2024 年 CCF-蚂蚁科研基金绿色计算算效专项 申报课题介绍

一、 算力服务

1. GPU 跨域跨异构训练研究
2. 面向推理服务的动态显存管理研究
3. Kata 机密容器 GPU 安全性增强
4. 面向高效智能信息服务的 RAG（检索增强生成）策略优化与应用研究
5. 最小化推理成本：同构及异构模型极致合并部署降本研究

二、 算力治理

1. 面向高效任务协作的多智能体框架研究与应用
2. 面向大模型集群的网络监测与诊断方案

三、 智算网络

1. 自适应多路径高性能网络传输协议
2. 高性能异构集合通信优化技术

四、 AI Infra

1. 面向大语言模型的混合位宽训练优化技术研究
2. 面向大模型的在离线 GPU 混部技术
3. 高效分布式推理与异构算力调度
4. 大规模异构计算环境下大模型训练性能和稳定性提升方法研究

五、 模型算法

1. 面向超长上下文的大模型绿色部署优化应用研究
2. 面向端侧的绿色小模型技术研究和应用
3. 多方时空大模型绿色联合计算
4. 基于大模型的多模态评测系统研究
5. 面向 APP 应用的终端大模型推理性能优化
6. 基于多源异构用户理解大模型的人群圈选
7. 风控模型算力优化研究
8. 面向异构场景下多模态大模型训推一体化最优规划研究与应用

一、 算力服务

1. GPU 跨域跨异构训练研究

课题背景

当下供应链安全由于特殊原因受到了极大的挑战，多品牌战略有效缓解了供应问题，但是带来了新的挑战。目前大模型训练对于算力资源的诉求日益增加，动辄上万卡的训练集群依然在向着数万甚至数十万卡的规模延生，对于资源拥有者而言如何能够有效的统一化资源的使用，将异构卡统一组合为统一的更大的资源池用于分布式训练是一个急需解决的问题。对于联合统一资源池进行训练，不仅需要解决跨机通信问题还需要解决调度算力一致性问题，以及不同卡对于算子精度等一致性的问题。

研究方向（可选 1～2 个方向深入研究）

（1）跨机集合通信

不同品牌的算力卡尤其是 NVIDIA（GPGPU 架构）与其他厂商例如 DSA 架构的算力卡，在通讯库上均有自有集合通讯库，如何实现跨卡通讯，尤其是高效通信是联合训练的首要问题。

（2）GPU 算力精度问题

对于不同架构的 GPU 加速卡，GPGPU 架构以及 DSA 架构，两种架构除了设计理念，架构上的区别外，在精度上也有不同表现，需要研究一种方法实现在不同精度的前提下实现分布式并行的方法。

（3）GPU 算效等价模型

不同 GPU 的算力不一致，受通讯效率以及拓扑结构等众多因素影响，如何能够基于算效一致性模型进行调度成为制约整个训练效率提升的重要因素。需要建立一个算力等效模型，为算力调度及算力平衡作为依据。

（4）跨域并行训练加速

探索如何合理选择和调整并行策略，根据实际带宽优化跨域网络传输，使得训练作业能够在不同的集群间高效并行运行，提升训练效率。

本项目可以实现异构 GPU 集群算力融合，在 AI 基础架构侧实现最终一致性，有效屏蔽异构带来的调度问题，通信问题，伸缩问题，集群规模等问题。

预期目标和产出

本研究计划开发出一个面向分布式大模型跨域跨异构训练系统，实现多集群异构卡（NVIDIA 与其他 DSA 架构加速卡）间高效整合，实现跨域训练效率损失不大于 20%，跨卡训练相比同卡训练性能（在归一化算力情况下）损失不大于 10%，具体指标根据所选方向可以有选择性。

（1）1 套分布式训练并行方案及原型软件/文档，能够实现跨域跨异构高效通信，并最终实现性能损失低于 10%（根据方向选定，产出物可有方向侧重点）；

（2）1 篇蚂蚁认可的高质量顶级学术会议 CCF-A 类论文；

（3）申请 2 项以上专利。返回首页

2. 面向推理服务的动态显存管理研究

研究背景

高效的显存管理对于提升大模型推理服务的吞吐量和降低 AIDC 运营成本至关重要。显存利用效率的提升也是业界关注的热点。从解决碎片化（如 PagedAttention）、动态内存管理（如 vAttention）、memory tiering（如 ServerlessLLM、AttentionStore）到分布式管理、策略优化和压缩技术（如 FlexGen），围绕显存优化的研究十分活跃。本课题旨在通过优化推理引擎的显存管理策略，提高模型混部成功率、长上下文应用效果和推理业务吞吐量。

研究方向

（1）显存消耗预测算法

研究如何在推理过程中结合众多的 GPU 指标以及流量特征预测显存的消耗量。

（2）按需申请显存的动态显存管理算法

研究如何在推理引擎不预先分配实际显存的情况下，按照请求的实际需求按需取申请显存，同时确保不会因为分配显存的延迟影响在线服务的时延。

（3）在离线推理场景下的混部策略

研究如何将在线推理服务和离线推理服务混部，最大化推理卡的 GPU 利用率，而且不影响在线服务的 SLA。

预期目标和产出

（1）动态显存管理的推理引擎原型

通过本课题的研究，产出一个能够按需申请显存的动态显存管理的推理引擎原型，启动时不预留显存，接收流量时按需申请显存，同时尽量减少因为分配显存导致的时延问题。需要提供 benchmark 报告，证明 TTFT，TPOT，RT 等指标符合性能要求。

