解决方案简介

英特尔® 至强® 可扩展处理器
英特尔® OpenVINO™ 工具套件分发版

英特尔联手中兴优化深度学习模型推理 实现降本增效

“近年来，AI 技术作为数字经济的核心驱动力，已成为各行业智能化转型的关键要素。Adlik 是由中兴通讯主导，在 Linux AI & Data 基金会中孵化的一项开源项目，旨在解决在 AI 应用落地过程中的各类挑战性问题，实现模型在特定硬件环境的快速部署、高效推理。中兴通讯使用 Adlik 模型优化、服务部署组件，与 OpenVINO™ 工具套件相结合的方案，在各类产品设备上实现了云边端多场景的 AI 应用部署，大幅提升了部署效率和硬件执行效率，创造了巨大的商业价值。在未来，中兴通讯将与合作伙伴一道，共同推动 AI 应用技术发展和创新，为数字经济提供坚实支撑。”

— 韩炳涛
中兴通讯 AI 平台总工

概述

伴随着数字化技术的持续创新以及数字经济的蓬勃发展，以深度学习（DL）为代表的人工智能（AI）技术在近年来实现了快速的场景化落地，也带来了巨大的算力需求。研究数据显示，随着大规模机器学习模型的出现，训练算力的需求急剧增加。此外，模型推理、数据处理等环节也带来了大量的算力消耗，用户需要从端到端、全流程的角度优化深度学习系统的设计，为企业智能化升级和数字经济提供强劲动力。

为了加速端到端的深度学习模型优化和部署，中兴推出了深度学习模型推理应用加速工具链 Adlik，能够加速深度学习模型从研发到上市的进程。Adlik 集成英特尔® OpenVINO™ 工具套件分发版（OpenVINO™ 工具套件），可以在基于第三代英特尔® 至强® 可扩展处理器的服务器上实现出色运行，提升模型计算效率，减少能耗，降低推理时延，助力在云边端等多环境的模型部署，满足不同 AI 工作负载高效运行的需求。

挑战：深度学习模型亟待性能深度优化

作为 AI 技术重要的分支之一，深度学习在近年来得到了广泛的应用，并在语音识别、计算机视觉和自然语言处理等任务中取得了巨大的成功，推动了各行各业的智慧化转型进程，基于深度学习框架延伸、构建智能生态平台成为组织的重要选择。在通信领域，深度学习技术已在流量预测、KPI 预测、故障诊断、根因分析、网络优化、加密流量分类等场景下得到广泛应用。

但同时，深度学习模型在落地时往往会遇到多重挑战，这些挑战包括模型部署、模型优化、硬件选择等各个方面。

● 如何高效部署模型

要部署深度学习模型，需要充分考虑到深度学习训练框架的差异，以及推理框架的差异。目前存在多种深度学习训练框架，导致不同框架训练后保存的模型格式均不相同，这一点也增加了模型部署的困难。同时，不同的芯片厂商为了在各自的芯片上取得出色的性能，都会提供自有的推理框架，不同的推理框架提供的 API 完全不同，且互不兼容，开发者想要实现异构芯片下的模型部署将变得非常困难。

● 如何优化模型，确保模型以卓越的性能运行

深度学习模型在落地时往往会遇到推理性能问题，例如计算时延高、吞吐量低，内存占用大等。在不同的应用场景和部署环境下，对于模型的优化目标不完全相同。例如，在端侧部署中，内存和存储空间均非常有限，因此模型优化目标是减小模型的大小；在自动驾驶场景下，计算平台算力有限，对模型的优化则侧重于在有限的算力下，尽可能提升吞吐量、降低时延。因此，要想满足在不同场景和部署环境下模型对推理性能的要求，就需要合适的模型优化工具以及优化策略。

