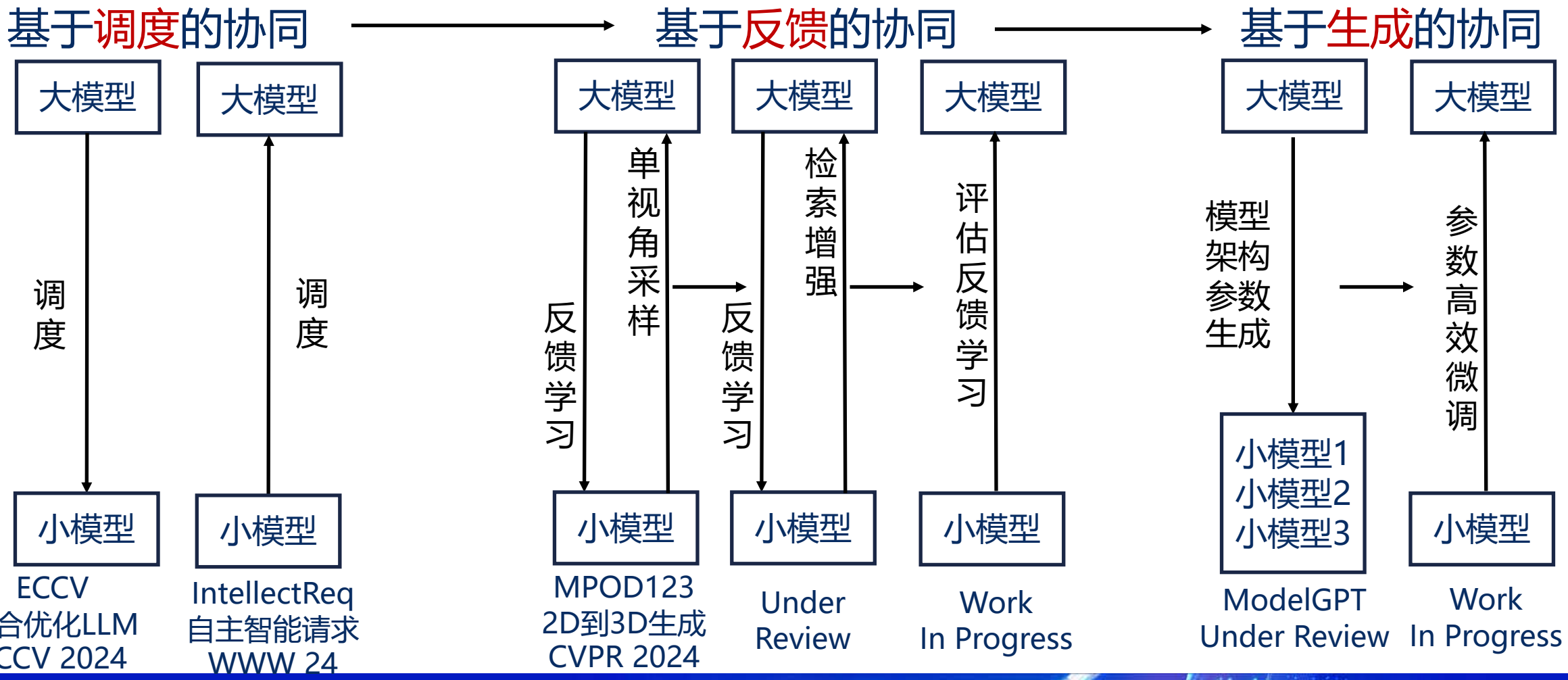
Jiangchao Yao, Shengyu Zhang, et al.
Edge-Cloud Polarization and Collaboration: A Comprehensive Survey, *IEEE Transactions on Knowledge and Data Engineering* (online, DOI: 10.1109/TKDE.2022.3178211)

PART 02

大小模型协同基础算法

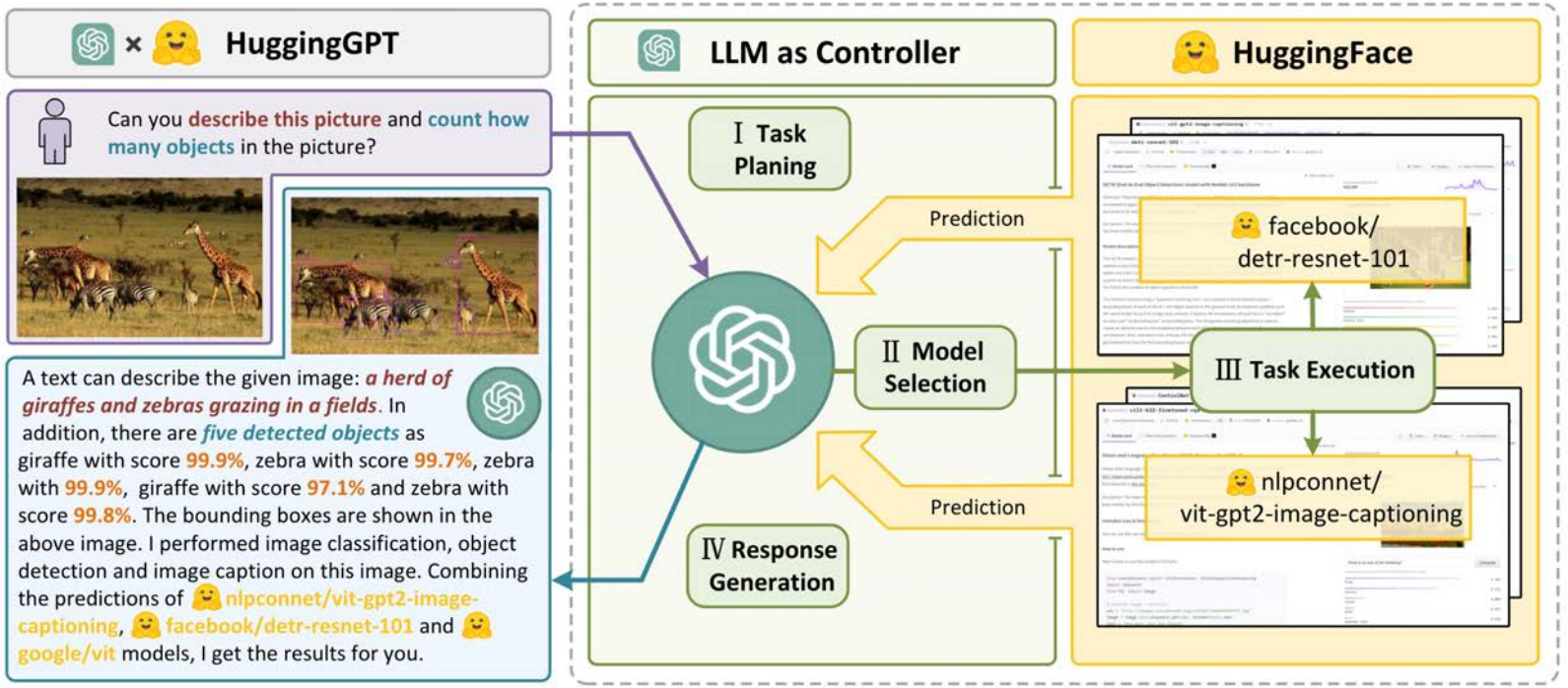
大小模型协同基础算法研究

联合应用平台既有的**特定业务小模型**与**云侧大模型**，将**端侧小模型轻量部署、快速响应、个性适配**的优势，和**云侧大模型认知推理、多模态理解、通用泛化**的优势进行互补



基于调度的协同: 大模型调度小模型

- **大模型驱动的多模态小模型调度与整合 – HuggingGPT**
  - 将ChatGPT作为不同模型之间的桥梁，为不同的任务调度不同的模型。
  - HuggingGPT以ChatGPT为控制器，专家模型为执行者的一个合作系统，实现**任务规划、模型选择、任务执行和回答生成**。
  - 在序列建模、图数据建模等任务上取得优越性能。



基于调度的协同: 小模型调度大模型

小模型调度大模型?

端侧小模型推理不确定性判断，自主请求调度大模型进行赋能

（cf. 大小模型端云协同）

基于反馈的协同：大模型反馈小模型

Observation

View-Conditioned Diffusion

Text-Conditioned Diffusion

MPOD123: One Image to 3D Content Generation Using Mask-enhanced Progressive Outline-to-Detail Optimization. CVPR 2024

基于反馈的协同：大模型反馈小模型

思路方法

- Stage 1

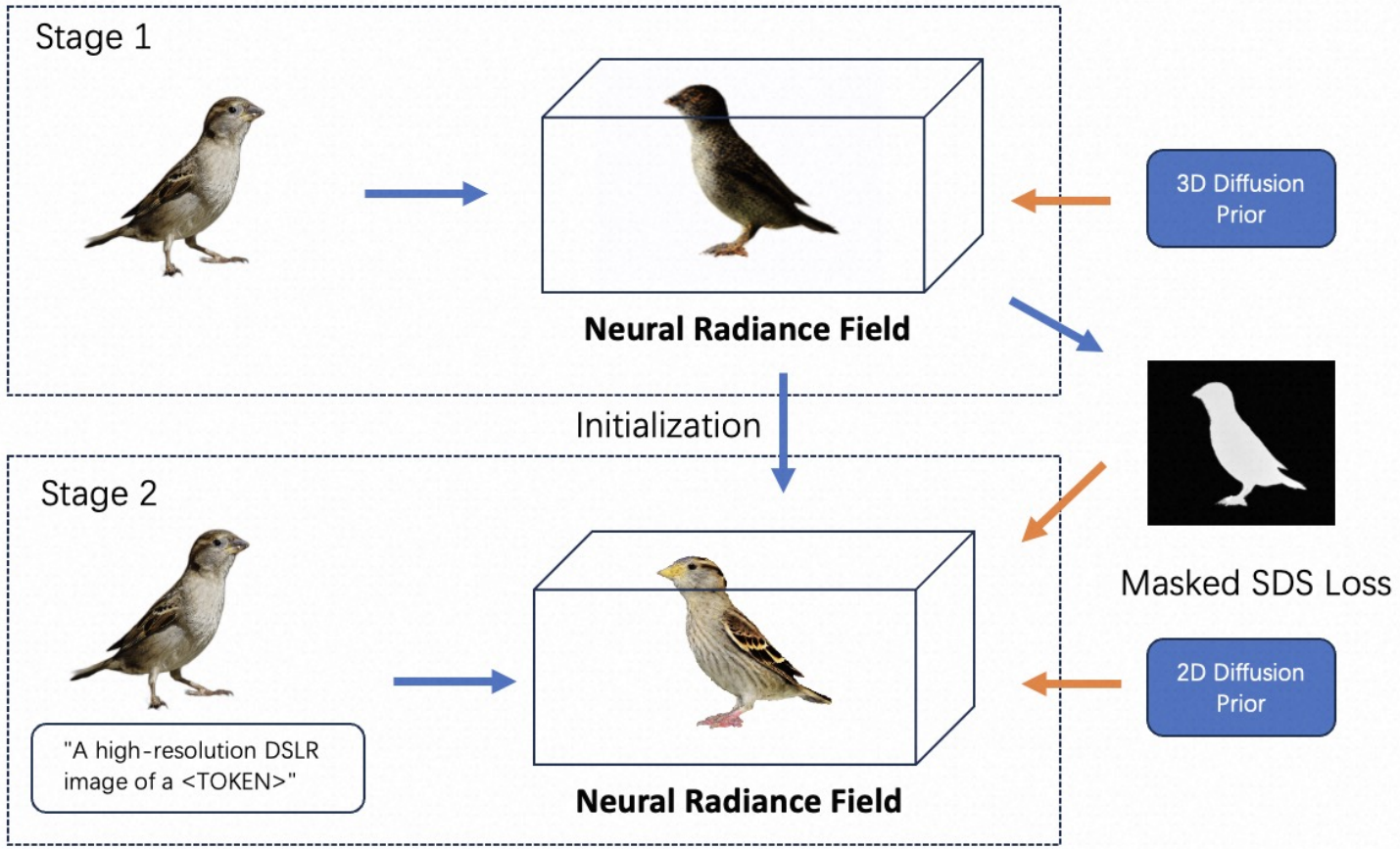
View-Conditioned Diffusion Priors

对于给定视角，能生成较好的形状，但在纹理等细节上质量差。

- Stage 2

Diffusion Inpainting Priors

基于Stage 1的NeRF模型生成Mask，重绘Mask部分。在保持Stage 1形状的同时，生成较好的纹理等细节。



MPOD123: One Image to 3D Content Generation Using Mask-enhanced Progressive Outline-to-Detail Optimization. CVPR 2024