（2）提升 GPU 利用率的混部策略

结合动态显存管理，提高 GPU 显存的利用率，减少显存的闲置问题，同时结合实际环境的流量特征提供 benchmark 证明该混部策略的有效性。

（3）1 -2 篇相关领域的创新专利或者软件著作权。

（4）发表 1-2 篇 CCF-A 类会议论文。

（5）交付一套完整可执行的原型系统及相关文档。交付一套完整的代码和文档，并鼓励整合开源，以促进技术交流和行业发展。

返回首页

3. Kata 机密容器 GPU 安全性增强

研究背景

Kata Containers 是由蚂蚁团队参与发起和维护的 OpenInfra Foundation 顶级开源项目，是云原生行业安全容器的标准实现。传统的 Kata 容器以虚拟机为安全边界，允许在容器中运行不可信的代码并保护容器基础设施不被攻击。TDX/SEV 是 Intel 和 AMD 推动的最新的 TEE（Trusted Execution Environments）实现方式，把内存加密技术和虚拟化技术结合起来，克服了上一代 TEE 技术的兼容性缺点，让 TEE 技术被广泛使用成为可能。

基于 TDX/SEV 等 VM-TEE 技术，Kata Containers 实现了一套机密容器方案，把 Kata 的安全边界从保护基础设施扩展到了同时保护容器执行内容。这项技术适合用在需要高数据保密级别的场景，非常适合用来保护在第三方提供的基础设施中运行和保存的蚂蚁业务数据信息。

同时，随着 AI 大模型训练以及大模型推理等在云上部署的需要，对于第三方用户的训练数据和推理数据进行保护也势在必行。因此，使 TEE 能够保护需要 GPU 处理的数据也成为一个必然。然而，在支持 GPU 的 TEE 方面存在几个关键挑战。首先，大多数 GPU 硬件缺乏机密计算（CC）特性、信任根和内存加密模块。其次，连接 CPU 和 GPU 的通道通常不受信任，当数据需要在 CPU 和 GPU 之间流动时，需要进行数据加密/解密。NVIDIA 虽然发布了几款支持 CC 特性的 GPU，但是由于对中国禁止售卖，所以国内很难拿到这样的 GPU 卡。

研究方向

（1）验证 TDX/SEV 的根信任机制和可信传导链路，对存在的问题提出修复建议或方案；

（2）基于 virtio 标准的构建可信 IO 通道，实现可信的虚拟化 IO 链路；

（3）提出适合 Kata 机密容器的通用 GPU 数据加密方案，推动形成行业标准。

预期目标和产出

在 Kata 安全容器和 CoCo 社区开源项目框架内合作完成相关开源项目（Kata, QEMU, Linux kernel 等）的特性增强，PR 和文档合入上游代码仓库。

（1）专利：1-2 项国内或国际专利。

（2）论文：1-2 篇 CCF-A 类或者同等级领域内顶级会议或期刊论文。

返回首页

4. 面向高效智能信息服务的 RAG（检索增强生成）策略优化与应用研究

研究背景

随着人工智能技术的快速发展，智能化信息检索和生成系统在各个领域中的应用日益增多。然而，现有的 RAG（Retrieval-Augmented Generation）策略在实际应用中存在一些瓶颈，例如检索内容的准确性不高、上下文不一致、系统性能不佳等问题。这些问题限制了 RAG 在智能客服、财保、医疗等场景中的广泛应用。学术界和工业界已提出了诸多优化方向，如上下文增强检索（Contextual Retrieval）和改进的检索算法（如 BM25），但这些方法仍有提升空间，特别是在大规模知识库下的实际应用中。

本课题旨在通过创新性的检索算法优化与上下文建模，提升 RAG 在实际应用中的检索精度和上下文一致性，并推动其在智能信息服务中的有效落地。

研究方向

（1）检索算法优化

研究如何优化检索算法，提升系统在大规模知识库中的信息检索质量。研究可参考但不限于现有的方法（如上下文检索、BM25、Dense Retrieval），探索如何在实际应用场景中有效减少噪声数据的引入，确保检索结果的相关性。

（2）上下文建模与生成优化

探索在多轮对话或长文本生成中，如何保持上下文的一致性和连贯性。此方向鼓励创新型方法，例如通过动态调整上下文窗口大小、优化模型与检索段落的交互方式等，提升生成内容的连贯性。

预期目标和产出

（1）智能检索系统原型构建

构建一个结合优化后的 RAG 策略的端到端智能检索与生成系统，能够高效检索外部知识库并生成高质量、上下文相关的个性化内容。

（2）检索与生成效率提升

- 检索精度提升：相较于优化前的基线系统（无优化或仅采用基础 BM25），优化后的系统应能将检索精度（例如，通过 NDCG 或 Recall 评估）提升至少 30%。
- 上下文一致性提升：在多轮对话或长文本生成任务中，优化后的模型应使上下文一致性得分（如 BLEU 或 ROUGE）相较基线系统提升 20%。
- 响应时间优化：在大规模知识库下，优化后的系统应能相较于未优化的基线系统减少系统响应时间至少 15%，确保在高并发请求场景下依然能保持较好的实时性。
- 生成错误率降低：通过检索和上下文优化，生成的语义错误率相较基线系统降低至少 10%。

（3）发表 1-2 篇 CCF-A 类会议论文。

（4）申请 1 项创新专利或软件著作权。

返回首页

5. 最小化推理成本：同构及异构模型极致合并部署降本研究

研究背景

大模型应用市场规模近几年得到了飞速发展，2024 年大模型应用市场规模将达到 217 亿。随着业务需求的不断增长，越来越多的企业提出了私有化部署和低成本要求，以满足数据合规与隐私安全。