● 如何选择合适的硬件平台

在实际业务场景中，既有对算力要求较高的深度学习模型，又有传统的机器学习模型，同时还需要能够支持大数据、云计算和虚拟化等多种业务的扩展。这会带来硬件平台选型方面的困扰：GPU 性能通常能够满足深度学习推理的需求，但是价格较高，需购买专用 GPU 硬件，其不仅会大幅增加部署成本，而且应用范围有限，灵活度较低。在部分场景中，如果能够直接使用 CPU 来进行推理，将有助于降低成本，提升灵活度，但这也依赖于硬件创新，以及软件层面的深度优化。

解决方案：中兴 Adlik + 第三代英特尔® 至强® 可扩展处理器 + OpenVINO™ 工具套件助力深度学习端到端优化与部署

为了满足通信领域不同的业务场景和复杂的 AI 工作负载，以及在不同的应用场景和部署环境下对模型推理性能的要求，中兴与英特尔合作，将 Adlik 推理工具链与第三代英特尔® 至强® 可扩展处理器和 OpenVINO™ 工具套件结合，打造了端到端的深度学习模型优化和部署方案。

Adlik 推理工具链

Adlik 是用于将深度学习模型从训练完成到部署到特定硬件，提供应用服务的端到端工具链，其应用目的是为了将模型从研发产品快速部署到生产应用环境。Adlik 可以和多种推理引擎协作，支持多款硬件，提供统一对外推理接口，并提供多种灵活的部署方案。

Adlik 整体结构如图 1 所示，训练好的模型可通过 Adlik 模型优化器进行优化，随后通过模型编译器完成模型格式转换，最终生成推理引擎支持的模型格式，部署到相应硬件平台上。

Adlik 能够为用户带来如下优势：

- 多种模型压缩、优化算法在实践中表现出出色性能：面对异构的部署硬件，提供更优的端到端方案；根据不同的应用场景，实现更优的推理性能（如时延、吞吐量等）和更优的模型管理。
- 内置多种高性能运行时，以供用户更快地按需选用；提供可拓展性强的 Serving SDK，可更快集成自定义推理运行时；提供灵活易用的推理 API，更快实现 AI 应用的构建、迭代。
- 简单方便的模型部署 pipeline，缩短模型上线周期，节省部署时间；统一的模型推理和管理接口，节省模型迁移成本。