基于反馈的协同：大模型反馈小模型

View-Conditioned Diffusion

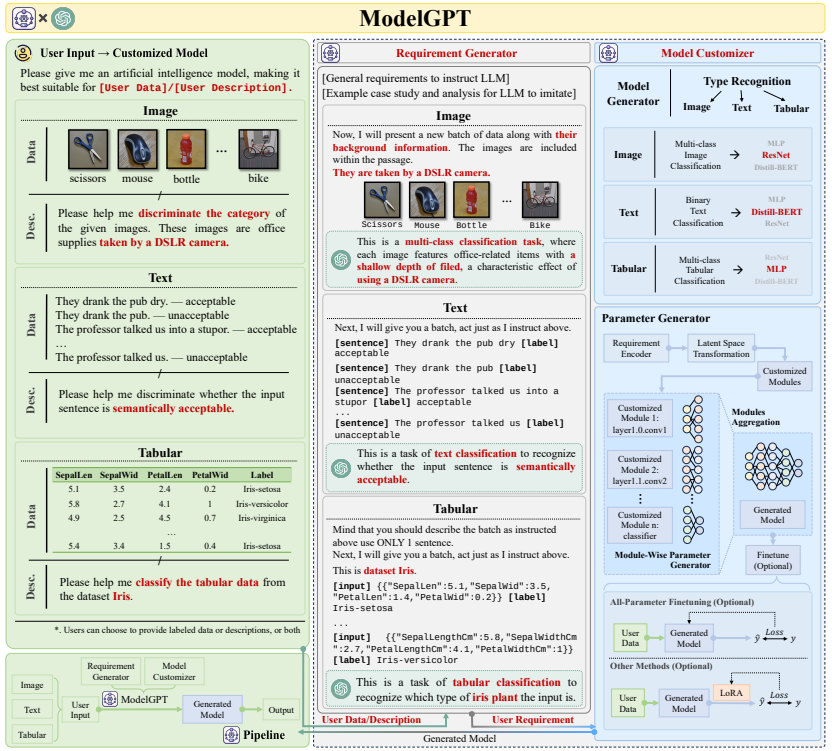
Text-Conditioned Diffusion

Our Trial

MPOD123: One Image to 3D Content Generation Using Mask-enhanced Progressive Outline-to-Detail Optimization. CVPR 2024

基于生成的协同：One（大模型） to All （小模型）生成

- **大模型驱动的小模型生成框架ModelGPT**
  - ModelGPT + 用户对**模型的需求描述** + **少量数据** =(推理生成) 开箱即用小模型。在All-in-One 的通用大模型范式之外，初步探索 **One-to-All** 的可能性，为更广泛的小数据、小算力（边端）、离线应用场景提供AI落地支撑。
  - 在**NLP，CV，和Tabular Data**典型数据集上进行验证，**性能超越Finetune**方法。



Algorithm 1 Pseudo-code of Parameter Generator $P(\cdot;\theta_p)$

Require: $A = \{(D_i = \{X_i, Y_i\}, r_i)\}_{i=1}^{N}$
Ensure: $\theta_p = (\theta_e, \theta_m, \theg_g)$ satisfies Equation (5)

```
i ← 1
for _ = 0 to ♯epoch do
    for (D_i, r_i) in A do
        for batch in D_i do
            Obtain θ_t with Equations (1) to (3)
            Use batch to compute the loss and update θ_t
            Compute the difference Δθ_t of θ_t
            Use Δθ_t to compute the gradients of θ_p
            Update θ_p
        end for
    end for
    Save best checkpoint according to Equation (5)
end for
```

Zihao Tang, Zheqi Lv, Shengyu Zhang, Fei Wu, Kun Kuang: ModelGPT: Unleashing LLM's Capabilities for Tailored Model Generation. CoRR abs/2402.12408 (2024)

基于生成的协同：One（大模型） to All （小模型）生成

- **大模型驱动的小模型生成框架ModelGPT**
  - 在**NLP，CV，和Tabular Data**典型数据集上进行验证，**性能超越Finetune**方法。
  - 给定用户的需求ModelGPT能够以至多先前范式（例如全参数微调、LORA微调）**270倍速度**快速生成定制好的人工智能模型。

Results on GLUE Benchmark (Distil-BERT)

Methods	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	DM	Score	♯Epoch	E2E Runtime
LoRA	**48.3**	91.0	84.9 / 80.3	81.2 / 80.0	68.9 / 87.3	80.5	33.1	**88.1**	52.8	**65.1**	0.0	71.5	20	216.1
Finetune	45.5	**91.3**	**86.6 / 80.8**	**82.1 / 80.9**	**69.2 / 87.8**	**81.8**	**80.8**	87.6	56.9	63.7	**35.6**	**74.4**	20	273.8
ModelGPT	39.5	88.9	85.3 / 78.4	80.9 / 80.3	63.3 / 83.5	77.8	78.0	84.6	69.5	64.4	28.0	73.4	**0**	**1.0**
ModelGPT-F	36.9	90.8	85.5 / 79.4	81.3 / 80.5	67.0 / 86.6	77.8	78.1	85.8	**70.0**	62.3	29.9	73.8	1	3.1

Results on Tabular Data (MLP)

Methods	Iris	Heart Disease	Wine	Adult	Breast Cancer	Car Evaluation	Wine Quality	Dry Bean	Rice	Bank Marketing	Average	♯Epoch	E2E Runtime
LoRA	93.3	**63.0**	67.3	54.7	95.9	71.3	55.0	88.9	92.5	89.8	77.2	20	46.2
Finetune	88.9	54.3	89.1	**55.2**	**96.5**	71.0	55.3	**90.6**	**93.1**	89.9	78.4	20	39.5
ModelGPT	**100.0**	60.9	**94.5**	54.7	95.3	**71.5**	54.1	85.0	92.5	89.8	79.8	**0**	**1.0**
ModelGPT-F	**100.0**	62.0	**94.5**	55.1	95.9	71.3	**55.4**	88.8	92.9	**90.0**	**80.6**	1	2.3

Results on Office-31 (ResNet-50)

Domain	Amazon			DSLR			Average			Webcam (ModelGPT is Zero-Shot)			♯Epoch	E2E Runtime
Methods	Acc	Acc@3	Acc@5	Acc	Acc@3	Acc@5	Acc	Acc@3	Acc@5	Acc	Acc@3	Acc@5		
LoRA	66.4	77.7	84.8	78.4	92.2	96.1	72.4	85.0	90.5	72.5	87.5	93.8	400	231.8
Finetune	67.5	79.2	83.7	84.3	98.0	**100.0**	75.9	88.6	91.9	**90.0**	**100.0**	**100.0**	400	257.6
ModelGPT	66.4	79.9	83.7	**92.2**	**100.0**	**100.0**	79.3	90.0	91.9	76.2	87.5	91.2	**0**	**1.0**
ModelGPT-F	**67.8**	**81.3**	**85.9**	**92.2**	**100.0**	**100.0**	**80.0**	**90.7**	**92.8**	77.5	90.0	91.3	1	1.2

基于生成的协同：同构大模型进化融合

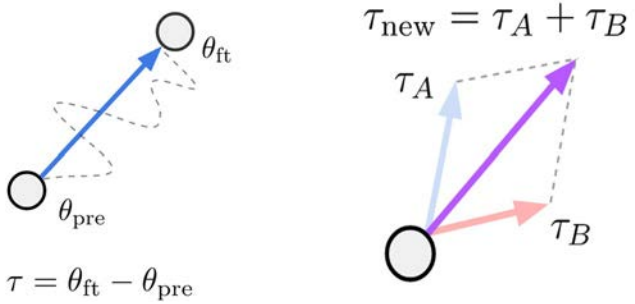
在以下数据集上进行进化搜索：
- GSM8K训练集的一个子集
- MMLU验证集的一个子集
- MBPP训练集

使用的源语言模型（LLMs）：
- DeepSeek-Math-RL
- DeepSeek-Coder-Ins-v1.5

方法：
- 算术任务
- TIES合并

进化方法：
- 协方差矩阵自适应进化策略（CMA-ES）

Task Arithmetic [1]



$$\theta_{\text{Merge}} = \theta_{\text{pre}} + (\lambda * \sum_{i=1}^{n} \tau_i)$$

TIES-Merging [2]

- Trim $\hat{\tau}_t = \hat{\gamma}_t \odot \hat{\mu}_t$
- Elect $\gamma_m^p = \text{sgn}(\sum_{t=1}^{n} \hat{\tau}_t^p)$
- Disjoint Merge

$$\tau_m^p = \frac{1}{|\mathcal{A}^p|} \sum_{t \in \mathcal{A}^p} \hat{\tau}_t^p$$
$$\mathcal{A}^p = \{t \in [n] \mid \hat{\gamma}_t^p = \gamma_m^p\}$$

Editing models with task arithmetic.
Ties-merging: Resolving interference when merging models.
Unconstrained Model Merging for Enhanced LLM Reasoning." arXiv preprint arXiv:2410.13699 (2024).

基于生成的协同：异构模型蒸馏融合

源 LLMs:

- Qwen2.5-Math-7B-Ins
- MetaMath-7B, MetaMath-70B
- OpenMath-Mistral-7B
- WizardMath-7B-V1.1
- CodeLlama-7B-Ins, CodeLlama-70B-In
- DeepSeek-Coder-Ins-v1.5 (7B)

结合不同结构和规模的模型能力

使目标LLM的输出分布Q与融合后的分布P对齐。

$$\mathcal{L}_{\text{Fusion}} = -\mathbb{E}_{(I_i,R_i)\sim\mathcal{D}}\left[\mathbb{H}\left(\mathbf{P}_{i,j}\|\mathbf{Q}_{i,\phi_j}\right)\right]$$

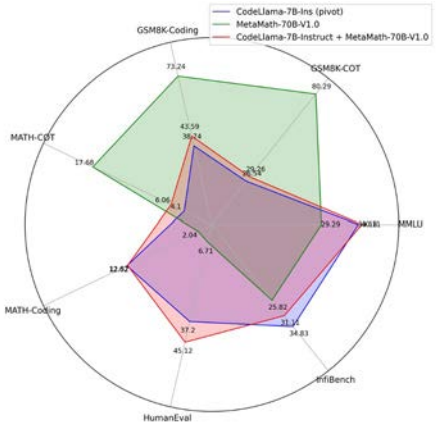
损失函数

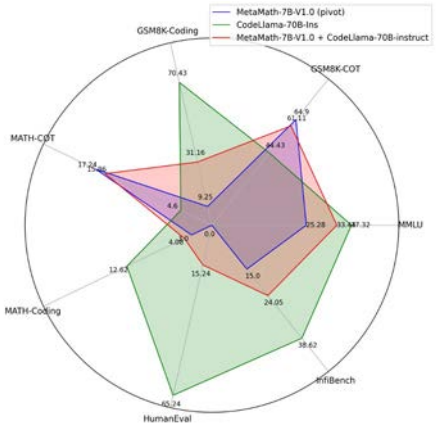
$$\mathcal{L} = \lambda\mathcal{L}_{\text{SFT}} + (1-\lambda)\mathcal{L}_{\text{Fusion}}$$

Unconstrained Model Merging for Enhanced LLM Reasoning." arXiv preprint arXiv:2410.13699 (2024).

基于生成的协同：异构模型蒸馏融合

Model		MMLU	GSM8K-COT	GSM8K-Coding	MATH-COT	MATH-Coding	HumanEval	InfiBench
Source Model	**Base Model/#Size**				**Source LLMs**			
Qwen2.5-Math-7B-Ins	Qwen2.5/7B	56.31	88.70 (95.2)	87.9	75.26 (83.6)	32.46	48.17	17.36
MetaMath-7B	Llama 2/7B	25.28	64.90 (66.5)	9.25	17.24 (19.8)	3.00	0.0	15.00
MetaMath-70B	Llama 2/70B	29.49	80.29 (82.3)	73.24	17.68 (26.6)	2.04	6.71	25.82
OpenMath-Mistral-7B	Mistral/7B	28.23	44.73	77.33 (80.2)	12.38	27.68 (44.5)	0.0	17.69
WizardMath-7B-V1.1	Mistral/7B	27.66	66.03	74.45	18.08	12.38	15.85	38.47
DeepSeek-Math-RL	Deepseek-LLM/7B	25.05	88.17 (88.2)	83.24	48.46 (51.7)	41.68	45.73	32.16
CodeLlama-7B-Ins	Llama 2/7B	39.18	26.54	38.74	4.1	12.62	37.2 (34.8)	34.83 (35.15)
CodeLlama-70B-Ins	Llama 2/70B	37.32	44.43	-	4.6	-	65.24 (67.8)	38.62 (42.82)
DeepSeek-Coder-Ins-v1.5	Deepseek-LLM/7B	49.78 (49.5)	56.33	73.31 (72.6)	12.28	29.12 (34.1)	68.90 (64.1)	56.67
Pivot Model	**Source Model**				**Fusion: Heterogeneous LLMs**			
OpenMath-Mistral 7B	CodeLlama-70B-Ins	53.32↑↑	71.49↑↑	80.13	20.22↑↑	24.02	47.56↑↑	42.21↑↑
WizardMath-7B-V1.1	CodeLlama-70B-Ins	46.74↑↑	76.65↑↑	74.90	24.04↑↑	15.46	53.66↑↑	40.29↑
Qwen2.5-Math-7B-Ins	CodeLlama-70B-Ins	55.86↑↑	83.62↑↑	71.19	56.42↑↑	26.80	50.00↑	17.73-
CodeLlama-7B-Ins	MetaMath-70B	40.11↑↑	29.26↑	43.59↑	6.06↑	12.52↑↑	45.12↑↑	31.11↑↑
CodeLlama-7B-Ins	OpenMath-Mistral 7B	36.98↑↑	29.34↑	42.38↑	6.26↑	13.68↑	45.73↑↑	29.87↑↑
CodeLlama-7B-Ins	Qwen2.5-Math-7B-Ins	38.28-	29.64↑	42.07↑	6.5↑	16.06↑	46.34↑↑	32.36↑↑
MetaMath-7B	CodeLlama-70B-Ins	33.44↑↑	61.11↑↑	31.16↑↑	15.86↑↑	1.48↓	15.24↑↑	24.05 ↑↑
OpenMath-Mistral 7B	Deepseek-Coder-Ins-v1.5	53.38↑↑	72.33↑↑	80.36↑↑	20.66↑↑	21.98↓	48.17↑↑	40.96↑↑
CodeLlama-7B-Ins	WizardMath-7B-V1.1	37.98↑↑	28.28↑	42.91↑	6.32↑	13.86↑	41.46↑↑	31.51↓