然而，一套完整的大模型 SaaS 应用软件如果要做到高准确率，就会依赖大量针对细分场景微调的模型，需要消耗大量的 GPU 资源，这无疑给企业带来了高昂的使用成本，导致大模型应用在私有化模式下无法普及。

研究方向

在 24GB 或更小显存下，实现模型的极致合并部署，包括:

（1）对同一个基础模型的不同 Lora 微调模型合并部署在同一张卡。

（2）将异构的基础模型合并到同一张卡，并且具备显存管理和限额能力。

（3）如何在低 QPS 场景几乎不损失性能的前提下，使用小显存部署比其容量更大的多个模型。

预期目标和产出

（1）研究并产出一套推理模型合并、显存管理算法，在低端推理卡（如 A10），实现同构和异构模型的极致合并部署。

（2）研究并产出一套显存与内存联合管理算法，在 QPS < 10 的小场景下，用低端推理卡（如 A10）部署比显存容量更大的多个模型，并且几乎不损失推理性能。

最终实现专有云场景下的最小化交付。

（3）1 篇顶级学术会议 CCF-A 类或 B 类论文。

（4）1 篇相关领域的创新专利或者软件著作权。

返回首页

二、算力治理

1. 面向高效任务协作的多智能体框架研究与应用

研究背景

随着人工智能技术的飞速发展，AI Agent 作为新兴技术正逐步成为推动 AI 应用落地的重要力量。同时为了支撑企业内部复杂场景的落地，在单智能体之上业界提出了类似于 AgentVerse、MetaGPT 等多智能体协作框架，促使这一领域进入新的发展阶段。多智能体系统是由多个自主的智能体组成的系统，这些智能体可以在一个共享的环境中相互作用、协调和协作，以完成复杂的任务。目前在多智能体的企业内部大规模场景落地探索中，主要面临几个主要问题：

（1）智能体协作稳定性

多智能体系统中的智能体必须能够稳定、高效地协作，尤其在处理复杂任务时，需要保证智能体之间的任务分配和依赖关系合理且可执行。当前的多智能体框架中，智能体在执行任务时可能因为信息缺失、任务重叠或冲突以及模型幻觉问题等，导致执行效率降低或中途失败。这就要求系统具备完善的任务分配机制和智能体间的协调策略，确保每个智能体在合适的时间点执行正确的任务。

（2）任务处理与评估机制

多智能体系统在任务处理的过程中，如何评估系统的整体性能和各个智能体的表现，是衡量系统是否稳定高效的核心问题。具体来说，评估应涵盖任务的完成速度、轮次消耗、任务质量，以及智能体在各个节点的执行情况。当前的评估标准大多集中于任务的完成情况，而忽视了对系统内各个节点的稳定性、智能体间信息传递的有效性等更细粒度的分析。因此，需要设计更全面的评估机制，以量化多智能体系统在复杂任务处理中的表现。

研究目标和产出

（1）研究多层次的任务调度和环境管理机制：如何在复杂任务处理过程中确保智能体之间协作的稳定性。并最终保证复杂任务的执行成功率在 70%以上。

（2）研究多智能体任务处理评测机制：构建评测模型、评测算法，针对多智能体运行过程中的整体目标完成情况、完成质量、子目标完成情况、完成质量、轮次消耗、指令遵从、幻觉率等方面进行评测，评测达到准确率 80%以上。

（3）发表 1-2 篇 CCF-A 类会议论文：在国际顶级会议上发表研究成果，提升项目的学术影响力。

（4）申请 1-2 项相关领域的创新专利或软件著作权：保护研究成果，巩固公司在 AI 和智能协作领域的技术优势。

返回首页

2. 面向大模型集群的网络监测与诊断方案

研究背景

为计算密集型任务的核心。大模型集群作为支持这些任务的基础设施，通常由数百到数千台计算节点构成。节点内部的计算资源（例如 GPU）通过高速的主机网络相连（例如 NVLink），其带宽可达 Tbps 级别（例如第五代 NVLink）；而节点间一般通过基于 RDMA 的高速网络相连，其带宽可达数百 Gbps。这种大规模的集群不仅对计算资源和通讯资源提出了高要求，还对监测与诊断能力提出了新的挑战。

监测与诊断方案的有效性直接影响到模型训练的效率和推理服务的稳定性。目前，大模型集群面临几个主要问题：

（1）复杂性与动态性

大模型集群的节点和连接数量庞大，且其流量模式与传统网络区别较大。例如，大模型训练流量的典型特点是能够预测、突发性强、峰值流量极大、具有周期性等，而推理流量表现为受到调度和分批处理的流量整形作用影响。随着任务的增加，网络负载和拓扑结构可能发生变化。这种动态性使得传统的监测方案难以适应。

（2）故障影响与检测延迟

在大规模环境中，由于分布式的集合通讯方式是主流，即使是单一节点故障和网络异常都导致性能下降甚至整个任务的失败。及时响应业务告警、快速定位故障区域以及分析故障类型成为了关键需求。