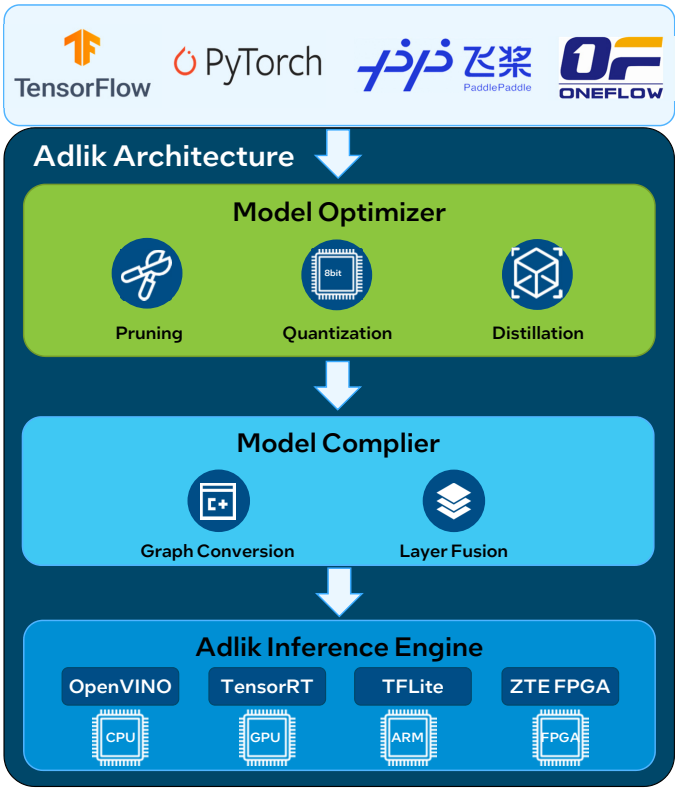

图 1. Adlik 整体结构

第三代英特尔® 至强® 可扩展处理器

第三代英特尔® 至强® 可扩展处理器拥有 8 至 40 个强大的内核，并提供多种频率、功能和功耗水平选项。与第二代英特尔® 至强® 可扩展处理器相比，性能、吞吐量和 CPU 频率都实现了显著提高，这为其处理 AI 推理负载提供了关键的性能基础。尤为重要的是，该处理器集成了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），能够充分提高计算资源和缓存的利用率、减少潜在的带宽资源瓶颈。多达六条英特尔® UltraPath Interconnect（英特尔® UPI）通道提高了平台的可扩展性，并改善了 I/O 密集型工作负载的 CPU 间带宽，在提高吞吐量和能效之间实现了灵活平衡。

处理器集成的英特尔® 深度学习加速技术能够在不改变现有硬件的前提下，提供足以运行复杂人工智能工作负载的灵活性。借助第三代英特尔® 至强® 可扩展处理器支持的 INT8，矢量神经网络指令能够通过充分利用计算资源、提高缓存利用率和减少潜在的带宽瓶颈来增强推理工作负载。同时，借助部分第三代英特尔® 至强® 可扩展处理器上提供的 16 位脑浮点（BF16）支持，能够进一步增强人工智能推理和训练性能。

作为 Adlik 推理工具链的关键组件，Adlik 模型优化器专注于结构化剪枝和 INT8 量化技术，其主要包括 3 个算法组件：剪枝、蒸馏和量化。Adlik 模型优化器支持多种深度学习框架训练的模型，训练完成的模型经过剪枝、蒸馏和 INT8 量化后，可通过模型编译器将其编译为不同推理引擎上的中间格式（IR）模型（如 OpenVINO™ IR 模型）。

Adlik 模型优化器有助于减小模型大小，在加速模型推理的同时，提供高准确率的模型，为开发者提供灵活高效的深度学习模型优化方案。

Adlik 模型优化器支持多种结构化剪枝方法，并可通过多节点，多 GPU 进行剪枝。通过使用剪枝方法，Adlik 模型优化器能够有效降低模型参数量和浮点运算次数（FLOPs）。为了解决传统的剪枝方法需要人工评估模型每一层敏感度，手动设置剪枝层以及剪枝层类型的痛点，降低使用剪枝技术的门槛和学习成本，Adlik 模型优化器支持自动剪枝方法：用户只需要指定网络类型（如 ResNet-50 等）和限制条件（如 FLOPs，Latency），Adlik 模型优化器就能自动决定模型每一层的通道数，得到在限制条件下最优的模型结构。

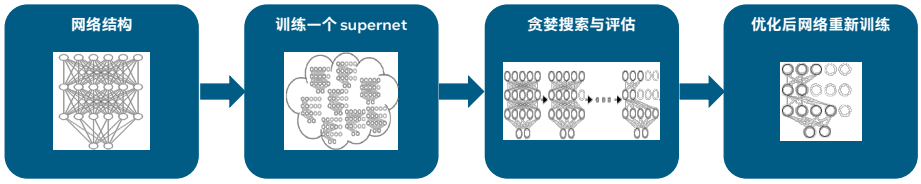

图 2. Adlik 模型优化器可实现自动剪枝

知识蒸馏是将已训练完成具有较高精度大模型的“知识”迁移到参数量小、结构相对简单的小模型中，可以使得小模型具有与大模型相当或接近的性能。通过剪枝和量化的方式获取的小模型，相对于大模型的参数量和 FLOPs 大幅降低，从而实现模型压缩与加速，但模型精度不免下降。Adlik 模型优化器采用知识蒸馏的方法，用适当的大模型去蒸馏小模型，提升小模型的精度，在加速推理的同时，也能达到业务所需推理精度要求。