融合后支点模型能力大幅提升！

- 合并数学和编码模型能够提升数学和编码能力，甚至超越原始编码模型的水平。这可能表明，通过大语言模型（LLM）的合并，获得了**组合能力**——即带有数学思维的编码能力，而不仅仅是个别技能的叠加。

- 当选择以优化复杂任务（例如数学优于编码）的LLM作为支点模型时，合并后的结果表现更为优异。特别是，如果枢纽模型是预训练模型，而与其合并的模型是Chat模型，那么在初始预训练后的微调过程对实现最优集成具有很高的重要性

Instruction tuning for large language models: A survey

Instruction Tuning for Large Language Models: A Survey

SHENGYU ZHANG, Zhejiang University, China
LINFENG DONG, Zhejiang University, China
XIAOYA LI, Shannon.AI, China
SEN ZHANG, Zhejiang University, China
XIAOFEI SUN, Zhejiang University, China
SHUHE WANG, Peking University, China
JIWEI LI, Zhejiang University, China
RUNYI HU, Zhejiang University, China
TIANWEI ZHANG, Nanyang Technological University, Singapore
FEI WU, Zhejiang University, China
GUOYIN WANG, Amazon, United States

This paper surveys research works in the quickly advancing field of instruction tuning (IT), a crucial technique to enhance the capabilities and controllability of large language models (LLMs). Instruction tuning refers to the process of further training LLMs on a dataset consisting of (INSTRUCTION, OUTPUT) pairs in a supervised fashion, which bridges the gap between the next-word prediction objective of LLMs and the users' objective of having LLMs adhere to human instructions. In this work, we make a systematic review of the literature, including the general methodology of IT, the construction of IT datasets, the training of IT models, and applications to different modalities, domains and application, along with analysis on aspects that influence the outcome of IT (e.g., generation of instruction outputs, size of the instruction dataset, etc). We also review the potential pitfalls of IT along with criticism against it, along with efforts pointing out current deficiencies of existing strategies and suggest some avenues for fruitful research.

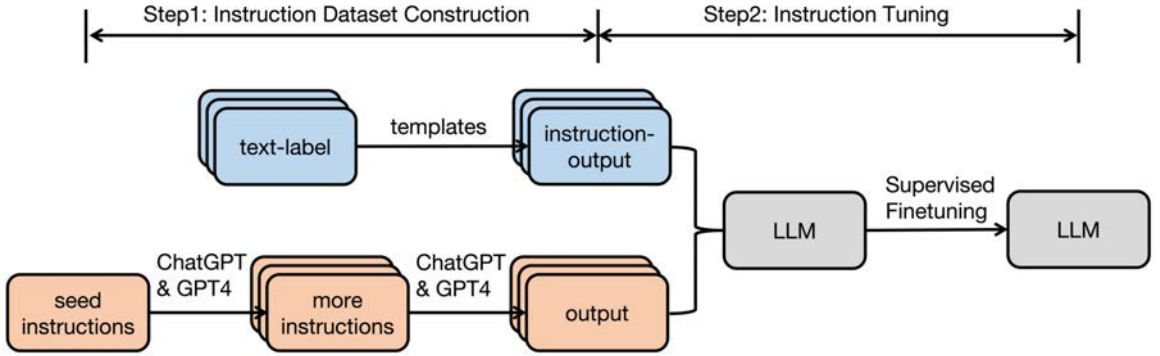
CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Neural networks**.

Additional Key Words and Phrases: Large Language Model, Instruction Tuning, Survey

ACM Reference Format:
Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2018. Instruction Tuning for Large Language Models: A Survey. In . ACM, New York, NY, USA, 38 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The field of large language models (LLMs) has witnessed remarkable progress in recent years. LLMs such as GPT-3 [18], PaLM [25], and LLaMA [130] have demonstrated impressive capabilities across a wide range of natural language tasks [2, 22, 45, 47, 70, 73, 87, 96, 103, 120–122, 132, 134, 135, 143, 144, 153–155, 165]. One of the major issues with LLMs is the mismatch between the training objective and users' objective: LLMs are typically trained on minimizing the



Shengyu Zhang, Linfeng Dong, Xiaoya Li, et al.
Under Review
(online, https://arxiv.org/pdf/2308.10792.pdf)

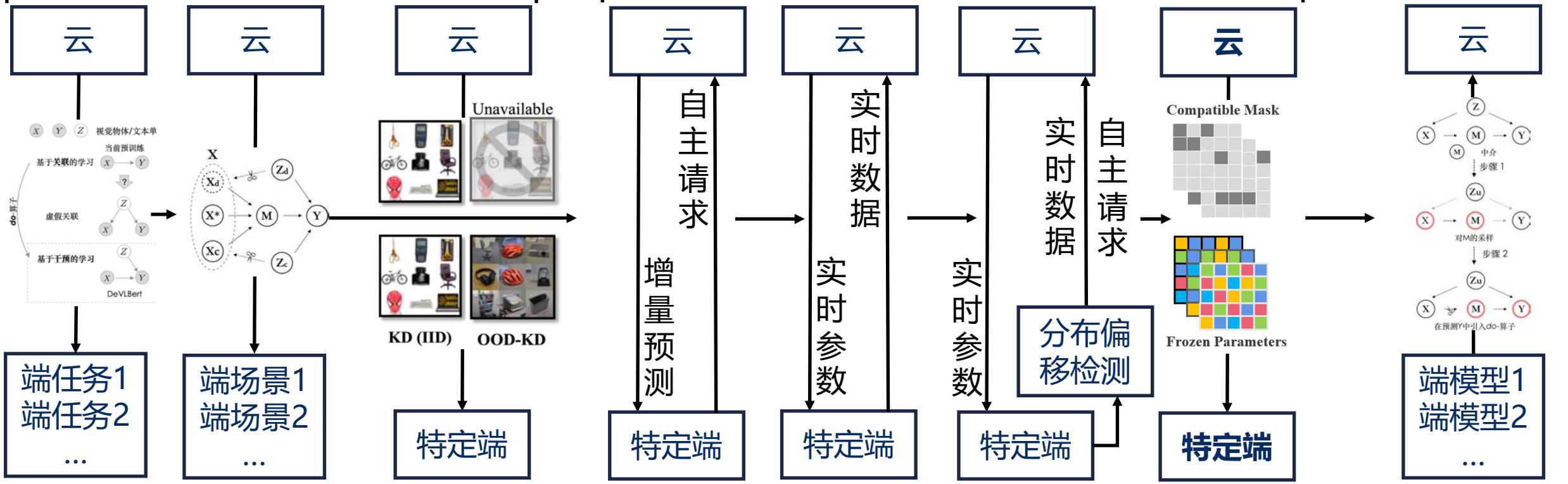
PART 03

大小模型端云协同智能

端云异构模型知识互迁与协同推断

Cloud to Device (C2D) → Cloud for Device (C4D) → Device to Cloud (D2C)

大规模因果预训练



DeVLBert
跨任务泛化
ACM MM 20

DeVADG
跨场景泛化
AAAI 23

AUG-KD
迁移压缩
ICLR 24

AdaRequest
自主请求
KDD 22

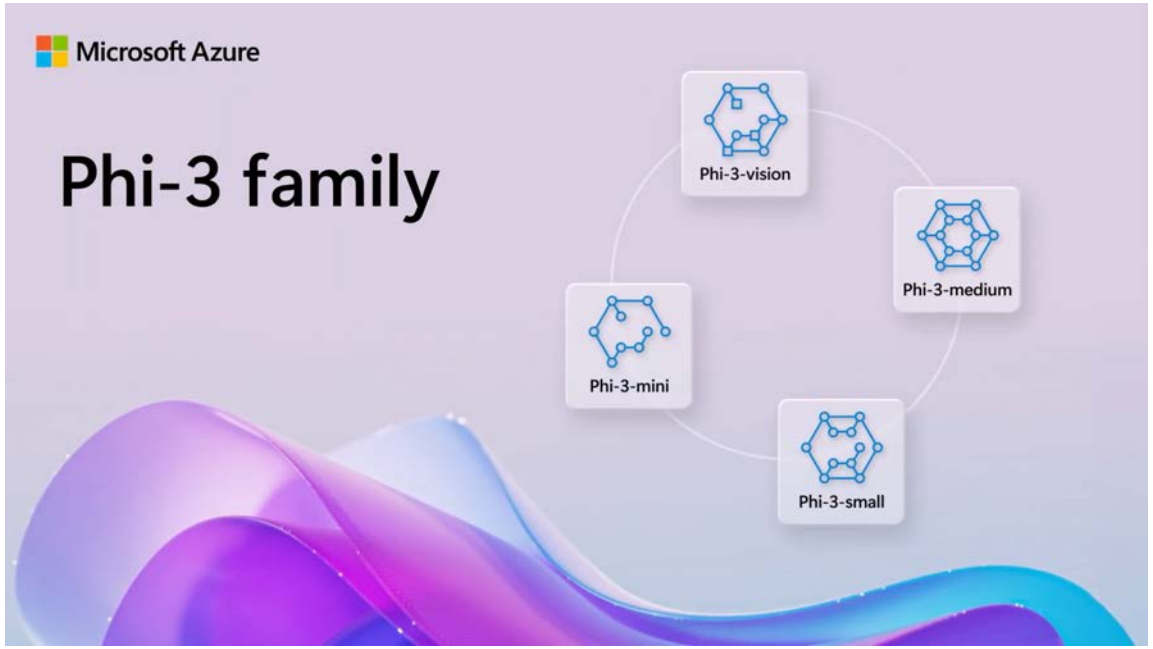
DUET
实时适应
WWW 23

IntellectReq
实时自主适应
WWW 24

DIET
高效定制
KDD 24

AdaRequest
因果去偏汇聚
TKDE 23

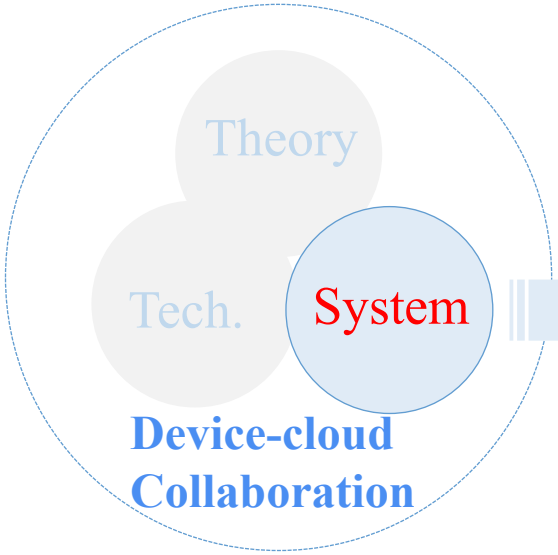
国内外现状 – 国外产业应用



微软Phi-3

Apple Intelligence

- 目前Phi-1已经进化到只有3.8亿参数的phi-3-mini，经过3.3万亿tokens的训练，其性能可以与Mixtral 8x7B和GPT-3.5相媲美，部署在手机上。
- Apple Intelligence：将生成式模型引入 iPhone、iPad 和 Mac 核心的个人智慧系统



Real-time

Perceptron Capacity

Efficiency

Research models have been deployed in Mobile Taobao, Huawei Music, and Tencent News

Huawei Music

Playlist Recommendation

Self-supervised music representation learning

From 2022.4

Average-Play-Rate/Time +5.1/5.8%

WWW 2022

Taobao, **Guess You Like**

Product Recommendation

Edge-cloud Collaborative Recommendation

From 2021.10

GMV +3% in **Double 11**, 2021

KDD 2022

Tencent News

Micro-video Recommendation

Cross-channel Recommendation

Average-Play-Rate +25%

CIKM 2022

Taobao, Single-image Reconstruction

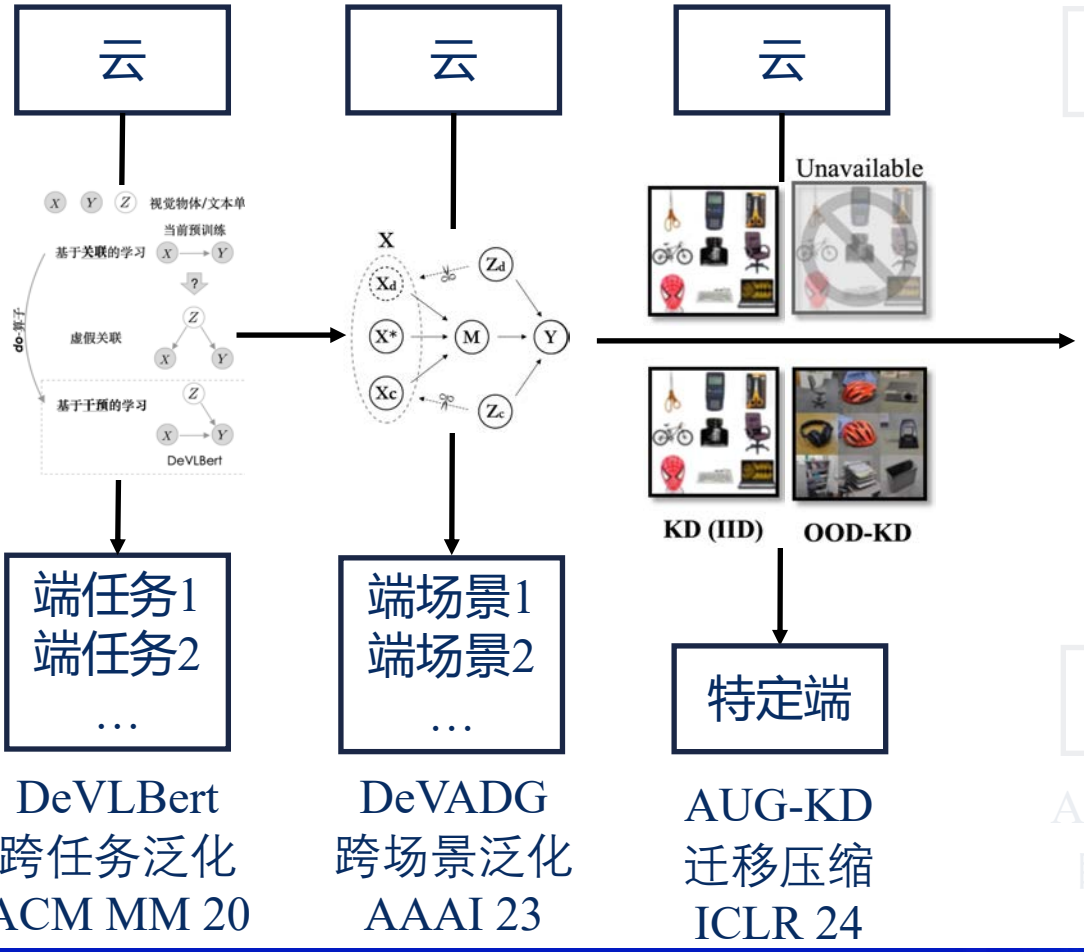
Audio-driven Talking face Generation

CVPR 2024

端云异构模型知识互迁与协同推断

Cloud to Device
(C2D)

