intel XEON

案例研究

直播服务
AI 虚拟形象
第四代英特尔® 至强® 可扩展处理器

携手英特尔，朝夕光年 A-SOUL 团队以高效 LLM 推理方案打造高品质 AI 虚拟形象

朝夕光年

“高效的 LLM 推理是实现 AI 羊驼出色表演的重要保证。在本次的合作中，我们与英特尔一起，基于第四代英特尔® 至强® 可扩展平台对 LLM 实现了非常完善的推理任务加速适配，并开展了完善的测试和对比。全新的 Super-fused LLM FP16/AMX BF16 推理加速方案让我们以很低的投入就能完成开发部署，并实现超出预期的推理性能和性价比。”

A-SOUL 团队
朝夕光年

前言概述

人工智能（Artificial Intelligence，AI）虚拟形象正在直播行业获得越来越多的关注，而表现优秀的 AI 虚拟形象背后必然是高质量的 NLP（Natural Language Processing，自然语言处理）能力提供的有效支撑。作为直播行业的领军企业，朝夕光年 A-SOUL 团队也正通过打造 AI 羊驼 - 阿花产品，为观众提供更具个性化的观看和互动体验。

为了让阿花具备良好的语言和行为成长曲线，A-SOUL 团队在后台交互式系统中，加入基于 LLM（Large Language Model，大语言模型）构建的 ChatAI 对话生成模型来为阿花提供 NLP 能力。而为了在良好的推理成本性价比下获得卓越的推理性能，从而保证阿花与观众之间的实时互动能力，A-SOUL 团队与英特尔合作，在引入第四代英特尔® 至强® 可扩展处理器作为新算力核心的云平台中，引入英特尔打造的 Super-fused LLM FP16/AMX BF16 推理加速方案。经过多轮测试表明，这一优化方案能让 A-SOUL 团队经济地获得超出预期的推理性能和性价比。在单实例场景下，AI 羊驼方案中不同 LLM 能获得 1.89 至 2.55 倍的推理性能提升；而在多实例场景中，推理性能可在单实例基础上进一步提升至原有的 1.16 至 1.2 倍。[1]

方案背景：实时互动的 AI 虚拟形象亟需 CPU 平台提供强劲推理性能支持

AI 技术的高速发展正让 AI 虚拟形象在直播行业获得更多青睐，与传统真人主播相比，AI 虚拟形象不仅无需专门的职业培训且不受工作时长限制，也能有效帮助电商等领域的用户提升运营效率，降低营销推广成本。同时形态各异的虚拟形象也可在 AI、CG（Computer Graphics，计算机图形学）等 IT 前沿技术的加持下展现出更酷炫的视觉效果，与观众的交互也更实时、更具感染力。尤为重要的是，AI 虚拟形象的语言与行为能

伴随与观众交流、交互的增多而不断成长变化，这会让观众始终有耳目一新的感觉，在产品黏性上更胜一筹。

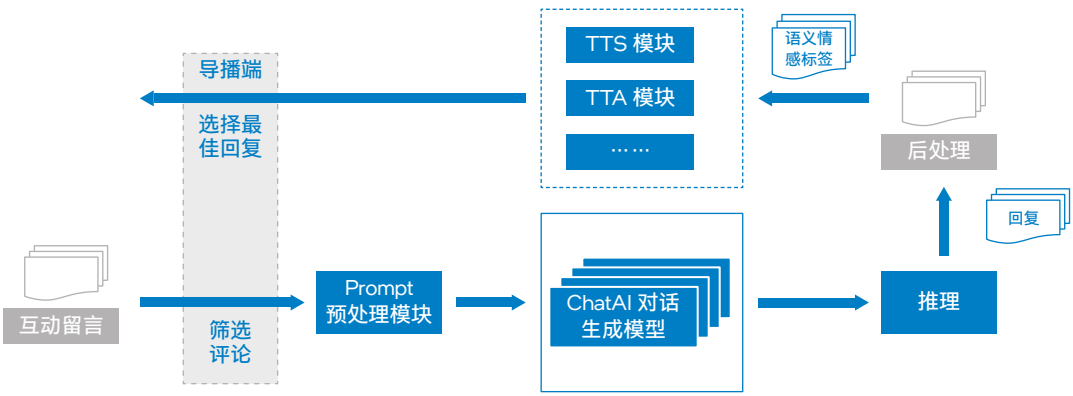
作为直播行业的翘楚，A-SOUL 团队在虚拟形象的打造和运营上也一直走在行业前列。如图一所示，AI 羊驼（角色名：“阿花”）是 A-SOUL 团队近期面向直播领域推出的一个交互式 AI 产品。除了有着“动物外形 + 萝莉声线”的二次元设定外，A-SOUL 团队也为其设定了完备的形象成长曲线，并具有动物和人两种形态。初期 AI 羊驼 - 阿花会以“非人类且未具备成熟心智”的形象出现，在进行持续的 NLP 训练之后，其逐渐能够根据与观众的交互内容，提供新鲜点 / 爆点内容输出的能力，并在取得一定关注度后，会偶尔以客串嘉宾的方式加入到 A-SOUL（一知名虚拟偶像女团）的团播节目中。