Adlik 模型优化器提供不同的蒸馏方法，能够应用于各种深度学习任务（如图像分类，目标检测等）。为了使用户更易于将知识蒸馏方法应用于不同的深度学习模型，Adlik 模型优化器使用基于输出响应的蒸馏方法，并对 YOLO 系列模型知识蒸馏方法进行了特别的优化。

Adlik + 第三代英特尔® 至强® 可扩展处理器 + OpenVINO™ 工具套件 + 实现端到端模型推理性能优化

如今，用户已经可以通过 Adlik+ 第三代英特尔® 至强® 可扩展处理器 + OpenVINO™ 工具套件，构建端到端的深度学习应用流程，提升推理性能优化效果，在 CPU 上实现高效的深度学习模型推理。

在典型的应用流程中，用户可以将训练好的模型输入到 Adlik 推理工作链中，通过 Adlik 模型优化器进行优化，随后通过模型编译器完成模型格式转换，最终生成 OpenVINO™ 工具套件所支持的 IR 模型格式，部署到基于第三代英特尔® 至强® 可扩展处理器的服务器上，实现端到端的性能优化。

OpenVINO™ 工具套件的 Post-Training Optimization Toolkit（POT）优化包含以下几个要素：

- 一个能在 CPU 上运行推理程序的 OpenVINO FP32/FP16 IR（Intermediate Representation）模型
- 有代表性场景的数据做为标定数据集（根据精度要求调整数据集大小）
- 精度校验所需的验证数据集和精度评价指标（Metric，例如，分类模型常用 TopK）

为了满足严格的精度要求，POT 提供了两种量化的算法：缺省量化 Default Quantization（DQ）和精度感知量化 Accuracy-aware Quantization（AAQ）。Default Quantization（DQ）提供了一种快速的量化的方法，尽可能将所有成为计算热点的层进行量化，量化后的模型推理性能最优，适合作为模型 INT8 量化的基准。在缺省量化之上，精度感知量化可调节预期精度下降范围，获得更高精度的量化模型。

POT 提供了以下两种使用方式，即命令行调用和 API 编程。命令行调用通过命令行运行相应配置文件来调用预定义的 Adapter，Pre/Postprocessing，Metric 等模块，这种方式适用于 Open Model Zoo 支持的标准模型的 INT8 量化（如 Resnet50）。无需编写代码就可快速量化。POT API 编程提供了数据加载、预/后处理、评价指标等基类模板，用户可以客制化重写以上功能模块，可以更加灵活地使用量化工具。

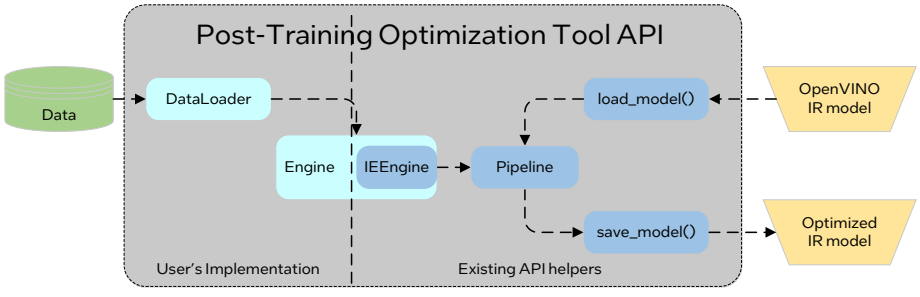

图 3. 基于 POT API 进行 INT8 量化的通用流程

OpenVINO™ 工具套件

OpenVINO™ 工具套件是一个由英特尔开源的工具包，用于优化和部署 AI 推理，支持包括英特尔 CPU、GPU（包括独立显卡和集成显卡）以及 VPU 在内的多个硬件平台的部署。