大规模因果预训练



DeVLBert
跨任务泛化
ACM MM 20

DeVADG
跨场景泛化
AAAI 23

AUG-KD
迁移压缩
ICLR 24

不同端设备存在**差异化任务功能**和**差异化使用场景**，云模型向端侧迁移部署面临着**跨场景、跨任务的泛化性**问题

面向未知端侧分布的压缩-适应联合

研究背景

- 大模型向端侧迁移部署往往采用知识蒸馏等压缩手段，传统知识整理方法假设大模型训练数据分布（压缩前）和小模型测试数据分布（压缩后）服从独立同分布假设（IID Hypothesis）。
- 实际应用中，源域数据和应用场景存在**分布偏移**，**导致压缩性能显著下降。**

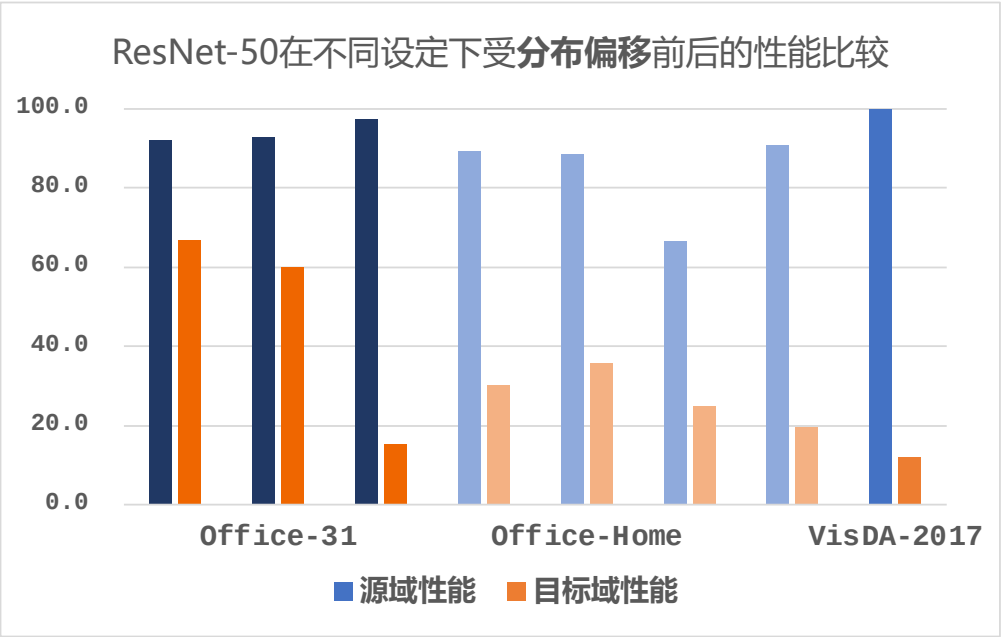
理论分析

独立同分布假设（IID Hypothesis）：**源域**P_s和**目标域**P_t（应用场景）独立同分布。在此情况下进行知识蒸馏，源域的知识可以很好地指导模型完成目标域的任务。

- 数据蒸馏的目标：

$$\min_{\theta_s} \mathbb{E}_{(x,y)\sim P}[D_{KL}(T(x;\theta_t) \parallel S(x;\theta_s) + CE(S(x;\theta_s), y)].$$

- 多数场景下，源域分布和应用场景存在**分布偏移**（$P_t \neq P_s$），违反独立同分布假设。
  - 情况1：$P \approx P_t$，对应无数据蒸馏方法（P_t 由生成器拟合），蒸馏出的目标模型并不适用P_s。
  - 情况2：$P \approx P_s$，源模型给出的知识不一定有效。



Zihao Tang, Zheqi Lv, Shengyu Zhang, Yifan Zhou, Xinyu Duan, Fei Wu, Kun Kuang: AuG-KD: Anchor-Based Mixup Generation for Out-of Domain Knowledge Distillation. ICLR 2024

面向未知端侧分布的压缩-适应联合优化

动机

- Main Idea：**利用目标域数据提取源模型知识**

挑战

- **选择性抽取源模型的知识?**
  - 源域和目标域的联合分布$P(X,Y)$存在差异。
  - 源模型（面向源域数据优化）在处理目标域数据时，可能会**给出错误的预测**，或者**无法反映出目标域各个类别间的关联**，进而误导目标模型。
- **源域数据的缺失**
  - 无数据知识蒸馏不稳定，难收敛。
  - 目标域数据中包含**域特定信息（Domain-Specific Information）**，可能有助于目标模型的学习。

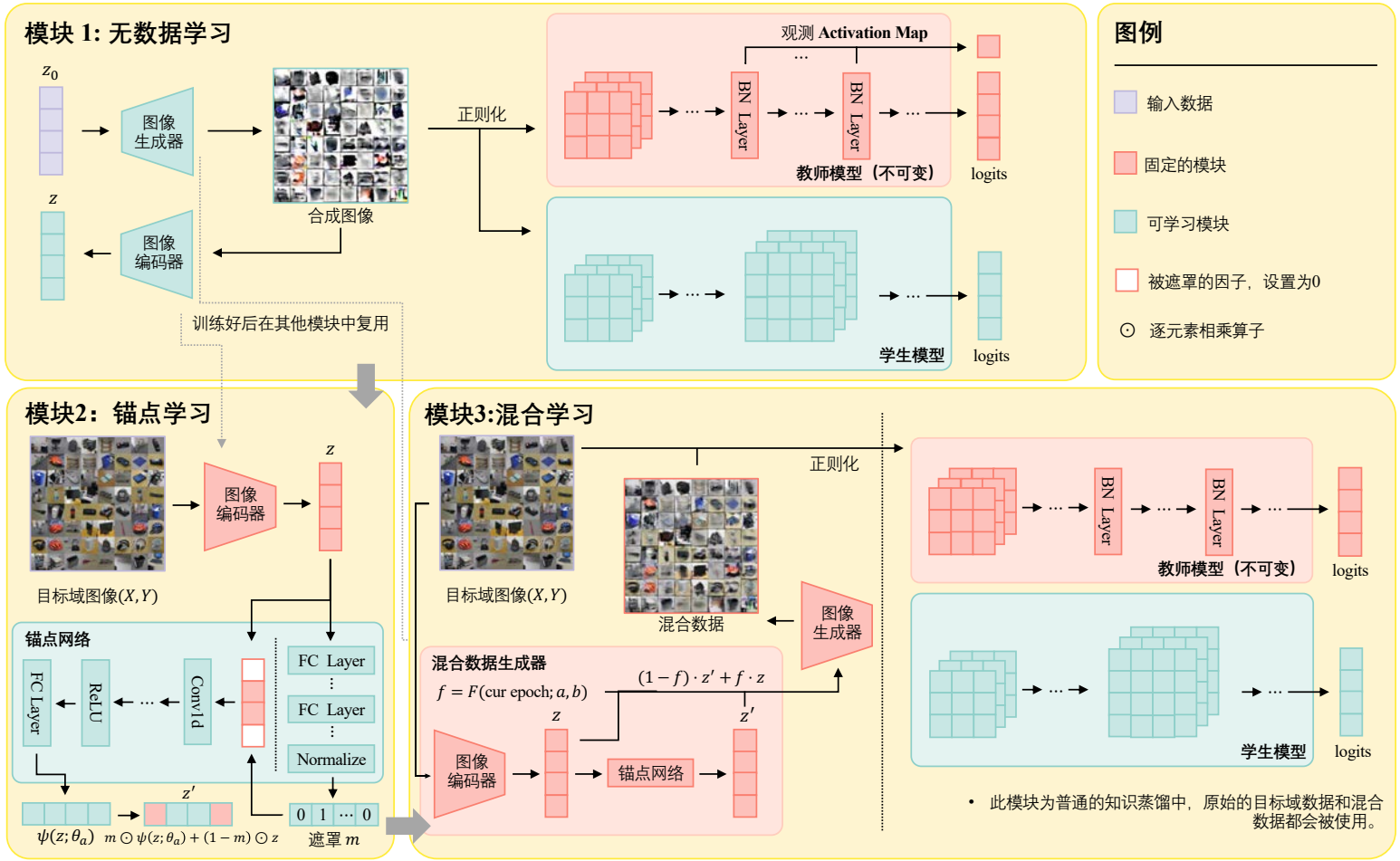
三种知识蒸馏设定的比较

Zihao Tang, Zheqi Lv, Shengyu Zhang, Yifan Zhou, Xinyu Duan, Fei Wu, Kun Kuang: AuG-KD: Anchor-Based Mixup Generation for Out-of Domain Knowledge Distillation. ICLR 2024

面向未知端侧分布的压缩-适应联合优化

算法模型 —— 整体架构

- 利用**不确定性驱动（Uncertainty-Driven）**且**样本特定（Sample-Specific）**的锚点，将目标域数据与源域对齐，使用生成式的方法逐渐将目标模型的学习过程从针对跨域知识的蒸馏转变为针对域特定信息的学习。
- 整体方法由**无数据学习**、**锚点学习**、**混合学习**3个模块组成。
  - **无数据学习模块**解决没有源域数据的问题。
  - **锚点学习模块**使得源模型可以传递正确的知识。
  - **混合学习模块**平衡针对跨域知识的蒸馏转与针对域特定信息的学习。



Zihao Tang, Zheqi Lv, Shengyu Zhang, Yifan Zhou, Xinyu Duan, Fei Wu, Kun Kuang: AuG-KD: Anchor-Based Mixup Generation for Out-of Domain Knowledge Distillation. ICLR 2024

面向未知端侧分布的压缩-适应联合优化

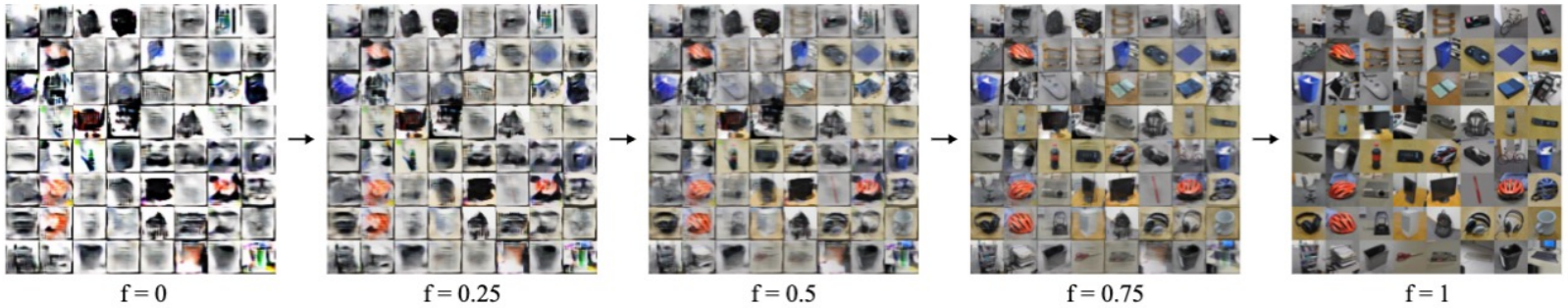
- 此模块为普通的知识蒸馏过程，**但是用于蒸馏的数据（混合数据和原始源域数据）会随训练进程改变。**训练初期混合数据会更接近源域数据，便于跨域知识的蒸馏；训练后期则会更像目标域数据，便于域的特定信息的学习。
- 混合数据的变化由阶段因子$f \in [0,1]$控制，它由一个单调不减的计划函数$F(\cdot; a, b): \mathbb{N} \mapsto [0,1]$通过当前的训练的epoch数决定，并满足如下性质：
  - $F(a \cdot \sharp\ epoch; a, b) = 1$
  - $F(0; a, b) = b$

