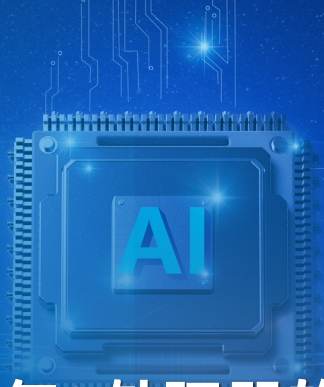
解决方案简介



携程采用基于英特尔® 处理器的 AI 推理优化方案提供高性能、经济的 AI 服务能力

携程旅行

“作为一家技术驱动的互联网公司，携程将人工智能技术作为企业持续发展的强大驱动力和竞争力，并为消费者提供更为优质、快捷的服务。通过与英特尔合作验证基于 CPU 的 AI 推理优化方案，我们找到了一条更具经济性、灵活性、稳定性，且全面易用、使用和维护成本更低的 AI 优化之路，能够支撑我们更深入地推动业务的智能化创新，赋能在线旅行生态。”

— 韩宝龙

携程大数据与 AI 应用研发部

概述

“人工智能（AI）+ 旅行”的融合正在成为旅游行业发展的重要趋势。得益于自然语言处理、机器翻译、计算机视觉、搜索排序等 AI 技术的广泛应用，在线旅行服务提供商已经能够根据用户偏好，为用户提供个性化的内容与服务推荐，改善服务的质量与效率，加速商业价值的挖掘。但同时，AI 模型训练、推理等过程需要规模庞大的算力基础设施，并带来了较高的总体拥有成本（TCO）压力。

作为一站式旅行平台，携程近年来加大了在 AI 创新方面的投资，将 AI 技术成功应用于酒店、机票、自由行、跟团游、签证、玩乐、租车等旅游度假的多个业务线，为全球用户提供一套完整的旅行产品、服务及差异化的旅行内容。为了在满足业务需求的同时降低成本压力，携程构建了基于英特尔® 至强® 可扩展处理器的 AI 推理算力平台，并通过高性能算子库、计算图优化、模型压缩、模型部署优化等方式，提升了 AI 推理性能。

挑战：化解 AI 算力瓶颈需要深度的 AI 推理性能优化

AI 应用的发展凸显了企业在算力方面的挑战。一方面，AI 技术正在日趋多样化与复杂化，为了适应不同的业务场景需求，企业常常需要融合使用传统机器学习、卷积神经网络、Transformer 等深度学习模型结构，以及知识图谱、图神经网络等技术。同时，AI 模型的深度、宽度以及结构复杂度也在不断提升，增加了企业的开发门槛，也使得 AI 算力调度、AI 性能优化更具挑战。

另一方面，需要由 AI 赋能的智能化应用正在迅速扩张，需要由 AI 模型处理的数据也在不断增长。在旅行服务行业，内容与广告个性化推荐、实时风控、机器翻译、智能客服、图像处理等领域正在越来越多地使用 AI 技术，以从海量的数据中生成高价值的商业洞察，从而带来了较高的算力基础设施建设成本。与此同时，上层应用对于 AI 模型推理也有着特定的服务级别协议（SLA）要求。企业需要在满足 SLA 要求的前提下，通过模型优化等方式，更好地发挥硬件的性能潜力，降低算力基础设施的 TCO。

要满足上述需求，企业首先要面临硬件平台的选择问题：虽然基于独立 GPU 的推理方案能够提供强大的算力，但未必是一个经济的选择。这是因为独立 GPU 不仅采购或租用成本相对较高，而且通常是以专用服务器的模式进行部署和运维，带来了较高的综合成本。考虑到旅行服务中大量的 AI 推理场景所需要的性能经过优化能够得到满足，采用 CPU 的方案将更具成本效益。

除了硬件平台选择之外，企业在 AI 模型推理性能优化方面也面临着以下瓶颈：

- 模型结构种类多，性能瓶颈差异较大，适用的优化方法各有不同，手动优化成本高、门槛高；
- 传统方式需要对模型进行逐个手动优化，可推广性差，技术覆盖面有限；
- AI 推理面向的硬件平台存在广泛差异，技术人员往往需要进行针对性调优，引发较高的人力成本和部署成本；
- 新模型的发布和迭代需要应用优化方法，涉及较高的沟通和接入成本，同时带来了性能的不稳定性；
- 模型压缩技术对不同模型的优化效果有所差异，可能需要进行模型的再训练，训练和数据准备流程较长，效率低下。