（3）多维度数据的处理

大模型集群的监测不仅需要关注计算节点的状态，还需综合考虑网络流量、存储状态等多维度的数据。这对监测系统的数据处理和分析能力提出了更高的要求。

研究目标

（1）研究充分覆盖大模型集群的网络监测机制，达到 99%故障覆盖率水平，有效降低网内和端侧故障识别和定位的时间，并在端侧与网内实现低峰值开销。

（2）研究面向大模型集群的故障诊断机制，有效降低故障类型判别时间，保证类型识别的综合准确率水平。

（3）研究自动化故障恢复机制，针对部分典型的网络侧及端侧问题，在故障严重影响业务质量的前提下，快速进行故障组件隔离与恢复，减少对于现有流量的影响。

预计产出

（1）一套大模型集群的网络监测与诊断的算法代码；

（2）1 篇 CCF-A 类论文；

（3）申请 1-2 项相关领域的创新专利或软件著作权。

返回首页

三、智算网络

1. 自适应多路径高性能网络传输协议

研究背景

大模型训练过程中，不同训练节点之间通过高速网络连接来同步模型参数、梯度等信息，流量具有低熵、周期性大象流等特征，传统 ECMP hash 负载均衡策略在 AI 数据中心里容易出现 hash 极化导致网络拥塞。

为了解决这种网络负载不均问题，网络侧目前存在两种常见思路：一种思路是在集合通信组网规划阶段通过控制器对通信 flow 流量路径进行规划来避免 hash 冲突，另一种思路是通过交换机芯片的包喷洒、flowlet、自适应路由等拥塞感知和自适应路径切换机制来实现流量负载均衡。这两种思路都存在一些问题，要么不具备混部任务等场景的普适性和快速响应网络变化，要么不能满足运维过程中的转发确定性。

端侧解决方案一般通过集合通信库层面建立多条 QP 连接的方式，将原本的端到端单条连接拆分多 QP 连接进行传输，增加网络中通信的熵值，结合网侧改进的 ECMP hash 算法，来降低 hash 冲突概率。部分解决方案可以在集合通信库层面感知多条路径上的网络状态，据此动态选择合适的路径进行通信。但是这种方式存在的问题是拆分多 QP 进行通信需要额外的性能开销，且集合通信库对于故障路径感知不敏感，无法做到快速感知并快速切换链路。

研究方向

（1）基于多路径的高性能网络传输协议

设计并实现基于多路径的高性能网络传输协议，协议层面原生支持多路径传输；

（2）基于多路径协议的拥塞控制算法

设计并实现一套基于多路径的拥塞控制算法，实现对多条路径的统一拥塞控制管理；

（3）基于多路径协议的路径调度算法

设计并实现一套基于多路径的路径调度算法，通过 ms 级感知网络运行状态，节点或者链路故障，各条路径的网络拥塞情况，智能并无感的进行路径切换；

（4）基于多路径网络传输的性能影响分析

分析上述方案在端到端部署场景下，整体的性能开销情况，包括但不限于网卡内存影响，网络包乱序，拆分多路径通信等对性能的影响；

预期目标和产出

本研究希望实现一套基于多路径的高性能网络传输协议，在避免多路径额外性能开销的同时，可以根据不同路径的网络拥塞状况自适应进行速率调整和路径切换，同时保障整体方案满足可运维的稳定性诉求。

（1）方案原型和仿真结果；

（2）发表 1-2 篇 CCF-A 类或者同等级领域内顶级会议或期刊论文；

（3）申请专利 1-2 项。

返回首页

2. 高性能异构集合通信优化技术

研究背景

AI 大模型已经成为引领下一代人工智能发展的关键技术。相比于传统模型，AI 大模型的参数规模庞大，需要在千卡、万卡的分布式集群环境下对海量数据进行训练。在训练过程中，不同集群服务器间需要进行频繁的数据交换，而这种交互具有广播式、超大流量、超低时延、超高频率、零容忍丢包和严格时间同步等特点，这对网络传输提出了极大性能挑战。另外在异构算力日益增长的趋势下，对于异构算力的连接也至关重要。同时，集合通信库 xCCL 作为连接整个 AIDC 算力能够被高效利用的关键，是底层异构硬件算力/网力和顶层 AI Infra 业务承上启下的位置。因此，研究高性能异构集合通信优化技术极为重要。

研究方向

（1）异构通信，通过研究异构跨芯的集合通信技术，实现 NVIDIA 系和国产芯片的算力连接，充分发挥异构算力聚合的能力。

（2）负载均衡，通过研究集合通信库的 QP 负载均衡、路径规划等端网协同的流量调度技术，实现流量在端-网全链路的均衡性，有效缓解流量冲突导致的拥塞问题。

（3）通信算子优化，通过分析集合通信算法以及传输流量特征，实现训练数据传输在空间和时间上的打散以及更优的集合通信算法，提升集合通信效率。

（4）网络容错，通过研究集合通信库的网络心跳保持、故障绕路等网络高可用技术，实现毫秒级网络故障容错能力，避免中断训练任务。

研究目标与产出

本课题目标在于探索面向 AI 大模型训练的高性能网络技术，旨在通过结合大模型训练流量特征，从异构通信、负载均衡、通信算子优化、网络容错等方面出发，实现 AI 大模型训练的高性能异构集合通信库技术。

（1）交付 1 套完整的系统代码及使用说明文档，基于提出的高性能网络传输技术体系，实现集合通信（AllReduce、AllGather 等）带宽利用率达到 95%以上；

（2）实现异构卡在同一个集群做集合通信，性能损耗<=20%；

（3）产出 1 篇 CCF-A 或蚂蚁认可的领域内顶级会议或期刊论文；

（4）申请专利 1 项。

返回首页

四、AI Infra

1. 面向大语言模型的混合位宽训练优化技术研究

研究背景

伴随着 ChatGPT 为代表的大语言模型(LLM)在自然语言理解、视频理解等领域的成功，如何将大语言的模型推向下一个高度是产业界和工业界共同关注的热点话题。然而，大语言模型的训练通常遵循 Scaling Law，追求更高的性能需要更多更优质的数据以及更大的模型规模，目前的 SOTA 大模型已经突破 400B 大关，但是支撑大模型训练的加速卡内存仍旧停留在百 GB 层次，如何缓和两者之间的矛盾，成为大模型预训练需要解决的关键难题。