算法模型 —— 混合学习

- 混合数据由下式，通过图像生成器生成：

$$x_m = (1-f) \cdot G\big((1-f) \cdot z' + f \cdot z\big) + f \cdot x$$

- 下图为不同阶段因子时的混合数据：



源域数据

目标域数据

Zihao Tang, Zheqi Lv, Shengyu Zhang, Yifan Zhou, Xinyu Duan, Fei Wu, Kun Kuang: AuG-KD: Anchor-Based Mixup Generation for Out-of Domain Knowledge Distillation. ICLR 2024

面向未知端侧分布的压缩-适应联合优化

Office-31: resnet34 → mobilenet_v3_small

Settings	Amazon, Webcam→DSLR			Amazon, DSLR→Webcam			DSLR, Webcam→Amazon		
	Acc	Acc@3	Acc@5	Acc	Acc@3	Acc@5	Acc	Acc@3	Acc@5
Teacher	92.2	96.1	97.0	93.1	96.2	97.3	97.7	99.6	99.8
	67.1	82.6	88.0	60.0	77.5	82.5	15.2	26.1	36.0
DFQ+	80.4±5.7	93.3±4.1	96.4±2.1	86.5±5.7	**97.5±2.0**	99.0±1.0	46.6±4.5	67.6±2.4	76.5±2.9
CMI+	67.1±3.5	86.6±4.3	92.9±3.0	70.0±5.3	88.0±5.1	94.3±2.1	35.9±2.3	56.1±5.1	65.0±5.6
DeepInv+	65.9±6.3	84.7±4.9	90.6±3.8	70.0±5.4	91.5±0.5	94.8±1.6	36.5±4.4	56.1±5.1	66.3±3.3
w/o KD	63.5±7.9	84.7±4.5	90.2±3.7	82.7±5.4	96.0±1.9	98.3±0.7	52.9±3.4	72.5±3.6	**79.9±2.2**
ZSKT+	33.3±5.9	55.3±11.8	65.9±11.5	33.0±8.1	55.3±14.3	66.8±16.2	23.7±5.3	42.7±7.1	53.7±5.9
PRE-DFKD+	68.3±19.5	87.8±14.3	91.8±13.3	66.5±20.9	82.0±17.3	88.9±12.9	28.4±13.3	46.4±19.0	55.9±20.8
Ours	**84.3±3.1**	**94.9±2.6**	**97.6±0.8**	**87.8±7.6**	96.3±1.8	**99.5±0.7**	**58.8±3.7**	**73.7±2.1**	79.7±1.5

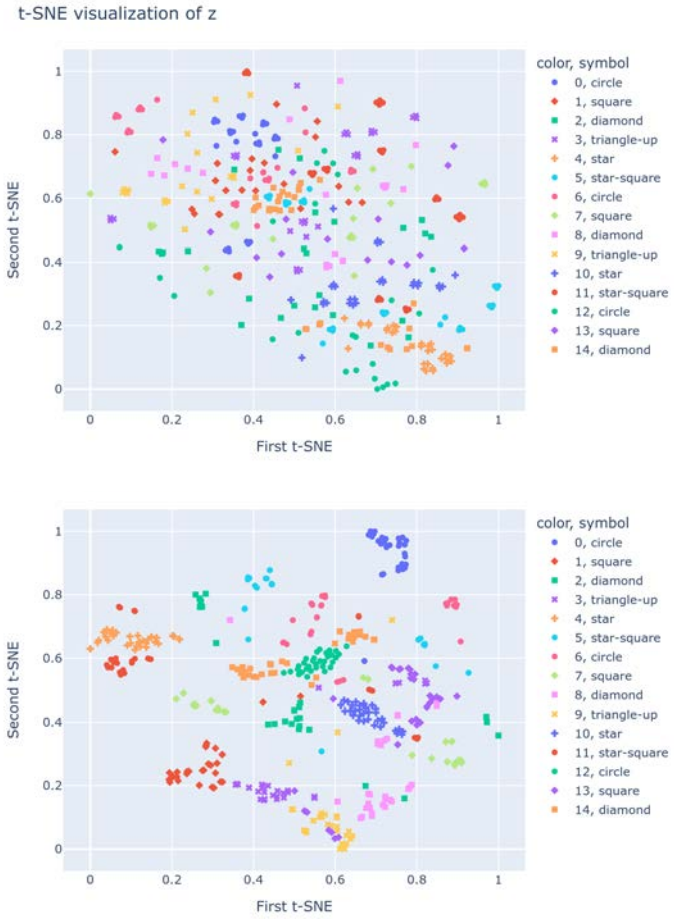
Office-Home: resnet34 → mobilenet_v3_small

Settings	ACP→R			ACR→P			APR→C			CPR→A		
	Acc	Acc@3	Acc@5	Acc	Acc@3	Acc@5	Acc	Acc@3	Acc@5	Acc	Acc@3	Acc@5
Teacher	89.4	93.2	94.5	88.6	92.5	93.8	66.7	75.8	79.6	90.9	94.4	95.4
	30.3	46.7	55.2	35.8	53.4	60.8	24.7	40.6	48.7	19.6	31.2	39.1
DFQ+	33.3±1.3	51.7±1.4	60.7±1.7	60.0±3.8	75.8±2.9	81.8±2.6	50.6±2.8	67.7±2.8	75.2±1.6	21.0±3.4	31.8±3.5	40.3±2.5
CMI+	16.4±1.2	29.0±0.4	37.0±0.7	48.8±1.5	63.9±1.4	70.3±1.6	35.3±1.9	51.2±2.0	58.4±1.7	13.4±3.0	21.4±2.7	27.5±2.8
DeepInv+	15.4±1.7	28.6±1.8	36.4±2.1	47.8±2.1	62.9±2.2	70.7±2.2	36.9±2.5	52.5±3.4	60.5±2.9	13.0±2.1	22.3±2.4	27.5±2.1
w/o KD	32.5±3.8	48.4±3.8	57.6±3.3	59.9±2.0	77.2±1.4	82.6±0.8	49.9±1.4	67.0±1.6	73.6±1.1	16.8±2.1	28.9±1.5	36.4±2.3
ZSKT+	15.5±3.3	29.7±4.4	38.5±4.5	11.9±5.6	23.5±9.5	32.0±10.9	7.8±2.9	19.5±5.3	27.5±6.7	7.9±3.1	17.6±3.6	26.7±3.0
PRE-DFKD+	22.3±3.7	36.9±5.0	44.9±5.2	34.4±9.5	52.4±11.2	60.5±11.0	38.4±7.9	57.9±11.2	65.4±11.1	9.0±2.7	20.4±3.7	27.5±5.0
Ours	**35.2±2.5**	**53.4±2.0**	**62.8±1.8**	**65.3±1.6**	**79.3±1.4**	**84.1±2.0**	**53.4±3.0**	**70.3±1.4**	**76.6±1.4**	**21.2±4.7**	**33.4±3.8**	**41.7±4.6**

VisDA-2017 (train→validation): resnet34 → mobilenet_v3_small

Settings	Teacher		DFQ+	CMI+	DeepInv+	w/o KD	ZSKT+	PRE-DFKD+	Ours
Acc	100.0	12.1	53.4±1.0	49.5±1.3	47.6±0.9	50.7±0.9	48.4±3.5	54.9±1.0	**55.5±0.3**
Acc@3	100.0	34.5	80.2±0.6	77.2±1.1	75.5±0.7	78.7±0.7	77.5±2.6	81.4±1.0	**82.1±0.2**
Acc@5	100.0	54.7	89.0±0.4	88.1±0.7	87.3±0.6	89.3±0.4	88.9±1.2	90.6±0.6	**91.3±0.1**

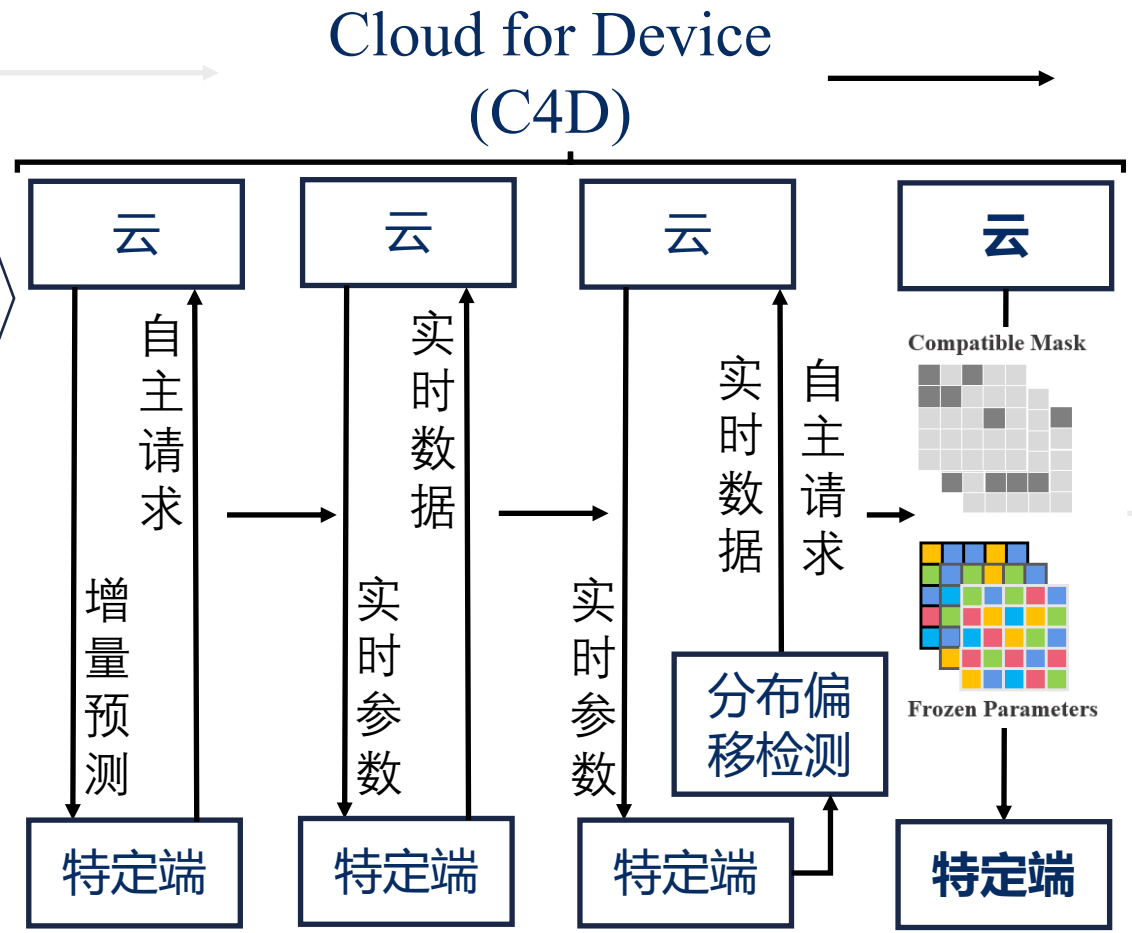
对遮罩前后的隐变量进行t-SNE分析结果



端云异构模型知识互迁与协同推断

解决端云大小模型在差异化**尺寸架构**和**优化目标**下的**协同推断**问题

云不直接执行任务本身，而是帮助端更好的执行既定任务



AdaRequest
自主请求
KDD 22

DUET
实时适应
WWW 23

IntellectReq
实时自主适应
WWW 24

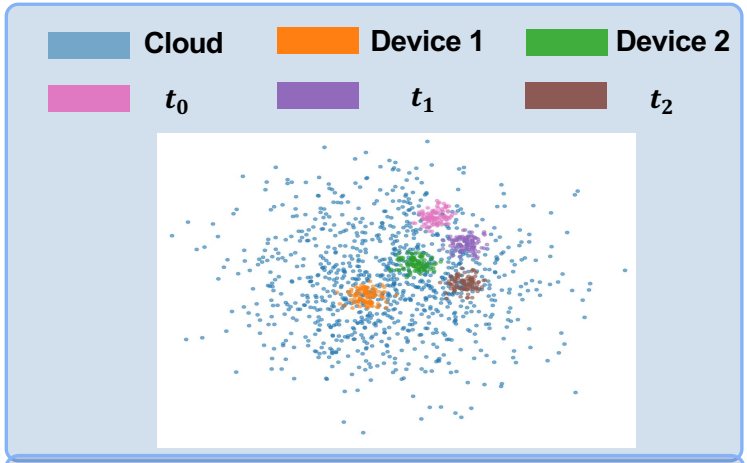
DIET
高效定制
KDD 24

基于端云协同的高效端模型定制

研究背景 现有端侧部署方案采用云侧大规模预训练，通过模型压缩后传输至端侧进行部署。然而多阶段训练、稠密信息传输给端侧动态复杂环境下的**高响应、低成本**自适应带来了巨大挑战

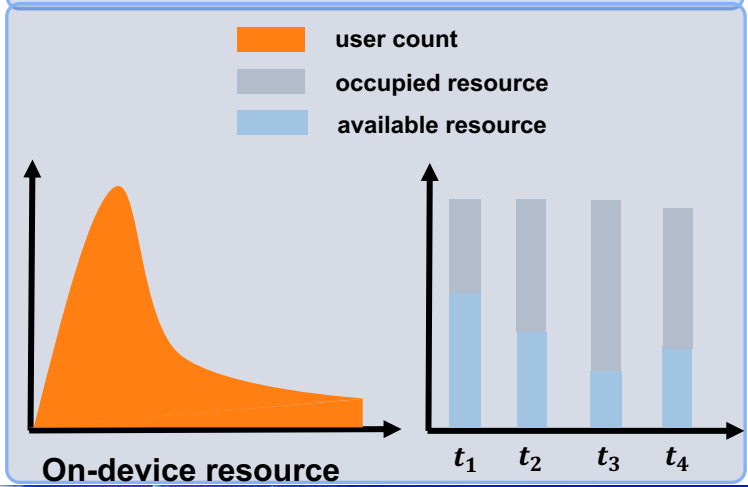
分布异质性

- **端云分布异质**：云侧全局数据分布体现平台整体共性与端侧特化分布存在偏移
- **端侧分布迁移**：端侧用户兴趣意图动态偏移，需要由云向端及时下发适配模型