为了充分利用英特尔相关硬件对深度学习的加速，提高模型推理性能，OpenVINO™ 提供低精度量化算法来进一步优化模型的工具 — 量化工具 Post-Training Optimization Toolkit（POT）和面向专用硬件加速的 Runtime。POT 量化工具将模型的权重和激活函数从 FP32 的值域映射到 INT8 的值域中，从而实现模型压缩，以降低模型推理所需的计算资源和内存带宽，提高模型的推理性能。不同于 Quantization-aware Training（QAT），POT 在不需要对原模型进行 fine tuning 的情况下进行量化，也能得到精度较好的 INT8 模型，因此广泛地被应用于工业界的量化实践中。

图 4. 端到端模型推理性能优化与部署流程

Adlik + 第三代英特尔® 至强® 可扩展处理器 + OpenVINO™ 工具套件的结合可实现模型推理的性能的显著提升。为了验证上述联合方案带来的性能提升，双方以不同场景中广泛使用的图像分类模型 ResNet50 和目标检测模型 YOLOv5 为例，进行了测试。参测的推理服务器为 ZTE 5300G4X，该服务器采用第三代英特尔® 至强® 铂金 8378 处理器，服务器详细配置如表 1 所示。吞吐量使用 OpenVINO™ 工具套件的 benchmark 测试工具，测试使用 38 个 CPU 内核。

表 1. 测试配置

	ZTE 5300 G4X
CPU	2*英特尔® 至强® 铂金 8378C CPU @ 2.80GHz
内核	2*38
内存	1024 GB 总内存 (16*64 GB DDR4 2933 MT/s)
硬盘	3* 3.5 TB
网卡	2*英特尔® 以太网连接 I210
软件	OpenVINO™ 2022.1 Adlik v0.5.0

测试效果如图 5 所示，在 ImageNet val 验证数据集上，ResNet50 剪枝模型经过蒸馏后精度略有提升。剪枝模型的吞吐量比原始模型提升了 2.74 倍[1]。INT8 量化后的模型的吞吐量比未量化模型提升了 2.96 倍[2]。经过 Adlik 剪枝蒸馏和 INT8 量化等方法优化后的 ResNet50 模型，在精度无损失的情况下，吞吐量比原始模型提升了 13.82 倍[3]。

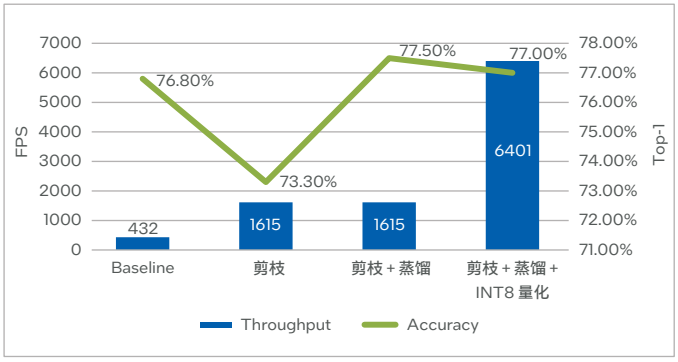

图 5. 图像分类 ResNet50 模型优化测试结果[4]

目标检测 YOLOv5m 模型优化测试结果如图 6 所示，在 COCO2017 验证集上，YOLOv5m 经剪枝蒸馏和 INT8 量化后的模型，精度损失在 1% 以内[5]。优化后的 YOLOv5m 模型吞吐量比原始模型提升了 3.39 倍[6]。

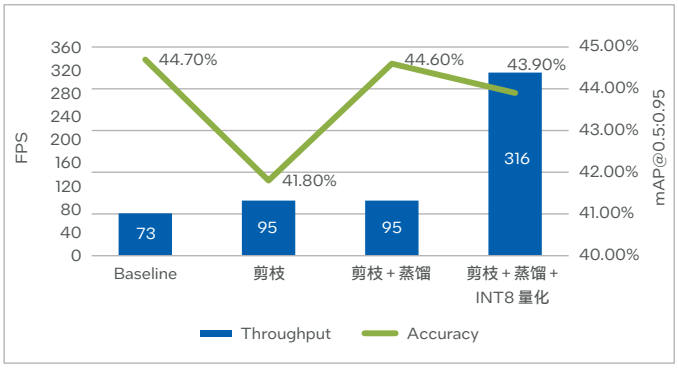

图 6. 目标检测 YOLOv5m 模型优化测试结果[7]