研究目标和产出

针对大语言模型日趋庞大的规模和加速卡有限的内存空间之间的矛盾，提出一套高效的混合位宽大模型训练框架，包括 1）针对大语言模型的混合位宽训练方法，研究大语言模型的权重、激活、优化器等的混合位宽优化，探索如何在保证精度的前提下缓解大语言模型训练的内存压力；2）针对大语言模型的混合位宽优化方法，研究如何结合加速器的运算特征、访存特征加速混合位宽计算，提升端到端训练效率。

（1）1 套在国产智能芯片上支持混合位宽大模型预训练的框架系统及详尽的用户手册；

（2）1 篇 CCF-A 类论文；

（3）1 项软件著作权。

返回首页

2. 面向大模型的在离线 GPU 混部技术

研究背景

大模型蓬勃发展对异构算力如 GPU 需求巨大，一方面线上推理服务有较严格的性能要求（包括首字延迟和生成速度等）和稳定性 SLA（如成功率）；另一方面流量通常存在潮汐现象，在波谷阶段流量稀少，预留并常驻使用 GPU 资源容易导致巨大的资源浪费和成本问题。

如何既满足大模型在线性能和稳定性要求，同时显著提升异构算力集群利用率、降低成本具有重要的研究价值和实际效益（亿级别）。

预期目标和产出

面向大模型多卡部署场景（典型 LLM >=70B 参数，文生视频模型 >=10B 参数），本课题希望深入探索并研究有效提升大模型混部的关键问题、核心技术与方法，实现 GPU 算力的按需分配和 SLA 保障。特别是以下技术方向：

（1）**动态混部：**优化大模型服务计算和显存管理，实现算力和显存的按需动态分配；保障在线大模型首字延迟、生成速度和成功率不影响的前提下，提高资源利用率。

（2）**灵活混部：**针对不同大模型的特点，探索融合混部技术，提升总吞吐。

（3）**训推一体弹性混部：**充分挖掘全体集群的存、算、传资源，优化调度和分布式策略，实现高效、弹性的训推一体混部，提升吞吐性能或降低成本。

最终，希望在以上领域有方法创新和效果对比验证。典型产出包括：

（1）原型系统实现和对比分析报告；

（2）核心方法产出 CCF-A 类论文 >= 1 篇；或发明专利 >= 2 项。

返回首页

3. 高效分布式推理与异构算力调度

研究背景

模型更大：典型大模型参数变大、上下文序列变长，对算力需求越来越大；以及推荐模型增长迅速，达到数百 GB 甚至上 TB，在模型加载、更新和推理都面临更大的显存和性能压力。

算力多元：算力呈现多元异构并存现象，包括 GPU、CPU 以及多种国产卡；同时加速卡内部也有多种不同的算力单元。多种算力通过不同级别的互联实现协同计算和加速（例如 cross-bus, C2C, NVLink/NVSwitch、PCIe/CXL、RDMA 等）。如何充分发挥多种算力或组合提供高效的并行或分布式机制，从而更快、更高效、更绿色，具有重要的研究价值和经济效益。

预期目标和产出

面向典型 AI 推理场景，包括大模型（语言、多模态）和推荐模型等，本课题希望深入探索如何更好使用多元异构算力，提升推理性能或资源效能的关键技术与方法。特别是以下两个技术方向：

（1）高效并行与分布式推理：高效使用多卡、多机算力，或组合使用多种类型算力（CPU、GPU 等），提高系统扩展性和性能。目标模型包括 3 类（覆盖至少一种）：1）LLM（>=70B 参数）；2）多模态文生视频（>=10B 参数）；3）推荐模型（>=100GB 参数）

（2）推理请求高效弹性调度：实现总吞吐显著提升或能耗、成本显著降低。

最终，希望在以上领域有方法创新和实测效果优化。典型产出包括：

（1）原型系统实现和对比分析报告；

（2）核心方法产出和 CCF-A 类论文 >= 1 篇；或发明专利 >= 2 项。

返回首页

4. 大规模异构计算环境下大模型训练性能和稳定性提升方法研究

研究背景

近期，蚂蚁集团在基础大模型建设方面投入了大量资源，其中用于大模型训练的计算资源在类型和数量上都得到了大幅扩充。除了继续增加 Nvidia GPU 的数量外，蚂蚁还引入了来自不同国产厂商的大批量加速卡。为充分利用这些算力资源，我们必须持续提升训练性能和稳定性。大规模异构计算带来的新挑战包括：