图一 AI 羊驼 - 阿花动物形态及人形态[2]

为实现上述语言风格和行为特征的变化曲线，A-SOUL 团队为其设计了一个全新基于 NLP 能力的交互式系统。如图二所示，系统自带的对话直播导播软件首先会获取观众的互动留言，经导播端筛选后输入到 Prompt 预处理模块，这一模块负责对提示语进行加工，同时过滤掉有害词语。而后预处理过的、具有结构化格式的输入数据会进一步发送到多个 ChatAI 对话生成模型中。

对话生成模型集群中有着多路微调过的模型，数据在这里通过模型推理后，系统会对所生成的回复进行后处理，提取语义情感并作为标签同步到用于音频合成的 TTS（Text to Speech，文本转语音）、用于文本动画生成的 TTA（Text to Animation，文本转动画）等模块。这些模块都经过了 A-SOUL 团队的深度优化，尤其是 TTA 模块在结合了最新 motion diffusion 技术之后，能让 AI 羊驼 - 阿花实现更多更有趣的语言和动作表达。同时，系统的内容安全与合规对齐模块也会对内容进行敏感关键词、偏见内容的校准，避免回复存在不公平性或歧视性。

基于目前对中文有着良好支持的 LLM，A-SOUL 团队在上述 NLP 工作流程中采用了已在大量开源中文语料上进行了预训练（Pre-trained）的中文模型作为系统的基座模型，并在流程中予以微调（Fine-tuning）。其中，预训练过程是采用自监督学习（self-supervised learning）方法在大规模无标签文本数据集上进行。这一过程中，阿花对话生成模型学习到了大量的语言知识，例如语法规则、语义信息等。而微调是在有标签的对话数据集



图二 AI 羊驼交互式工作流程

上进行，AI 羊驼对话生成模型能根据不同风格的语料，从中进一步学习特定任务的知识，例如对话任务中的上下文理解和回复生成等。值得一提的是，A-SOUL 团队在方案中选择了开源可商用的 LoRA、QLoRA 及 RLHF 等 LLM 微调技术，参数量可控制在百亿级或更少，既可实现方案的轻量化部署，也能应对直播所需的实时性要求。

多维度的技术优化，是保证这一系统高效运行的重要前提。如针对微调过程中可能出现的过拟合现象，模型未完全理解输入语境，或可能对输入数据中的偏见进行过拟合等问题，A-SOUL 团队正通过对数据处理、模型训练以及内容安全与合规对齐模块的优化来逐一解决。而在方案的性能表现上，A-SOUL 团队也面临着以下挑战：

- **海量算力需求以及由此产生的计算成本。**特别是在系统的预训练阶段，数以亿计的参数和数据集处理需要基础承载平台具备强大的算力支持和突出的内存性能；
- **直播场景对实时性的严苛要求，需要系统能够快速生成内容，这对推理性能提出了巨大的挑战。**众所周知，LLM 之所以被称为大语言模型，一个重要原因就是其有着极为庞大的参数量（有些模型甚至可达千亿级别），这意味着系统需要大量的计算资源来开展推理，而在计算资源有限的情况下产生的过长推理时延，会使对话失去实时性效果。

虽然采用“堆算力”的方式可以部分应对上述性能挑战，但这在经济性上无疑得不偿失。尤其在 AI 羊驼 - 阿花方案这一相对轻量级的 LLM（参数规模在 10B 以内）上，过于厚重的算力堆砌会压低推理成本性价比，进而带来阿花运营成本的极大增长。

