专题出品人：黄国平 博士

腾讯AI Lab 专家研究员

腾讯翻译负责人，腾讯AI Lab 专家研究员，毕业于中国科学院自动化研究所，研究方向为机器翻译、自然语言处理。长期专注于交互翻译和机器翻译的研究与应用，在 ACL、AAAI、IJCAI、EMNLP 等人工智能领域顶级会议与 TASLP 等顶级期刊发表论文20余篇。

专题：LLM推理加速和大规模服务

地点：爱那里厅 1（三层）

随着 LLM 参数量的不断增加，100B 甚至 200B 或者更大规模的模型都在不断出现。但在实际业务中，7B~13B 模型的大规模服务在请求延时和服务成本方面都面临巨大挑战。所以如何通过工程或者模型的方法，在不增加硬件成本的前提下降低请求延迟，同时减少高 QPS 时的资源需求，就成为非常关键的课题。

本专题会邀请一些知名公司专门从事 LLM 推理加速和大规模服务的团队分享他们在 LLM 服务方面的经验，希望能对大家的工作带来一些帮助。

搭建端到端的机器学习平台最佳实践

by 郑培凌

PayPal
Tech Lead

by 韩红芳

PayPal
AI/ML平台解决方案研发经理

Cosmos.AI 是 PayPal 公司一个端到端的 ML 服务平台，目标用户涵盖了公司内部不同业务部门（风控，合规，客户服务等）的 ML 相关人员（包括数据科学家，业务分析师，开发人员和 MLOps 工程师）。Cosmos.AI 平台专注于提供一个自助式的，可靠的，可伸缩的，高效和安全的模型构建，训练，部署和决策平台，来标准化机器学习开发的生命周期（MLDLC）。随着 GenAI 技术的不断推广和进步，我们的平台也融入了对 LLM 的支持。

本场演讲将从平台架构，技术，功能，用例等各方面介绍 Cosmos.AI 服务平台，以及其中的一些最佳实践。

演讲提纲：

1. 数据整理与特征工程
2. 模型开发训练
   - 训练代码版本控制
   - 训练数据集管理
   - 训练实验跟踪
3. 模型部署
   - 高级部署
   - 模型优化
4. 对 GenAI 的支持
   - LLM 模型部署
   - LLM 模型语义缓存
   - LLM 模型推理加速
   - LLM 模型提示管理
5. GenAI 的应用
   - GenAI 小助手

听众收益点：

- 了解 Paypal 在 ML 平台构建上的解决方案和工程经验
- 了解 GenAI 技术在 PayPal AI/ML 平台的架构和实现
- 了解 GenAI 技术在 PayPal AI/ML 平台的应用和探索

高效且高度可配置的大模型推理引擎与服务

by 史树明 博士

腾讯
AI Lab自然语言处理中心总监、T14级专家研究员

大模型推理服务的速度、吞吐率、成本和易用性对于大模型的商业落地至关重要。本次分享将介绍如何设计和实现一款高效的、高度可配置的、容易部署的大模型推理引擎与服务。

听众收益：

- 了解常见的大模型GPU推理加速技术；
- 了解实用的大模型量化技术；
- 了解大模型GPU/CPU混合推理技术以及多卡推理技术；
- 高度可配置推理引擎的设计与实现。

LLM 模型压缩与推理加速实践

by 陈磊

小红书
中台技术部技术专家

自 ChatGPT 发布以来，大语言模型（LLM）以其令人惊艳的推断和生成能力震惊了世界，强人工智能时代的到来近在眼前。由于 LLM 巨大的参数量与计算任务，在满足线上服务延迟和吞吐要求方面面临较大的技术挑战，以 GPT-175B模型为例，它拥有1750亿参数，如果采用半精度（float16）存储的话至少需要320GB的存储空间。再考虑到 kv-cache 等其他存储需要，部署该模型进行推理至少需要5～6个A100GPU（每卡80GB显存）。巨大的存储与计算代价是横在 LLM 模型落地面前的一大难题。

面对这些问题与挑战，我们团队从模型压缩、算子加速以及分布式推理框架等多个方向上齐头并进，实现对 LLM 模型线上推理的综合提速，我们在多款 LLM 上已取得了业界一流水平的推理性能。

本场演讲将从模型压缩、算子加速以及分布式推理框架研发等多方面介绍小红书在 LLM 模型推理优化问题上的一些最佳实践。

演讲提纲：

1. 模型压缩
   - weight 量化
   - kv-cache 量化
   - weight & activation 量化
2. 推理框架
   - continuous batch inference
   - paged attention
3. 算子加速

听众收益点：

- 了解到 LLM 模型服务在实际业务部署过程中所面临的问题与挑战；
- 多种 LLM 量化压缩算法在实际业务模型上的实现与改进；
- LLM 持续批量推理等功能的开发与落地调整；
- 共同探讨与展望未来可能的改进方向；

云原生场景下 Fluid 加速 AIGC 工程化实践

by 车漾

阿里云
高级技术专家

人工智能生成内容（AIGC）和大型语言模型（LLM）在近一年内方兴未艾，进一步提升了大众对生成式模型的期望值。然而，正如 Gartner 报告中所提到的：“启动 AI 应用程序试点项目看起来轻而易举，但将它们部署到生产环境中则极具挑战性”，AIGC 模型推理服务相比于传统的模型，在云上的工程化落地存在许多挑战，包括如何应对模型复杂的架构和规模，计算资源需求，弹性扩缩容和模型更新等问题。而模型从存储加载到 GPU 的性能制约了弹性伸缩和模型频繁开级等核心场景。

在本次分享中，我将介绍在 Fluid 项目作为云原生AI场景下的数据和任务编排框架，在 AIGC 模型推理工程化落地方面做了许多优化探索的工作，包括简化云原生 AI 场景的分布式缓存管理和运维，降低资源成本；以及优化推理服务读取模型数据的效率，加速模型加载过程。我们也会演示如何通过 Fluid 将一个 LLM 模型的推理加载速度提升近7倍，同时提供缓存弹性的能力，避免资源浪费。

演讲提纲：

1. 背景与挑战
   - AI/大模型应用云原生化的趋势
   - AIGC 模型推理服务在云原生场景下的痛点和挑战
2. Fluid 对于 AIGC 模型推理场景的优化实践
   - Fluid 简介
   - 可自运维的计算侧分布式缓存
   - 数据流编排实现大模型缓存的自动化管理
   - 数据访问性能优化
3. 用户案例 & 演示
4. 总结和展望

听众收益点：

- 理解 AIGC 模型推理服务在云原生工程化落地中的挑战
- 对 Fluid 项目的了解，包括其背景、目标和优化方法，以及解决缓存管理、资源成本降低和模型加载效率的具体策略
- 学习解决云上 AIGC 模型推理服务挑战的实际方法，可以应用在类似场景和问题，并且通过演示了解实际效果
- 探讨未来展望和可能的改进方向