（1）在超大规模（如万卡资源）训练中，慢节点、网络拥塞等问题严重影响训练效率；

（2）超大规模训练过程中，上下游的各类稳定性问题被放大，影响训练的稳定性，进而降低有效训练时长；

（3）不同厂商的加速卡具有各自的特性，例如特有的算子亲和性和网络通信特性，这些特性对训练性能和稳定性的影响各不相同，需进行针对性优化；

（4）不同厂商的基础软件栈存在差异，需要进行相应的适配；

（5）在各种针对性优化的背景下，需要确保训练代码的可维护性，保证快速切换硬件的能力；

（6）需要融合硬件厂商提供的技术栈与蚂蚁自身的 AI Infra 技术栈，以实现最佳整体效果；

（7）超大规模训练需要与算法工程师协同，权衡训练效果与效率，达到整体优化。

预期目标和产出

（1）1-2 篇 CCF-A 类 full paper；

（2）1-2 个发明专利；

（3）可在蚂蚁落地的训练策略源代码。

返回首页

五、模型算法

1. 面向超长上下文的大模型绿色部署优化应用研究

研究背景

模型服务化落地场景中，除了模型参数规模增大对显存占用增多，更大的显存来自于超长上下文的推理阶段产生的 KV 缓存。典型的超长下文场景包括多文档检索，文档摘要，多轮对话等。仅仅采用传统推理并行技术（模型并行/流水线并行），一方面需要大量的推理显存，成本过高；另一方面，底层硬件并行通信损耗难以为超长上下文场景提供可扩展的高吞吐服务。因此，对于超长上下文的支持也成为大模型学术界和工业界的研究热点。头部大模型公司争先推出支持百万超长上下文的服务入口，学术界也在探索从不同维度优化超长上下文的性能。 作为产学研的切入点，超长上下文推理优化既可以提升核心技术竞争力，又可以快速落地支持业务需求，我们期望如下方向的研究(包括但不限于):

（1）优化显存占用，使得长下文在 GPU 上可执行，推进如下技术的创新

- KV 缓存压缩优化：基于不同 token 对应 KV 缓存重要性分析，动态淘汰历史 KV 缓存，结合低 Bit 量化技术，降低上下文对于显存的占用;
- 输入压缩优化：采用输入词裁剪去除不重要 token、保留语义信息的总结、主旨 token 软压缩等方式；甚至结合 Encoding 的前置模型的输入压缩，降低最终输入给目标大模型的上下文长度;

（2）解决显存问题之后，进一步提升推理速度，推进如下技术的创新

- 模型架构优化：研发基于后训练或者预训练的新 attention 架构，在模型结构上降低推理时每个 token 对应的 KV 缓存占用和推理速度
- 新型的投机采样技术，比如 query 可以自适应计算，提前退出推理;

（3）优化服务链路：服务链路上将 prefill & decoding 阶段分离，采用分布式资源池化技术，充分利用 CPU cache、DRAM、SSD 等缓存资源，联合调度优化、弹性扩缩容策略，提升缓存复用率，降低超长上下文推理 TTFT;

预期目标和产出

（1）部署服务所支持的上下文长度提升：定义明确的上下文长度、模型推理精度、推理性能指标；和业界 SOTA 方案，有一定对比

（2）清晰的落地策略：提供可执行的源代码，或者集成到通用框架(TGI, TensorRT-LLM, vLLM 等)；提供清晰的实践报告和可行的落地方案。

（3）蚂蚁认可的 CCF-A/B 类会议论文 1+ 篇;

（4）提出 1+项创新提案专利申请。

返回首页

2. 面向端侧的绿色小模型技术研究和应用

研究背景

近年来，大模型在业界取得了显著成功，然而其庞大的参数量和复杂的注意力机制带来了巨大的资源消耗与成本。具体而言，注意力机制的时空复杂度为序列长度的平方，导致部署与推理过程中对计算资源的需求极高。端侧模型在微软 Windows，苹果 iPhone 和以及笔记本厂商的 AIPC 战略中逐步体现出重要性。端侧场景通常资源有限且需要快速响应，基于经典 transformer 的 LLM 的高复杂度对端侧使用构成了重大挑战。

在这些场景中，理想模型不仅应保持出色的性能，还应具备较小的参数量和时空复杂度。因此，研发性能优秀且计算效率高的端侧小模型，成为一个亟待解决的课题。在行业内，微软的 Phi 系列小模型通过使用高质量数据训练，证明了小模型也能取得优良效果。同时，Mamba 和 gated linear attention 系列的研究发现，整合有限的隐状态和灵活的门控机制，能够有效降低计算复杂度，从二次方缩减到线性，同时保持模型性能。这一领域的持续深入研究，旨在实现人人可用的小模型目标。

面向绿色端侧小模型的研究，不仅能显著提升技术的竞争优势，还能推动模型在实际业务中的快速落地应用。这一领域亟需探索的新理念和新方案，将为未来的智能设备与应用拓展全新的可能性。基于上述背景，我们期待在以下领域进行深入研究（包括但不限于）：

--下一代小模型架构研究：传统的注意力机制复杂性高，不适合端侧应用。目标是设计低复杂度且性能有保障的新型模型架构，以适应端侧环境的时空复杂度需求。

--训练端侧小模型的数据选择策略：高质量数据对提升模型表现的重要性已得到验证。然而，目前缺乏系统化的数据选择方案。考虑到端侧小模型的容量有限，迫切需要精心挑选数据进行训练，以最大化模型性能。

预期目标和产出

（1）通过优化模型架构和训练数据，研发出具有行业领先水平（SOTA）的绿色端侧小模型；

（2）蚂蚁认可的 CCF-A/B 类会议上发表 1+篇高质量论文；

（3）提出 1+项具有创新性的专利申请。

返回首页

3. 多方时空大模型绿色联合计算

研究背景

大数据时代人类活动与地理环境相关的时空数据被多种媒介所记录：如地图软件、车载定位、卫星遥感等。这些时空数据对于刻画区域以及用户的真实经营状况和信用水平有重要价值，可以有效的服务网商业务，如基于时空数据识别客户的职业、是否种植户、以及识别地区的富裕程度等。如何利用不同媒介不同渠道的时空数据，通过安全合规的联合建模是当前的重点工作，如何平衡联合建模环境下的有限算力资源（缺乏 GPU）和时空大模型的加工计算复杂度，是当前面临的核心挑战。一方面，时空数据规模巨大，基本 PB 级别，往往需要剔除其中的高速运动点、噪声点以及聚类采样，同时能够尽量保留有价值信息。另一方面，业界对于大模型的多方联合建模处于起步阶段，主要采用模型压缩方案，包括量化、剪枝、蒸馏、低秩分解等，我们需要折中考虑位置大模型的压缩率和效果损失。因此，需要针对时空数据预处理、数据降噪和压缩、模型压缩和部署等一系列问题提出一套整体联合建模解决方案。