为了在良好的推理成本性价比下实现更优的方案效果，A-SOUL 团队计划引入更经济的 CPU 推理平台以及更有针对性的优化方案，并开展多方位的模型优化及硬件加速，但这并非易事。为此，A-SOUL 团队与英特尔一起，在第四代至强® 可扩展处理器的基础上，引入英特尔打造的 Super-fused LLM FP16/AMX BF16 推理加速方案，有效实现了 PyTorch 框架在 LLM 推理上的优化。后续的验证测试证明，优化方案在提升推理吞吐能力上有着明显的效果，在多轮验证测试中，单实例场景下，AI 羊驼方案中的不同 LLM 可取得 1.89 至 2.55 倍的推理性能提升。而在多实例场景中，由 IPEX（Intel® Extension for PyTorch，面向 PyTorch 的英特尔® 扩展）带来的优化，可令推理性能在单实例基础上进一步提升至原有的 1.16 至 1.2 倍，达到了 A-SOUL 团队对优化方案的预期。[3]

优化方案：基于第四代英特尔® 至强® 可扩展处理器的 Super-fused LLM FP16/AMX BF16 推理加速方案

凭借更优的矢量指令和矩阵乘法运算，第四代英特尔® 至强® 可扩展处理器可为各种场景提供出色的 AI 推理和训练加速，显著提升 NLP 等深度学习工作负载的性能。为更大程度地激发新一代处理器在推理负载上的性能表现，英特尔与 A-SOUL 团队一起，基于新平台推出 Super-fused LLM FP16/AMX BF16 推理加速方案，对用于 LLM 推理的 PyTorch 框架实现了专门的优化。

PyTorch 是目前部署 LLM 的主流 AI 框架之一，其在 AI 羊驼 - 阿花方案的部署、运行中有着不可替代的作用。例如在 AI 羊驼 - 阿花方案中 LLM 的实际部署中，需要使用 Hugging Face Transformers 中基于 PyTorch 开发的文本处理模块 model.generate()，因此整个优化方案必须围绕 PyTorch 模型和框架 API 展开。

在优化方案构建前，用于 AI 羊驼 - 阿花方案的 PyTorch LLM（参数规模为 1.4B）推理任务的吞吐性能约为 20 词元 / 秒 / 每处理器[4]，这距离 A-SOUL 团队预设的部署性能要求尚有一定差距。究其原因，是因为 PyTorch 在 CPU 平台上无法完全释放出第四代至强® 可扩展处理器在 LLM 推理任务上的全部潜能，这是因为**缺乏对 CPU 平台 LLM 推理优化的支持**。

虽然 PyTorch 2.0 开始具备了基于 CPU 平台的模型推理优化能力（基于 torch.compile 模块），但该模块在 CPU 平台上还

是只能对一些静态且精度为 FP32 的模型有效。由于 LLM 推理任务中的 MHA (Multi-Head Attention，多头注意力) 计算依赖于随生成词元自增长的缓存矩阵，这使得 torch.compile 模块需要面向生成词元最大窗口 (通常定义为上千，如 4,096) 生成庞大的执行代码，且优化模型所需的时间长达数小时。因此，目前 PyTorch 框架无法有效对基于 CPU 平台的 LLM 推理优化提供支持。

得益于第四代至强® 可扩展处理器提供的英特尔® AVX-512_FP16 及英特尔® AMX BF16 加速指令，其可以完美支持并加速 LLM 推理。基于这一新特性，英特尔推出了全新的 Super-fused LLM FP16/AMX BF16 推理加速方案，可用于弥补上述 PyTorch 在第四代至强® 可扩展处理器上进行 LLM 推理任务时的性能不足。新方案主要有以下三大优化“杀手锏”：

- **英特尔® oneMKL 加速计算**

Super-fused LLM FP16/AMX BF16 推理优化方案使用了英特尔® oneMKL (Intel® oneAPI Math Kernel Library，英特尔® oneAPI 数学内核库) 提供的矩阵乘算子来加速推理计算，能够在减少权值存储空间的同时降低内存带宽压力，在保持精度的前提下显著提升推理性能；

- **FP16 Flash Attention 算法加速 MHA 计算**

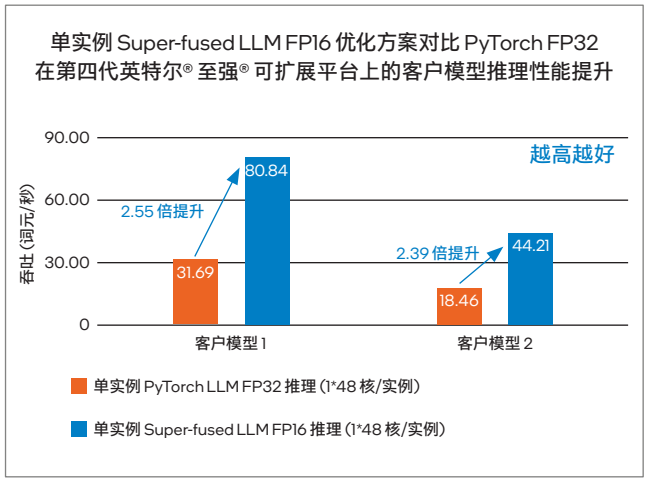
据统计，AI 羊驼 - 阿花方案 LLM 推理中 MHA 计算占比约 25%[5]，包含了大量内存拷贝操作。为此，英特尔开发了 Flash Attention 加速算法，通过算子融合及减少内存操作来降低模型中的 MHA 计算占比，从而提升推理性能；