解决方案：基于英特尔® 至强® 可扩展处理器的携程 AI 推理算力平台

为了降低优化、部署和迭代成本，提高工作效率，并实现稳定性能，携程尝试评估基于英特尔® 至强® 可扩展处理器的 AI 推理算力平台，旨在为算法模型提供更全面易用、稳定性更好、使用和维护成本更低的优化解决方案。

英特尔® 至强® 可扩展处理器内置人工智能加速功能，并已针对工作负载进行优化，能够为各种高性能计算工作负载、AI 应用以及高密度基础设施带来一流的性能和内存带宽。同时，采用矢量神经网络指令（VNNI）的英特尔® 深度学习加速（英特尔® DL Boost）能够有效提高 AI 推理的表现，这使其成为进行深度学习应用的卓越基础设施。

在基于英特尔® 至强® 可扩展处理器的硬件平台层基础上，携程构建了 AI 推理算力平台，该平台还包括引擎框架层、推理优化层、算法模型、应用场景。

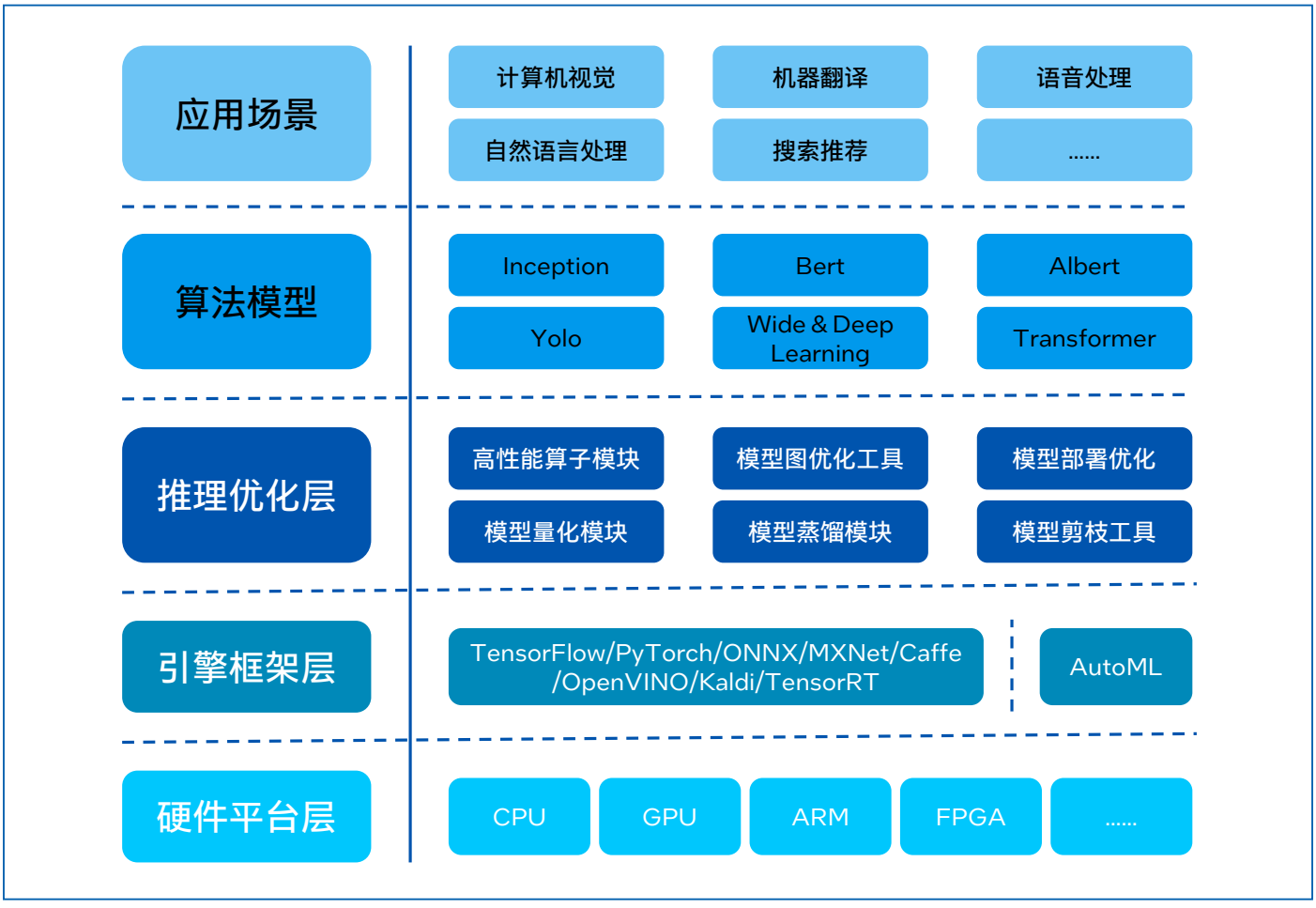


图 1. 携程 AI 推理算力平台架构

引擎框架层

主要涵盖 Tensorflow、Pytorch 等人工智能主流框架。

推理优化层

主要涵盖携程结合业务场景和实际需求进行自主研发的优化技术，包含高性能算子库、图优化和修改工具以及量化蒸馏等模型压缩模块。

算法模型

包含业界常用模型，例如 Resnet、GoogleNet、YOLO 等以卷积为主要结构的 CV 模型，Bert、Albert 等以 Transformer 为主要结构的 NLP 模型。

应用场景

主要体现在旅游场景中的实际应用，例如智能客服平台、机器翻译、搜索排序等应用。