预期目标和产出

基于时空大模型研究探索一种更高效更低资源的多方联合建模方案

（1）时空大模型在 TEE CPU 环境下高效联合建模：要求时空大模型压缩的同时保证效果；

（2）大规模时空数据的高效处理能力：要求同时减少信息的损失；

（3）一套满足实际业务应用的原型系统；

（4）一篇专利；

（5）发表一篇蚂蚁认可的 CCF-A/B 类会议高质量论文。

返回首页

4. 基于大模型的多模态评测系统研究

研究背景

随着多模态应用场景和数据规模的高速增长，针对多模态数据的人工智能算法设计日益复杂和多样化，传统评测框架已难以全面覆盖实际需求。在此背景下，基于大模型方法的细粒度自动化多模态评测系统将发挥重要作用。通过大模型算法优化多模态数据的采集、标注以及多模态算法的评测流程，不仅能够提升评测系统在不同数据模态与算法下的准确性与效率，还能帮助开发者快速定位和解决问题，优化算法性能和效果。

本项目旨在突破现有评测工具的局限，依托支付宝终端技术部的海量数据采集与分析能力，以及其对广泛业务场景的覆盖，构建一套全新的大模型评测方案与指标体系，降低传统评测所需的大规模人工成本，提升各业务场景下的评测效率和能力，同时助力算法优化更快更准，减少算法迭代次数和训练成本，助力业务高速发展。

技术价值：1）研究并建立基于多模态数据理解的评测大模型算法研究；2）建立完备的多模态场景评测数据集基线，使更多业务可以更快接入体验。

业务价值：通过精确的评测指标和计算工具实现 AIGC、音视频等业务的算法优化迭代次数的降低，减少训练调优成本，助力业务快速、绿色发展。

预期目标和产出

（1）开源评测系统：包含多模态评测数据、全面的大模型评测指标及其计算工具/可调用的计算接口；

（2）申请发明专利至少 2 项；

（3）产出 CCF-A 类论文 1 篇。

返回首页

5. 面向 APP 应用的终端大模型推理性能优化

研究背景

端 AI 使用海量用户设备完成本地模型推理，这种天然的分布式计算大大降低了 AI 应用的计算成本，是绿色计算重要的一环。在大模型应用爆发式增长的时代，将部分算力转移到终端手机成为智算领域新的课题。随着模型参数量的增加，端 AI 的价值背后也面临的更加严峻的挑战。国内外厂商和研究机构的技术发展也非常迅速，其中 3B 参数基础模型已经具备较强的能力，利用最新的终端芯片进行推理，使得端大模型也已经达到商用水平。但是在例如支付宝这类 APP 中，端 AI 面对着有限资源、模型下载、碎片化环境的额外挑战。因此，本课题从 APP 应用落地角度出发，从以下几个方面进行深入研究。

研究方向

（1）极致的模型压缩算法：APP 端大模型采用的是运行时下载的方式，物理尺寸太大不仅会影响运行时内存，更关系到触达到终端的成功率。相较于业界常用的 4bit 量化，APP 端模型压缩需探寻 2bit 以下的量化技术。

（2）场景模型构建：基于量化后基模，探究面向多场景垂类应用的模型微调方案，避免基础模型的频繁更新和保持多任务下的效果稳定性。

（3）推理性能优化：在 GPU/CPU/NPU 碎片化计算环境下，探究内存占用等资源极致优化和推理速度提升。基于不同硬件的特性，综合考虑基础模型压缩算法、垂类建模算法和引擎实现之间的相互影响，探究联合优化方案。

（4）端云协同新模式：结合智能助理等场景特点，探讨大模型生态下的端云协同模式，包括端云功能一致的模型内协同、端云完成不同任务的跨模型协同、端云场景交互的跨场景协同等方式，实现可行的业务落地方案。

预期目标和产出

（1）大模型压缩方案：通过本课题的研究，在精度损失小于 10%情况下，实现 1B ~ 3B 参数模型 2bit 以内的压缩技术，实现物理尺寸和内存占用的极致优化。

（2）高效的模型推理 SDK：在终端异构环境下，实现上述定制化量化方案的高效推理，可覆盖到支付宝 APP 60%以上机型。

（3）端云协同方案的业务应用：提出一套适用于支付宝 APP 业务的端云协同方案，将研究成果在智能助理等业务中进行验证和推广。

（4）产出 CCF-A 类论文 1-2 篇。

（5）申请发明专利 1 项。

（6）面向端侧大模型推理的源代码，性能超过业界公开 SDK。

返回首页

6. 基于多源异构用户理解大模型的人群圈选

研究背景

风控、营销、用增是支付宝内部重要的业务方向。随着相关需求的不断增加，对人力和计算资源的要求越来越高。传统的方法需要累积样本、训练模型等过程，需要消耗比较大的人力资源和计算资源。面临成百上千的人群圈选诉求，传统的方式已经无法快速支持，因此需要找到更加快速、绿色的方法进行人群产出。