- **算子融合及计算缓存复用**

传统 PyTorch 推理中，需要借助大量计算缓存来存储模型算子的中间计算结果。Super-fused LLM FP16/AMX BF16 推理优化基于新方案，模型实现了 PyTorch Transformer 融合算子，并按模型推理运行时的输入，更为准确地估算缓存所需空间，从而实现融合算子间缓存复用并提升推理性能。

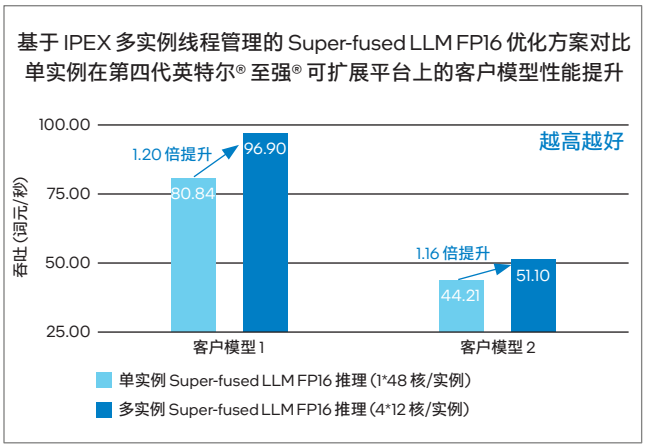
验证成果

为验证优化方案对 LLM 推理任务的性能提升，英特尔与 A-SOUL 团队一起开展了多轮测试验证工作。首先在单实例下，第四代英特尔® 至强® 可扩展处理器上的两种 LLM 推理任务在加入 Super-fused LLM FP16/AMX BF16 优化方案后，如图三所示，每处理器的吞吐量各自从 31.69 和 18.46 上升到 80.84 和 44.21，分别提升达 2.55 倍和 2.39 倍，优化效果非常明显。[6]



图三 单实例下 Super-fused LLM FP16/AMX BF16 优化方案效果对比[7]

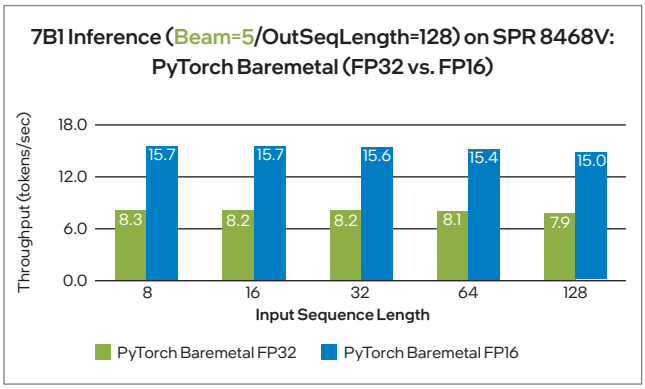
进一步地，双方也验证了在多实例部署的场景下，优化方案对 LLM 推理整体吞吐性能的提升。在多实例部署的场景下，优化方案进一步引入了 IPEX 来优化多实例线程管理。在给定实例数和每实例对应的处理器内核数后，IPEX 可自动根据当前环境绑定线程来运行多实例 LLM 推理任务。如图四所示，例如在一颗英特尔® 至强® 铂金 8457C 处理器 (48 处理器内核) 上，可运行 4 个 LLM 实例 (每个实例使用 12 个处理器内核)，推理吞吐量可从 48 内核单实例的 80.84 词元 / 秒进一步提升至 4 实例共计 96.90 词元 / 秒，提升达 1.2 倍。而在另一个模型的测试中，推理吞吐可从 48 内核单实例的 44.21 词元 / 秒进一步提升至 4 实例共计 51.10 词元 / 秒，提升达 1.16 倍。[8]



图四 Super-fused LLM FP16/AMX BF16 优化方案下单实例 / 多实例性能对比[9]