携程 AI 模型推理性能优化实践

为了尽可能地提升 AI 推理性能，释放硬件潜力，携程进行了推理优化。主要的优化思路为两点：一是通过调整/简化模型结构，或改进算法以降低算法复杂度；二是优化软件执行效率，使用硬件优势特征，提升硬件执行效率。

携程的 AI 模型推理性能优化流程如图 2 所示，训练的原始模型依次通过高性能算子模块、计算图优化模块、模型压缩模块、模型部署优化，以实现推理性能。

- **高性能算子库**

  该模块主要实现了常用的算子以及激活函数，包含卷积、全连接层、batch norm、softmax 等基础算子，以及 Transformer encoder、Decoder 等合并后的经典的模型结构。该高性能算子库基于 Tensorflow 实现，采用 C++ 实现，支持 CPU 平台的优化。

- **计算图优化**

  主要进行计算图搜索，修改替换模型图结构，合并生成新的模型文件进行推理部署，同时包含常用的图优化和修改工具。

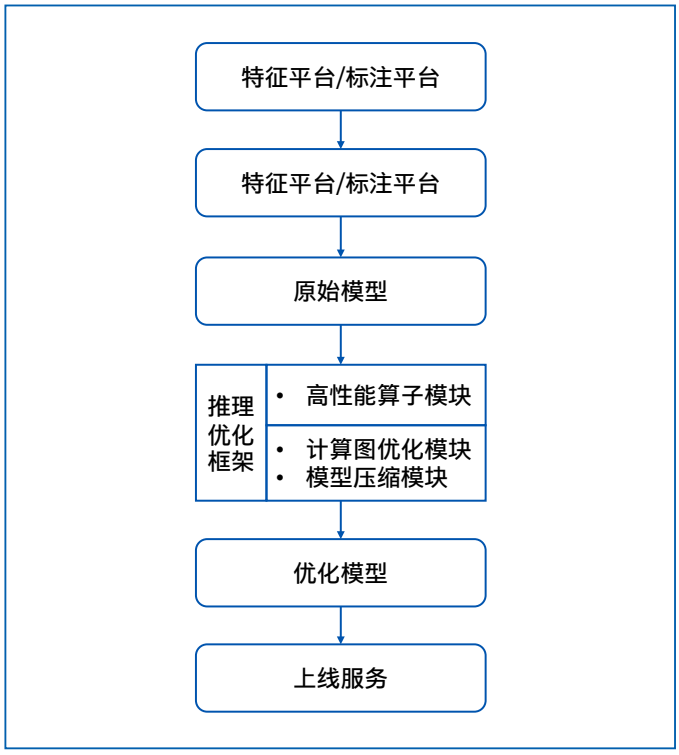


图 2. 携程 AI 模型推理性能优化流程图

验证：新一代英特尔® 至强® 可扩展处理器显著加速 AI 模型推理性能

为了验证优化的性能表现，携程以机器翻译应用的 Transformer 模型为例进行了验证。在该测试中，携程除了使用上述的优化方法，还分割模型并使用 jit 跟踪方法来提高性能。同时，在 batchsize 超过 16 个的情况下，使用 mm op 进行优化。