资源异质性

- **端侧计算资源有限**：大量长尾用户移动设备算力有限，难以支撑本地训练微调
- **端云通信资源有限**：频繁下发稠密适配模型消耗大量通信带宽资源，降低响应



Fu K, Zhang S, Lv Z, et al. DIET: Customized Slimming for Incompatible Networks in Sequential Recommendation. KDD 2024 Research Track

基于端云协同的高效端模型定制

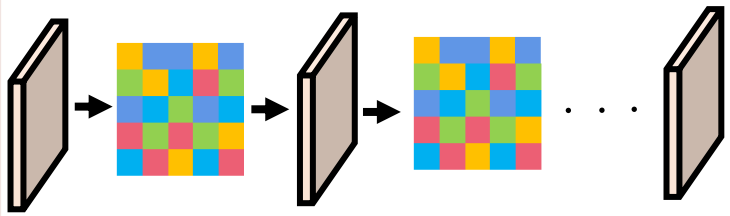
研究问题　研究基于端云协同的低通信开销、高响应速度端模型定制算法。

创新方法

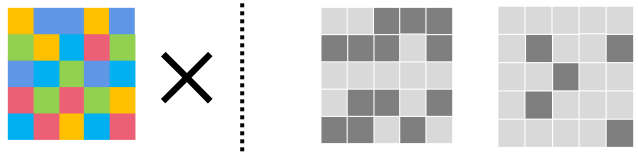
高效模型表示构建：基于神经网络彩票假说，将云向端训练压缩过程转化为传输适配子网二进制掩膜

高效适配子网搜索：云侧学习建立实时数据到端侧个性子网掩膜的映射，仅需前向推理即可高效响应

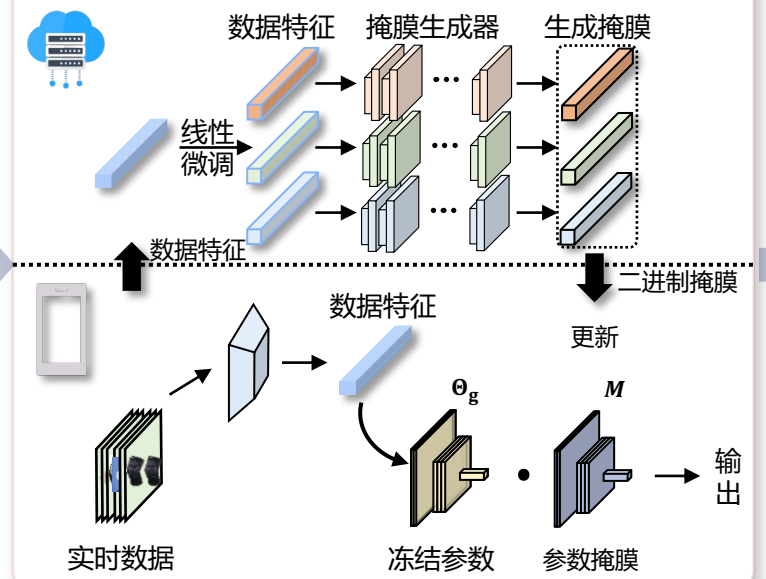
彩票假说理论



利用掩膜进行选择（一层参数多掩膜）

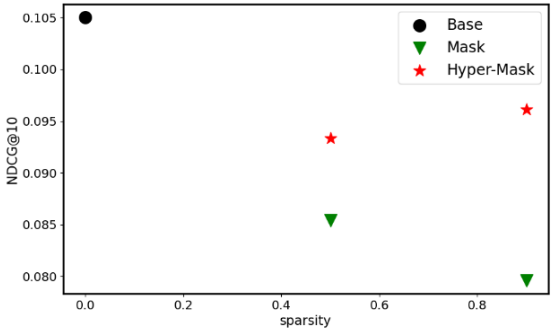


端云子网搜索



模型效率提升

优势 \ 方法	Base	Ours
低传输延迟	✗	✓
低存储成本	✗	✓
低推理时延	✗	✓



低时延低成本下得到相似的表现

Fu K, Zhang S, Lv Z, et al. DIET: Customized Slimming for Incompatible Networks in Sequential Recommendation. KDD 2024 Research Track

基于端云协同的高效端模型定制

应用验证

当前推荐系统存在的问题

通信开销大
云端分布差异大
端侧兴趣变化快
设备计算资源有限

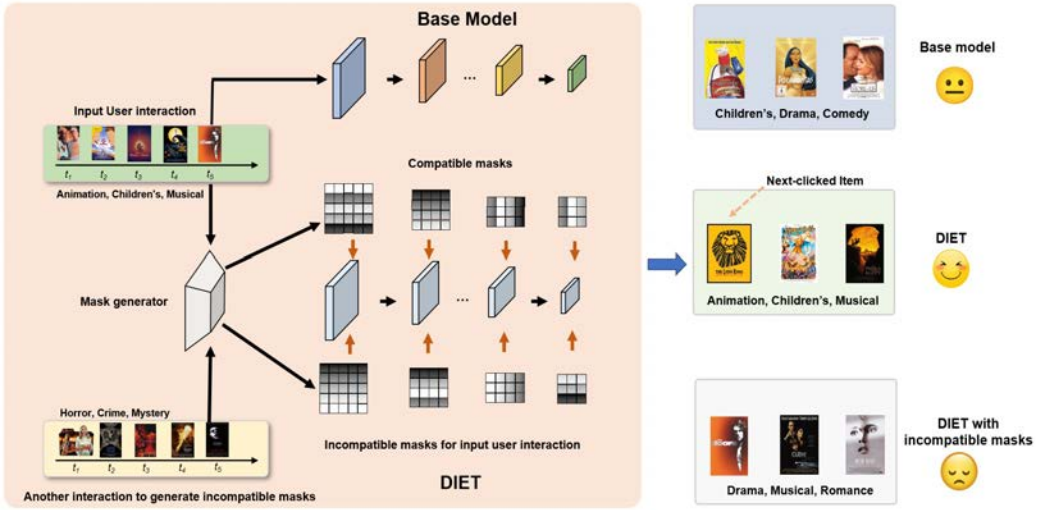
瘦身子网模型压缩
端侧实时兴趣提取
适配子网生成传输
…

端侧个性子网搜索

共性-个性协同
大-小模型协同
…

降低模型由云向端下发的传输开销至**原始大小的3%**
端侧模型能力提升的同时**推理速度提升5倍**

Model	Method	Dataset							
		Amazon-CD				MovieLens-100k			
		NDCG↑	Hit↑	Params↓	FLOPs↓	NDCG↑	Hit↑	Params↓	FLOPs↓
SASRec	Base	0.0386	0.0529	1.3107	0.2086	0.0517	0.1077	1.3107	0.2086
	DIET	**0.0425**	**0.0590**	**0.0410**	**0.1154**	**0.0635**	**0.1319**	**0.0410**	**0.0416**
	Improv. ↑	10.96%	11.53%	× 31.97	× 1.81	22.82%	22.47%	× 31.97	× 5.01
Caser	Base	0.0310	0.0424	0.4922	0.0586	0.0569	0.0719	0.4922	0.0586
	DIET	**0.0356**	**0.0488**	**0.0154**	**0.0294**	**0.0617**	**0.0771**	**0.0154**	**0.0488**
	Improv. ↑	14.84%	15.09%	× 31.96	× 1.99	10.42%	7.32%	× 31.96	× 1.76