而在 A-SOUL 团队开展的面向 AI 羊驼 - 阿花方案的实战测试中，上述性能提升也获得了充分验证。A-SOUL 团队分别选择了新方案中某模型的 1B、2B 和 7B 三个参数量进行了测试。其中参数量为 1B 的情况下，推理吞吐性能从 30+ 词元 / 秒提升至 84+ 词元 / 秒，参数量为 2B 时，推理吞吐性能从 15 词元 / 秒提升至 42.6 词元 / 秒，与前述测试结果基本相符。[10]

而在参数量为 7B 的模型上，A-SOUL 团队分别测试了在不同输入序列长度下 (Input Sequence Length) 优化方案带来的性能提升。如图五所示，在未优化的 FP32 精度下，不同输入序列长度下的推理吞吐性能均在 8 词元 / 秒上下，加入优化方案后，FP16 精度下的推理吞吐性能均达到 15 至 15.7 词元 / 秒之间，性能提升达接近 2 倍。[11]



图五 Super-fused LLM FP16/AMX BF16 优化方案效果对比[12]

获益与展望

AI 羊驼 - 阿花形象历经形象设计、模型制作、引擎适配、交互设计等多重流程，会逐渐从后台走进人们的视野，A-SOUL 团队与英特尔在第四代至强® 可扩展处理器上，引入 Super-fused LLM FP16/AMX BF16 推理加速方案所实现的效果也将逐渐崭露头角，为 A-SOUL 团队与英特尔双方都带来巨大获益。

- **性能获益**[13]**：** 在面向 AI 羊驼方案的不同 LLM 上，单实例场景下推理吞吐性能都有了 1.89 至 2.55 倍不等的提升，多实例场景下性能则可以进一步提升至原有的 1.16 至 1.2 倍；
- **成本获益：** 从实际测评数据来看，CPU 平台完全胜任对参数规模为 10B 及以下的 LLM 推理任务，这帮助 A-SOUL 团队能以更低的成本满足方案所需的推理性能要求，达到降本增效的目的。同时，优化后的 CPU 平台在部署运维时，环境的配置安装也更为简单，降低了部署和运维的人力成本；
- **生态获益：** 本次合作引入的 Super-fused LLM FP16/AMX BF16 推理优化方案完全基于 PyTorch 框架开发，完整继承了 AI 羊驼 - 阿花方案中 LLM 的文本生成模块，与 PyTorch 模型推理接口完全一致。相比其它 LLM 加速引擎，使用者无需为调用推理优化方案进行额外的代码开发，更易部署和落地，这为基于 CPU 平台的 LLM 推理提供了良好的实践。

随着 LLM 在 NLP 领域获得越来越广泛的运用，LLM 推理的高效性也为更多人所关注。本次 A-SOUL 团队与英特尔的合作有效验证了第四代英特尔® 至强® 可扩展处理器及其优化方案对 LLM 推理任务的加速能力，并成功应用到了 AI 羊驼方案的实际部署中，这为今后双方更为深入的合作及大范围推广打下了坚实基础。

面向未来，双方还计划在高效微调训练、加速 LoRA/QLoRA 等的训练时间、更多类型的模型结构适配以及更有效的加速算法 (例如更低精度的 FP8 方案) 上开展合作，让 AI 在直播领域的落地变得更为高效，从而为观众带去更为多姿多彩的直播盛宴。

[1, 3, 13] 1-node, 2x 英特尔® 至强® 铂金 8457C 处理器, 48 cores, HT On, Turbo On, Total Memory 1024 GB (16 slots/ 64 GB/ 4800 MT/s [run @ 4800 MT/s]), <American Megatrends International, LLC. 0B.01.02.02.00>, < 0x2b000161 >, < CentOS Steam 8>, <kernel-5.18.0>, <gcc 12>, <pytorch-2.0>, <OneMKL-2023.0.0>;
其中部分数据援引自A-SOUL团队未公开的内部测试，如欲了解更多详情，请联系 A-SOUL 团队。
[2]图片来源于 A-SOUL 团队。
[4, 5, 10, 11, 12] 数据援引自 A-SOUL 团队未公开的内部测试，如欲了解更多详情，请联系 A-SOUL 团队。
[6, 7, 8, 9] 1-node, 2x 英特尔® 至强® 铂金 8457C 处理器, 48 cores, HT On, Turbo On, Total Memory 1024 GB (16 slots/ 64 GB/ 4800 MT/s [run @ 4800 MT/s]), <American Megatrends International, LLC. 0B.01.02.02.00>, < 0x2b000161 >, < CentOS Steam 8>, <kernel-5.18.0>, <gcc 12>, <pytorch-2.0>, <OneMKL-2023.0.0>;