基于大模型的人群圈选通过融合领域专家经验，只需一句话的描述，通过一句简单的SQL 产出人群，对资源的消耗极小（约 1CU），相比传统方法节省资源几十倍，并能够快速的为业务提供端到端的人群圈选服务，助力业务拓展和降本增效。

技术价值：该方法探索了用户多源异构数据预训练算法，并创新的将用户行为与自然语言描述对齐，以一种带预测能力的方式更准确地识别目标人群。这种方法使得用户可以用简洁的语言描述其需求，而系统能够理解并圈选符合条件的人群。科研基金将进一步增强在结构化数据预训练、自然语言对齐上的技术深度。

自然语言处理的应用：该项目展示了自然语言处理（NLP）在用户行为分析中的实际应用，提升了用户交互的自然性和便捷性，降低了技术门槛。

大模型的优势：利用大规模预训练模型，可以有效处理复杂的语言理解任务，从而实现更高的圈选准确率。大模型的能力使得系统能够理解多样化的语言表达，提高了灵活性。同时冷启动效果好，不需要复杂的特征和模型设计。

行为数据的深度挖掘：项目能够结合用户的行为数据进行深度分析，识别潜在的用户特征和兴趣点，为后续的个性化推荐或市场分析提供支持。

预期目标和产出

（1） 1 篇 CCF-A 论文投稿；

（2） 完成 2 篇以上专利；

（3） 开发基于 Agent 和自然语言交互的系统，实现低资源消耗下有效的人群圈选。

返回首页

7. 风控模型算力优化研究

研究背景

近年来 AI 能力不断突破，模型越来越多的应用到业务场景中。然后在提供模型服务时，成本巨高不下，需要重点突破以下难点：

（1）如何对模型进行量化压缩，在不影响模型效果的同时尽少模型计算量和模型大小，并能对主流模型的量化压缩提供方法和框架。

（2）推理引擎优化，如何在推理过程中让算力性能最优，吞吐量最大？

（3）算力资源运筹优化，线上流量是不断变化的，如何依据流量变化动态调配适当的资源，减少资源浪费的同时保障线上服务的稳定性。

（4）异构算力推理，线上有着不同公司、不同的推理卡类型，如何使用异构算力进行模型推理？

由于风控的全场景、高响应要求，成本和耗时是制约多模态数据和大模型广泛在风控场景应用的最重要原因，降低推理成本、提升 GPU 资源利用率、RT 和稳定性是现阶段提升蚂蚁风控智能化水平的重要部分，对其他场景高实时响应的场景也有借鉴意义。

在技术上层面上，大模型由于其在各种任务中的出色表现而引起了广泛的关注。然而，大模型推理的大量计算和内存需求对其在资源受限场景的部署提出了挑战。业内一直在努力开发旨在提高大模型推理效率的技术。在量化压缩、推理引擎优化、算力资源运筹优化、异构算法上持续攻坚和突破，能降低推理成本的同时，也能持续保持行业领先。

通过算法优化、硬件加速、分布式推理、资源运筹优化、推理成本下降和运态模型部署模式与模型性能优化等技术手段实现目标。当前研究多集中在单点的模型压缩，或者集群的混部优化，工程优化与业务割裂，风控业务中存在多模型串联、多模型流量波动差异、cpu/gpu 混合计算、大模型和小模型混合编排等特点，因此大部分的研究不能解决实际问题，我们期望探索一套在风控场景具有普适性的模型部署架构方案，用于部署时的理论指导，包括但不限于如下研究方向：（1）大小模型协同推理架构优化（2）基于 agent 的日志分析与性能瓶颈定位技术（3）基于流量与资源预估的模型混部技术（4）模型合并与动态弹性路由技术。

预期目标和产出

（1）两篇顶会论文；

（2）一套模型量化压缩的代码系统；

（3）一套算力资源运筹优化的算法；

（4）通过在智能凭证业务上应用，整体模型 RT 提升 20%，助力业务完成智审率 90%；

（5）模型推理成本下降 30%。

返回首页

8. 面向异构场景下多模态大模型训推一体化最优规划研究与应用

研究背景

消金公司存在多个机房，包括 CPU、GPU 等多种算力，其中 GPU 算力存在多种不同型号。在保证数据安全的前提下，数据可以在跨机房、跨服务器流动。随着大模型在消金多个场景的落地，对异构算力的合理化利用提出新的挑战。消金大模型应用场景包括：智能营销中文案生成、跨域用户行为特征表征、交互式自证材料解析、自证材料真实性检测、投诉纠纷化解方案生成等，这些场景的数据覆盖文本、图片、视频、图谱等多种模态数据。不同场景存在共用和独享大模型基座的情况，场景对模型训练紧迫程度、推理时效各不相同、服务稳定保证不相同，实践中训练和推理共存，推理场景居多，训练较少。期望通过研究给出一套合理的大模型应用适配方案、算力分配和任务调度策略，实现计算/推理资源的最大化利用，减少碳排放。

研究方向

（1）多种算力混合部署和调度，实现消金公司多机房、多型号 GPU 的资源统一调度能力。

（2）训练和推理弹性部署，实现训练环节与推理环节灵活切换。

（3）最优规划与调度，根据不同场景的优先级、时效、算力需求、吞吐量等多种因素合理规划和制定算力调度策略。

预期目标和产出

（1）交付一套完整可执行的系统；

（2）产出 1 篇公司认可的高质量会议(CCF-A 类)论文；

（3）申请专利 2 项。

返回首页