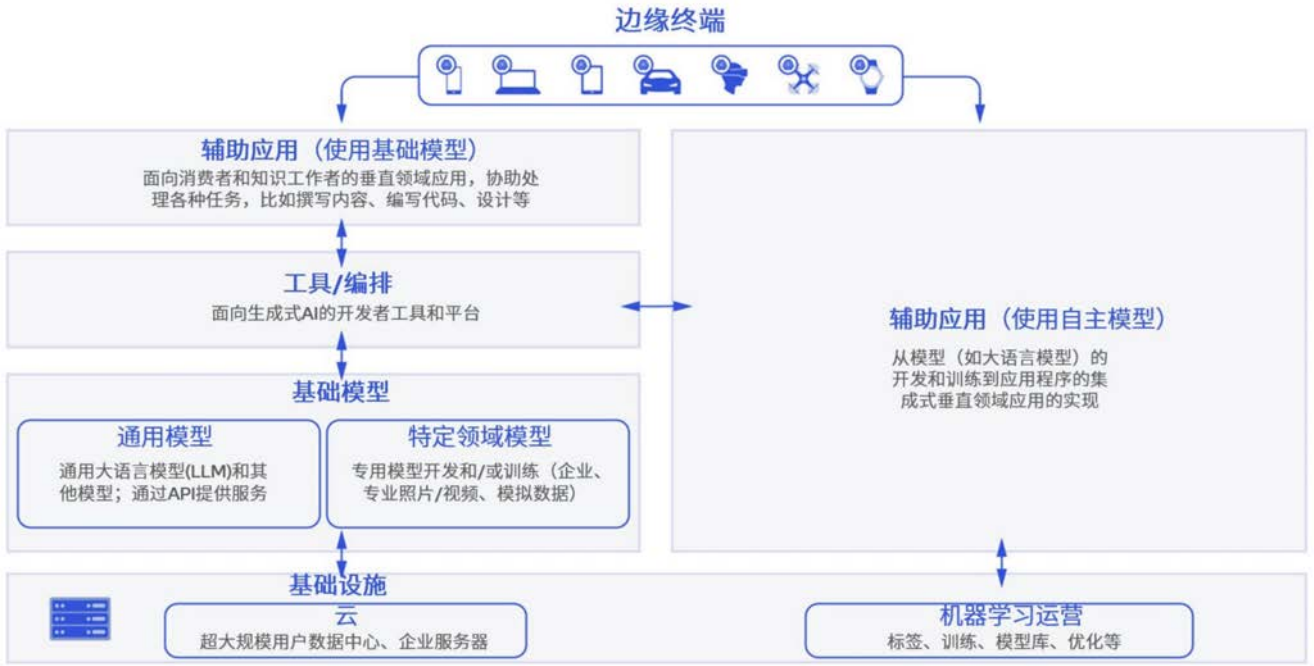


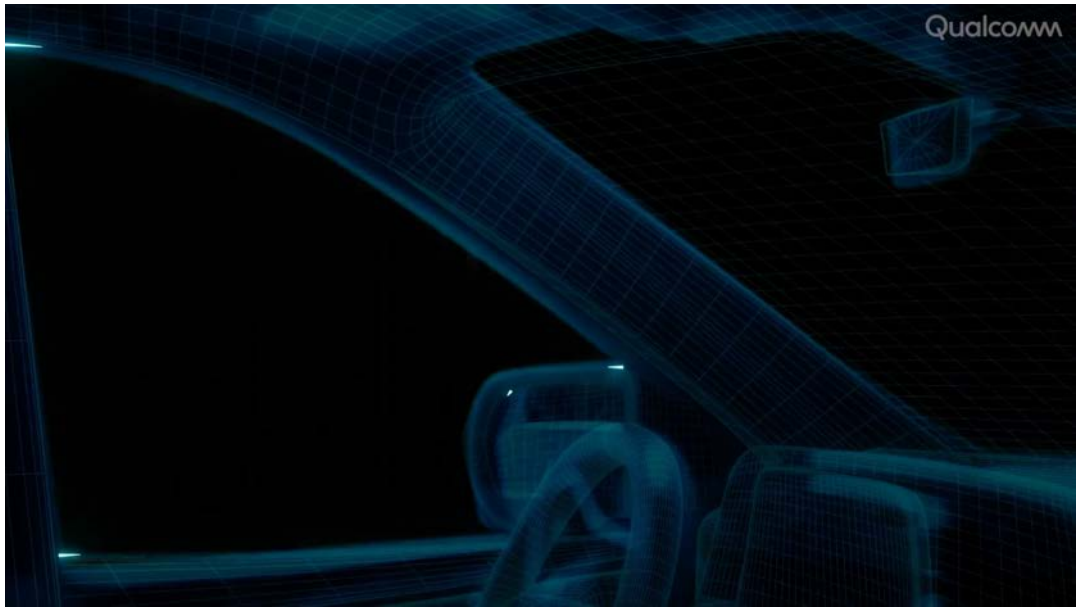
突破了端云协同计算在**分布偏移、资源受限**设备上训练推理效率局限

PART 04

案例分析

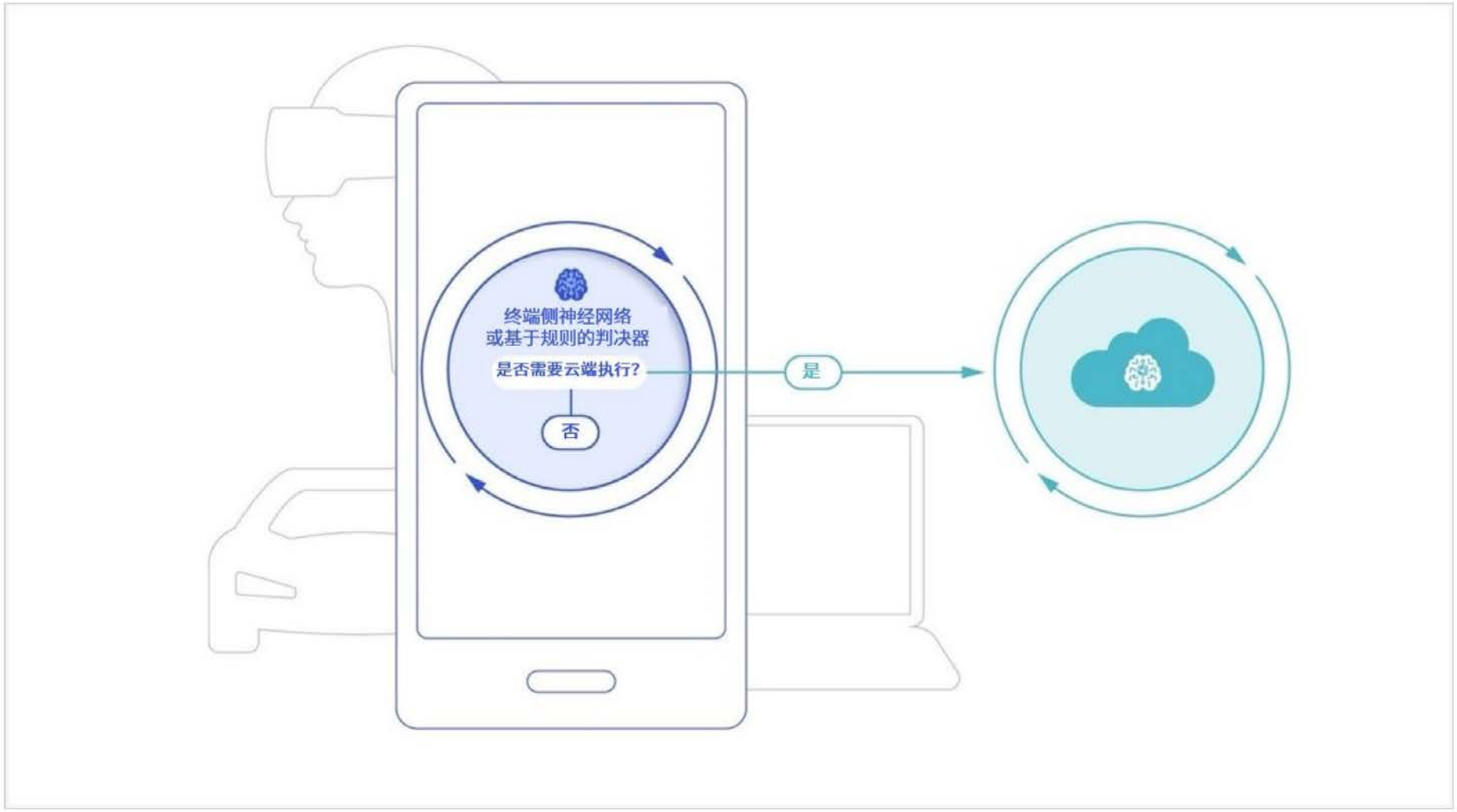
高通：生成式端云混合智能





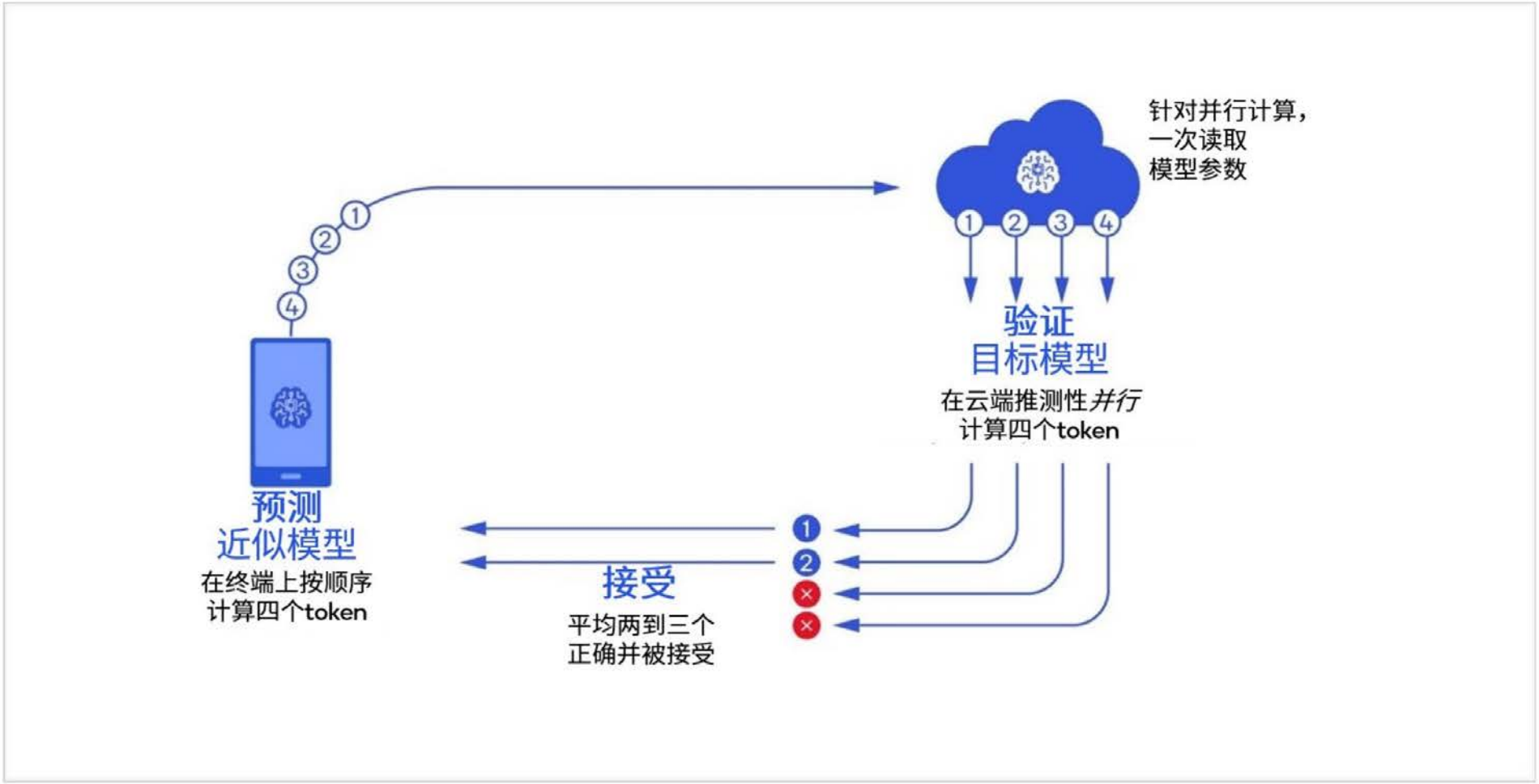
- 混合AI指终端和云端协同工作，在适当的场景和时间下分配AI计算的工作负载，以提供更好的体验，并高效利用资源。在一些场景下，计算将主要以终端为中心，在必要时向云端分流任务。而在以云为中心的场景下，终端将根据自身能力，在可能的情况下从云端分担一些AI工作负载。

端云协同智能



-- 高通《终端侧AI 和混合AI 开启生成式AI 的未来》

端云协同智能

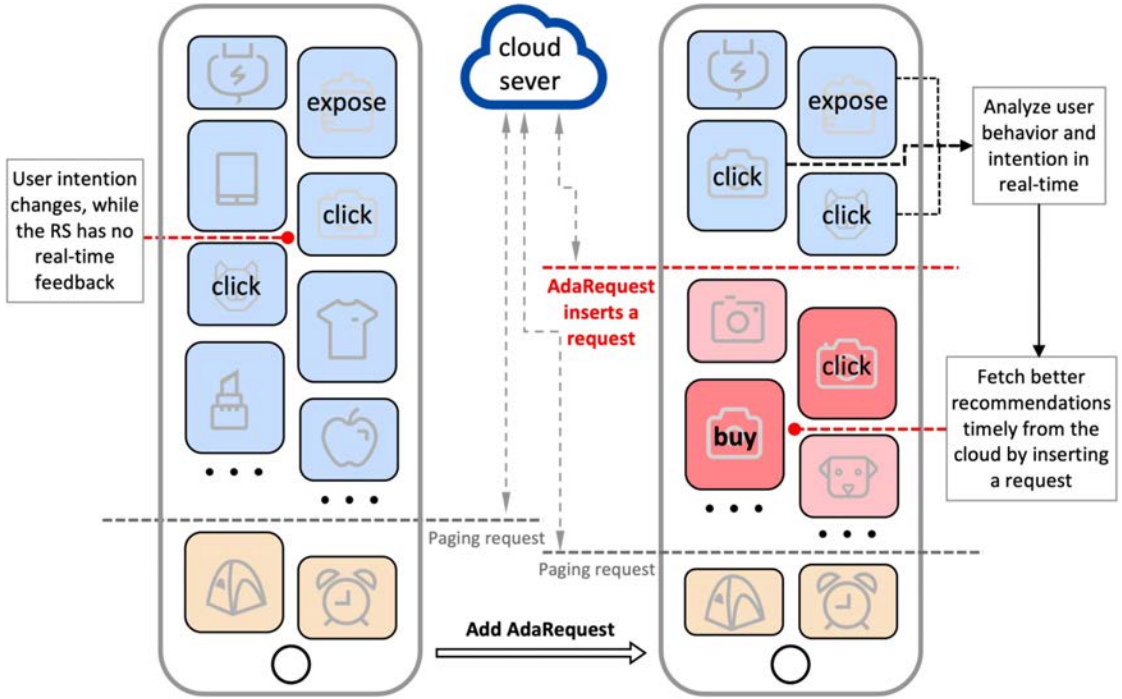


-- 高通《终端侧AI 和混合AI 开启生成式AI 的未来》

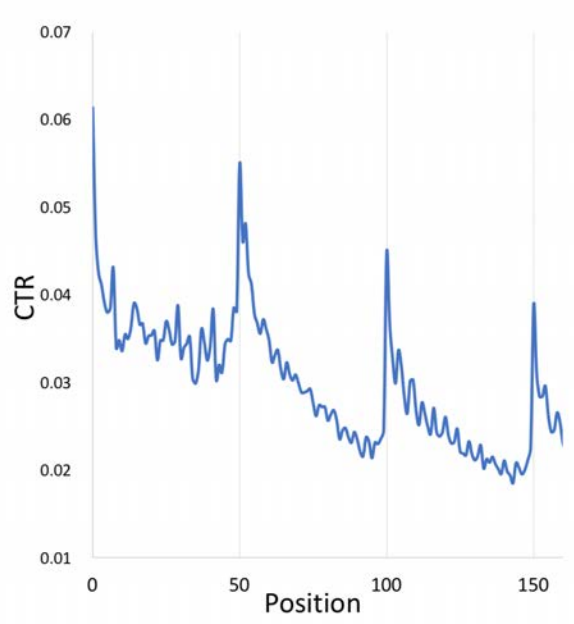
端云大-小模型协同推断算法

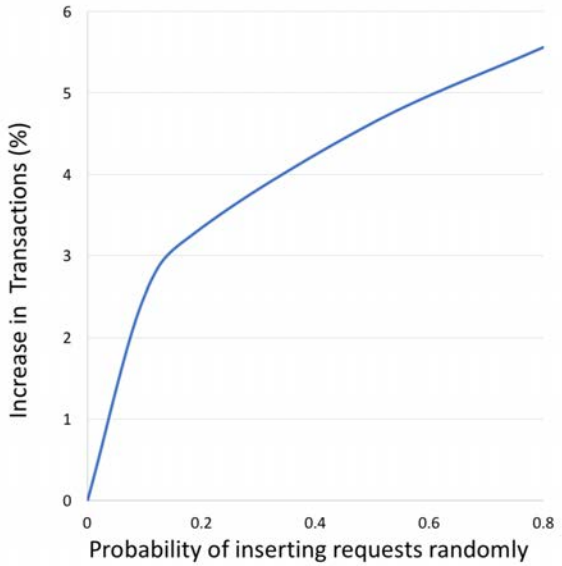
- 动态变化的端环境导致资源有限情况下云模型的延迟响应，导致端侧服务与端侧环境的不匹配，损害用户的服务体验