在该测试中，携程使用固定算例的平均响应时延作为测试数据，tokens 为 10，batchsize=1，优化后和优化前的性能对比如下图 4 所示。

其中，原始性能是基于 Tensorflow1.14 测试而来的基准数据，在 CPU 平台框架层优化和编译运行时等多层优化实现，图 5 是 Transformer 翻译模型基于 T5 平台使用模型压缩和高性能算子库优化之后的对比结果，图中给出的是 token 最大长度为 80 时，不同 batchsize 大小的吞吐量性能优化结果。

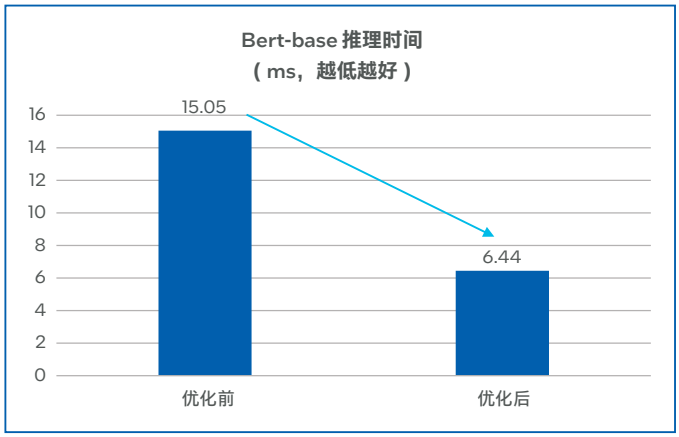


图 3. Bert-base 模型优化前后的性能对比[1]

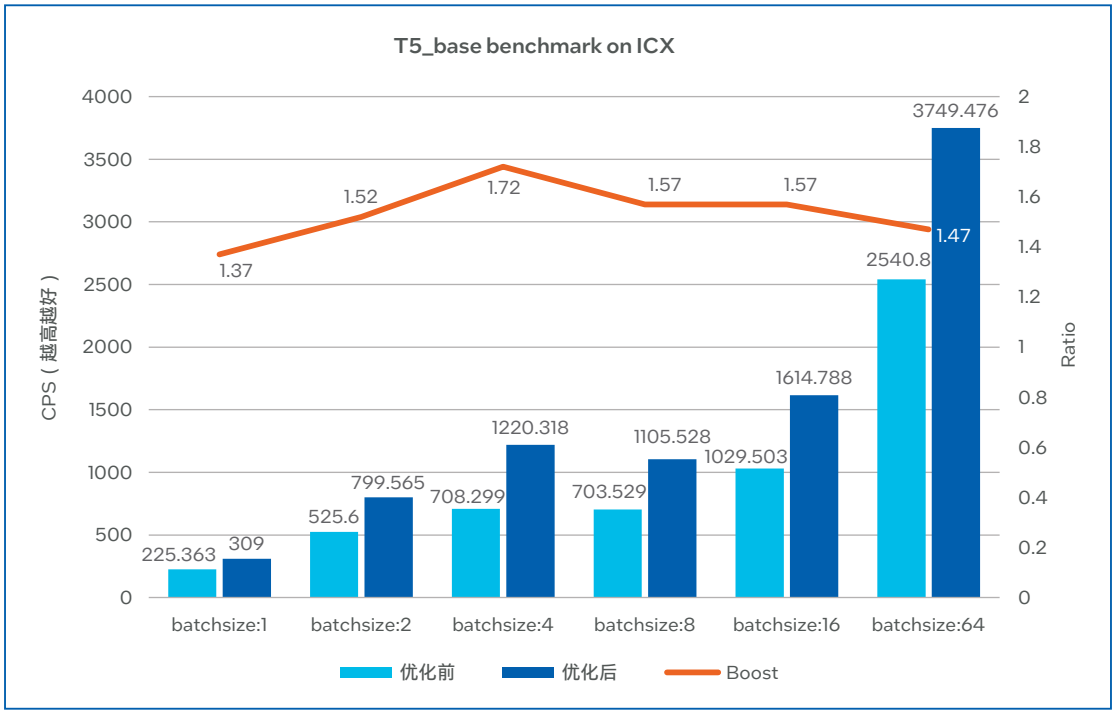


图 4. Transformer 模型基于 T5 平台使用模型压缩和高性能算子库优化前后的对比结果[2]