[1,2,3,4,5,6,7] 数据援引自中兴通讯内部测试结果。测试配置：2*英特尔® 至强® 铂金 8378C CPU @ 2.80GHz，2*38 内核，1024 GB 总内存 (16*64 GB DDR4 2933 MT/s)，3* 3.5 TB 硬盘，2*英特尔® 以太网连接 I210，OpenVINO™ 2022.1，Adlik v0.5.0。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。

以上两个典型的深度学习模型测试结果表明，Adlik + OpenVINO™ 工具套件优化能够显著提升深度学习模型推理性能。在第三代英特尔® 至强® 可扩展处理器强大 AI 算力的基础上，可以为多种场景下的深度学习应用提供强大的性能表现。

收益：助力行业智慧化变革

通过部署 Adlik + OpenVINO™ 工具套件 + 第三代英特尔® 至强® 可扩展处理器融合方案，用户能够实现端到端的深度学习模型推理性能优化，加速深度学习应用进程。具体来说，该方案为用户带来了如下价值：

显著提升使用 CPU 进行深度学习推理的性能

方案通过自动剪枝、蒸馏、量化等优化手段的应用，显著提升了深度学习推理在 CPU 上的性能表现，能够满足不同场景的深度学习应用对于性能的需求。

有助于控制深度学习系统的总体拥有成本（TCO）

方案能够充分挖掘第三代英特尔® 至强® 可扩展处理器的性能潜力，降低基础设施构建所带来的硬件采购、运维、能源耗费等成本。同时，这一方案支持用户高效利用现有的 CPU 资源，而无需采购专用加速器。

降低跨架构部署与运行复杂度

Adlik 集成 Intel OpenVINO™ 工具套件，提供“一次编写，随处部署”特性，让 AI 应用除了在 CPU 上运行之外，还能够将转换后的模型运行在不同的英特尔® 硬件平台上，显著简化了构建与迁移过程。在 Adlik 0.4.0 版本上，Adlik + OpenVINO™ 工具套件发布了在基于英特尔® 酷睿™ i5 处理器的设备上的测试结果，验证了 AI 应用在英特尔® CPU 和 iGPU 部署间可以做到无缝切换，显著提升部署效率。

目前，该端到端的深度学习模型优化和部署方案，已随中兴通讯的各类产品在电信领域“规划、建设、维护、优化、营销”等业务场景中广泛使用，优化后的模型能够有效提升运营商 CPU 推理服务器资源利用率，支撑运营商全面实现自智网络，同时助力运营商的节能减排。

展望

AI 已经成为推动数字化变革的关键技术，为运营商等各个行业提供了强大的智慧能力，AI 的创新有赖于 AI 生态的持续构建，以及软硬件的高效融合。Adlik + 第三代英特尔® 至强® 可扩展处理器 + OpenVINO™ 工具套件融合方案提供了基于 CPU 的深度学习推理优化方案，能够有效应对复杂的深度学习算法所带来的挑战，并降低用户在性能优化方面的时间成本与技术门槛。

未来，中兴与英特尔将进一步围绕深度学习端到端性能优化、AI 应用在异构平台上的部署与运行等方向，进行深度合作，帮助用户打通深度学习应用的全流程，真正实现高效率、低成本的 AI 应用落地，助力不同行业实现智慧化转型，为数字经济发展提供强劲动力。

关于中兴

中兴通讯股份有限公司，是全球领先的综合通信解决方案提供商，中国领先的通信设备上市公司。主要产品包括：2G/3G/4G/5G 无线基站与核心网、IMS、固网接入与承载、光网络、芯片、高端路由器、智能交换机、政企网、大数据、云计算、数据中心、手机及家庭终端、智慧城市、ICT 业务，以及航空、铁路与城市轨道交通信号传输设备。

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 ZTE中兴