手机淘宝商品推荐系统



用户点击率在云模型响应后陡升

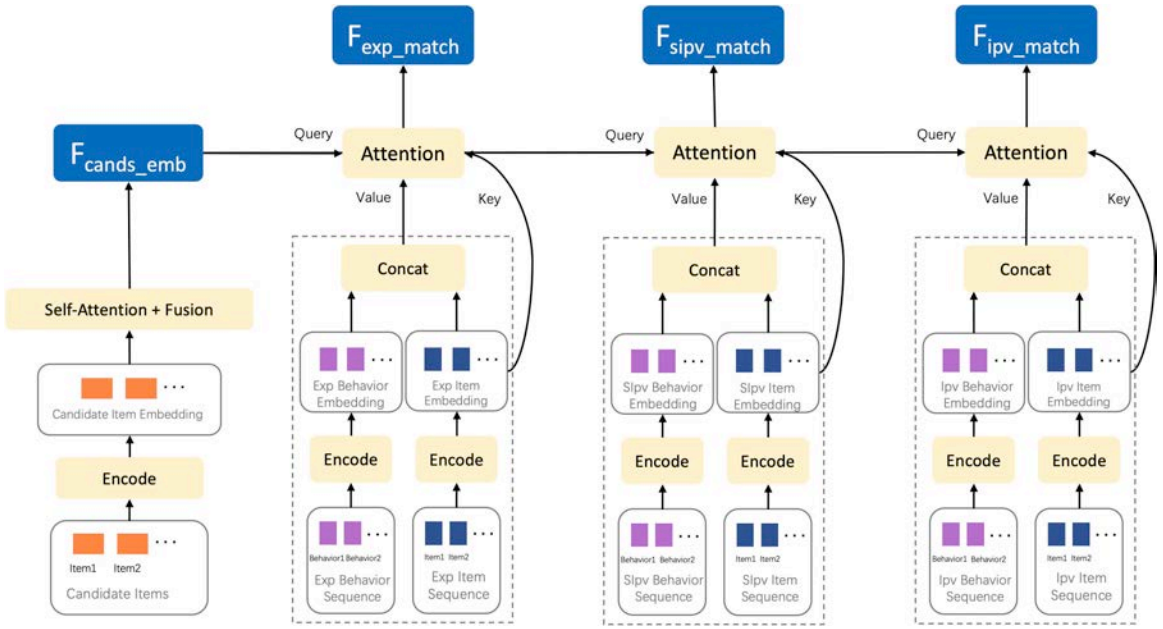


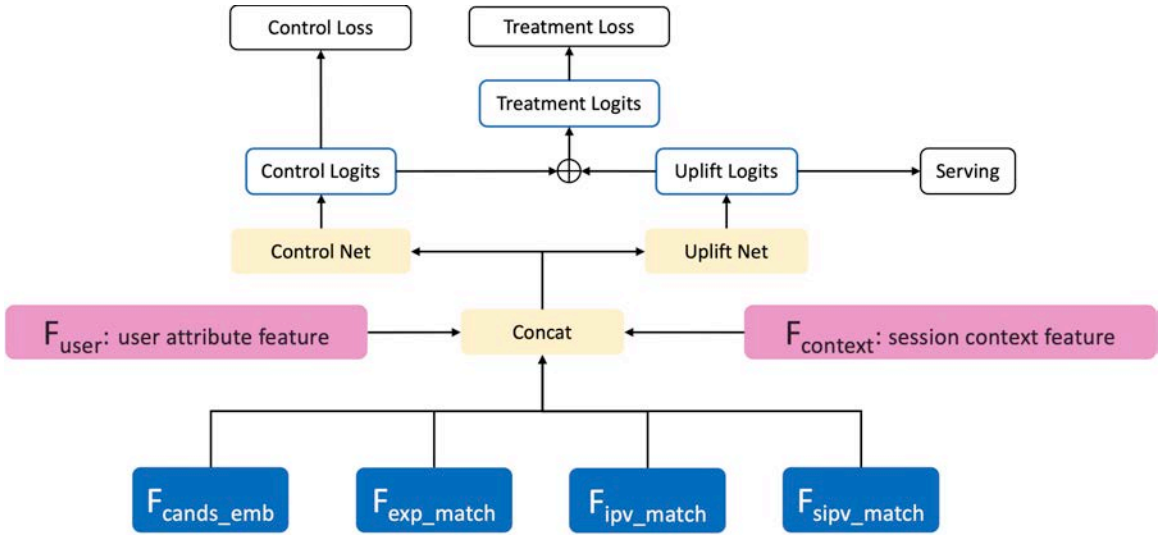


Xufeng Qian, Yue Xu, Fuyu Lv, Shengyu Zhang*, Ziwen Jiang, Qingwen Liu, Xiaoyi Zeng, Tat-Seng Chua, Fei Wu. Intelligent Request Strategy Design in Recommender System, KDD 2022

端云大-小模型协同推断算法

- 端设备部署小模型实时检测端环境变化（用户兴趣意图变化）
- 通过因果潜在结果模型预估请求大模型响应价值
- 动态规划对云侧大模型的请求,最大化资源有限时的线上收益。





Xufeng Qian, Yue Xu, Fuyu Lv, Shengyu Zhang*, Ziwen Jiang, Qingwen Liu, Xiaoyi Zeng, Tat-Seng Chua, Fei Wu. Intelligent Request Strategy Design in Recommender System, KDD 2022

端云大-小模型协同推断算法

当前推荐系统存在的问题

通信开销大
隐私破坏风险
隐时反馈噪声多
无法实时感知用户

因果结构学习机制
因果潜在结构框架
不确定性预估方法
…

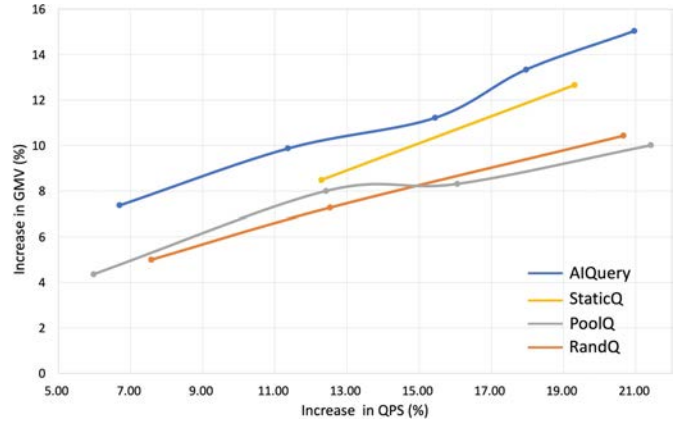
因果+端云协同

共性-个性协同
大-小模型协同
隐私-效率协同
…

直接经济效益
(购买率)

PR in N	NoQ	RandQ	PoolQ	StaticQ	AIQuery
10	0.889	1.174	1.177	1.130	**2.289**
20	1.660	2.197	2.164	2.045	**4.173**

平台经济效益
(商品交易总值)



Xufeng Qian, Yue Xu, Fuyu Lv, Shengyu Zhang*, Ziwen Jiang, Qingwen Liu, Xiaoyi Zeng, Tat-Seng Chua, Fei Wu. Intelligent Request Strategy Design in Recommender System, KDD 2022

IMRec: 用户视角下的多模态端云协同新闻推荐

目标

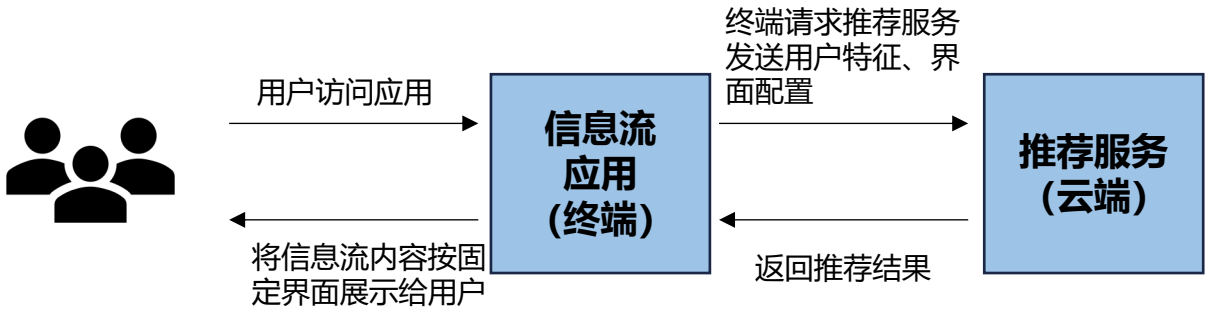
- 通过**用户端**侧界面元素对**用户的多模态关注点**进行挖掘，增强**云端**新闻推荐CTR预估性能。

动机

- 新闻是多模态载体，用户的兴趣往往会受到用户端侧的“第一印象”的影响。
- 想法：引入从用户**端侧**界面抽取的**视觉印象信息**预测用户兴趣点。

挑战

- 如何对“第一印象”合理具像化，易于深度学模型处理？
- 如何对具像化后的“第一印象”进行建模？
- 如何抽取和利用“第一印象”中空间位置、相对大小、风格等元素对多模态语义信息进行建模？



Xun, Jiahao, Shengyu Zhang, et al. "Why do we click: visual impression-aware news recommendation." In ACM MM 2021.

DisCover: 基于解耦的端云协同歌曲检索

目标

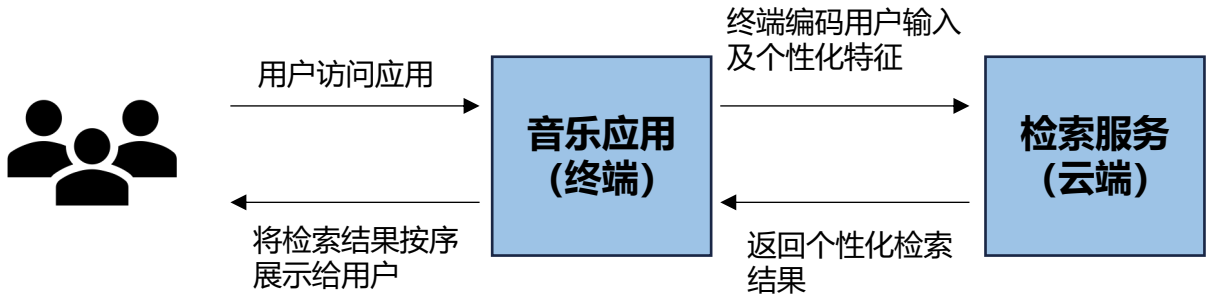
- 去除音频中的**偏差**并结合**端侧用户画像**，提高云端个性化音乐检索的精准性。

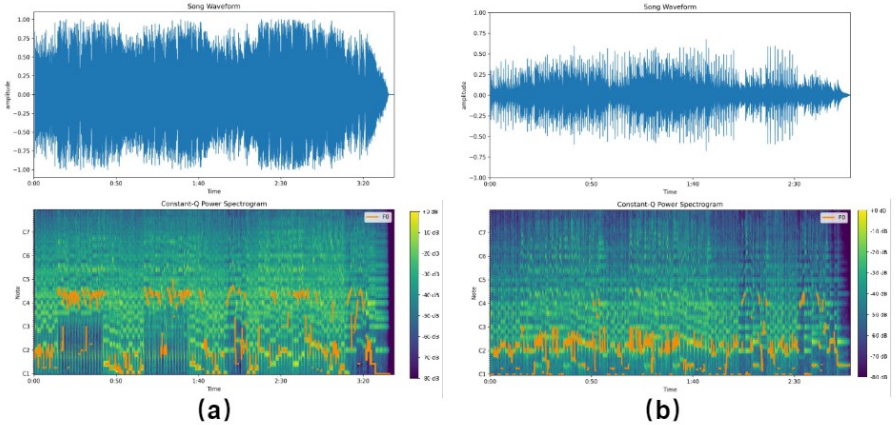
动机

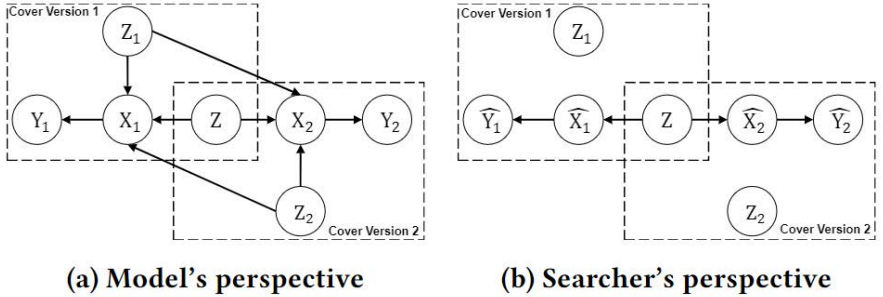
- 歌曲中往往存在许多**冗余的干扰信息**，且用户兴趣往往隐藏在**端侧的用户特征**中。
- 想法：利用**因果解耦**及**端侧用户特征分析**，提升端侧音频编码器质量。

挑战

- 同一首歌的不同版本也存在**较大差异**，这些差异具体现在音调、节奏、音色和类型等音乐属性（翻唱信息）上。因此，如何能利用同歌组学习到与**翻唱信息无关的本质的特征**是该任务的挑战。







Xun, Jiahao, Shengyu Zhang, "Discover: Disentangled music representation learning for cover song identification." In SIGIR 2023

PART 05

总结

总结

云向端泛化

面向异构端设备、复杂端场景的大规模预训练

端向云汇聚

融合不同场景、不同能力的统一模型

协同推断/训练

结合端智能和云智能的优势互补，端云一体化

端云协同的分布式机器学习系统

异构设备：摄像头、手机、机器人
复杂能力：听觉、触觉、视觉
动态环境：天气、活动、光照

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing

K+峰会

K+峰会 · 敦煌站
K+ 思考周®研习社
时间：2025.08.29-30

K+峰会 · 上海站
K+ 金融专场
时间：2025.10.17-18

K+峰会 · 香港站
K+ 思考周®研习社
时间：2025.11.25-26

K+峰会详情

AiDD峰会

AiDD峰会 · 上海站
AI+研发数字峰会
时间：2025.05.17-18

AiDD峰会 · 北京站
AI+研发数字峰会
时间：2025.08.08-09

AiDD峰会 · 深圳站
AI+研发数字峰会
时间：2025.11.28-29

AiDD峰会详情

AiDD AI+研发数字峰会 第5届
AI+ Development Digital summit

利用AI技术深化计算机对现实世界的理解

推动研发进入智能化时代