[1] 数据援引自携程于 2022 年 9 月开展的测试。测试配置：双路英特尔® 至强® 金牌 5318Y 处理器 @ 2.1 GHz，24 核，512 GB 总内存，CentOS 8，Kernel 4.18.0-348。英特尔并不控制或审计第三方数据。请您审查该内容，咨询其他来源，并确认提及数据是否准确。

[2] 数据援引自携程于 2022 年 9 月开展的测试。测试配置：双路英特尔® 至强® 铂金 8358 处理器 @ 2.6 GHz，32 核，512 GB 总内存，CentOS 8，Kernel 4.18.0-348。英特尔并不控制或审计第三方数据。请您审查该内容，咨询其他来源，并确认提及数据是否准确。

收益：实现性能与经济性的更佳平衡

得益于深度的 AI 模型推理性能优化，以及英特尔® 至强® 可扩展平台的基础算力，携程实现了预期的性能优化效果。在这一方案落地之后，预计将会为携程带来如下收益：

在特定的 SLA 要求下，降低 AI 推理应用的 TCO

在通过本轮优化之后，携程的 CPU 服务器 AI 推理性能得到提升，能够满足大量 AI 推理场景对于时延等 SLA 指标的要求，避免了在昂贵的专用 AI 加速器方面的支出。

提升基础设施的敏捷性与灵活性

通过本轮优化，携程能够高效利用现有的 CPU 服务器，根据实际负载需求进行灵活调度，而无需为 AI 推理新增需求部署专用服务器。

为 AI 推理性能优化提供了标准的参考流程

本方案提供了一套标准、可自动完成的参考模型优化流程，能够赋能更多的 AI 应用。

展望

在数字经济快速发展的背景下，随着 AI 理论和模型的日益完善，应用场景对模型精度等推理服务质量的要求不断增加，模型结构和计算复杂度将会越来越高。与此同时，数据的爆发式增长以及用户对 AI 推理服务的性能需求将导致企业 AI 推理算力供给与需求之间的瓶颈不断提升。从成本和效率多个角度考虑，强化 AI 推理性能优化都将是重要的发展趋势。

业界开始探索以基于 Transformer 的统一多模态（视觉、NLP、音频等）的模型结构为基座，并基于通用基础模型库向多样的下游任务进行精调的实践范式，这一范式由于在成本和效率之间得到了很好的平衡，得到越来越多的业界实践。

携程与英特尔合作验证了 AI 模型推理方案在英特尔® 至强® 可扩展处理器上的应用潜力，在成本、性能方面实现了平衡。双方还计划进一步开展合作，包括探索对采用第四代英特尔® 至强® 可扩展处理器集成的英特尔® AMX 进行深度优化等，以进一步释放硬件在 AI 应用中的潜力。

关于携程

携程集团（Trip.com Group）是全球领先的一站式旅行平台，公司旗下的平台可面向全球用户提供一套完整的旅行产品、服务及差异化的旅行内容。集团能够提供超过 120 万种全球住宿服务，480 多家国际航空公司的服务，以及超过 31 万项目的地内活动。同时，集团与超过 3 万家其他合作伙伴一起，致力于满足客户不断变化的需求。

关于英特尔

英特尔（NASDAQ: INTC）作为行业引领者，创造改变世界的技术，推动全球进步并让生活丰富多彩。在摩尔定律的启迪下，我们不断致力于推进半导体设计与制造，帮助我们的客户应对最重大的挑战。通过将智能融入云、网络、边缘和各种计算设备，我们释放数据潜能，助力商业和社会变得更美好。如需了解英特尔创新的更多信息，请访问英特尔中国新闻中心 newsroom.intel.cn 以及官方网站 intel.cn。

intel

实际性能受使用情况、配置和其他因素的差异影响。更多信息请见 www.Intel.com/PerformanceIndex

性能测试结果基于配置信息中显示的日期进行测试，且可能并未反映所有公开可用的安全更新。详情请参阅配置信息披露。没有任何产品或组件是绝对安全的。

具体成本和结果可能不同。

英特尔技术可能需要启用硬件、软件或激活服务。

英特尔未做出任何明示和默示的保证，包括但不限于，关于适销性、适合特定目的及不侵权的默示保证，以及在履约过程、交易过程或贸易惯例中引起的任何保证。

英特尔并不控制或审计第三方数据。请您审查该内容，咨询其他来源，并确认提及数据是否准确。