# 中兴 Adlik 与第三代英特尔®至强®可扩展处理器助力深度学习模型推理降本增效

## 引言



Adlik是人工智能领域深度学习模型推理应用加速工具链，致力于为开发者提供端到端的深度学习模型优化和部署方案，加速人工智能技术工程化落地。第三代英特尔®至强®可扩展处理器能够加快对人工智能、数据分析、高性能计算等多种复杂工作负载的开发和部署，助力不同行业实现 AI 赋能和数字化转型，为企业智能化升级和数字经济提供强劲动力。

为了提升模型计算效率，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 工作负载的高效运行，Adlik结合第三代英特尔®至强®可扩展处理器为产业开发者提供端到端的深度学习模型优化和部署方案，为企业深度学习技术落地应用降本增效。

## 背景和挑战

深度学习技术已在人工智能的任务中取得了巨大的成功，如语音识别、计算机视觉和自然语言处理等任务，同时各行各业也在深度学习技术的赋能下向智能化转型。如在通信领域，AI技术已在流量预测、KPI预测、故障诊断、根因分析、网络优化等场景下广泛落地。深度学习模型在落地时往往会遇到各方面的挑战。

如何高效部署模型？有两点需要考虑：一是深度学习训练框架的差异，二是推理框架的差异。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实现在自家芯片上取得最优的性能，都会提供各自的推理框架，不同的推理框架其提供的 API都完全不同，且互不兼容，当开发者想要实现异构芯片下的模型部署将变得极其困难。

如何确保模型以最优的性能运行？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是不完全相同的。例如，在端侧部署，内存和存储空间都是非常有限的，因此模型优化目标是减小模型的大小。如在自动驾驶场景下，其计算平台算力有限，对模型的优化需达到高吞吐量和低时延的要求。因此，如何满足在不同场景和部署环境下模型对推理性能的要求，对于模型优化工具的功能和特性有较高的要求。

如何选择合适的硬件部署平台？在通信领域不同的业务场景中，即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。如果在 GPU 上部署，则需购买专用 GPU硬件，其不仅会极大增加部署成本，而且应用范围有限，灵活度较低。在一般 CPU 上部署可能无法满足推理性能的要求。如何满足不同 AI 工作负载，同时减低降低部署门槛和成本，选择合适的硬件平台是至关重要的。

## 端到端的深度学习模型优化和部署方案：Adlik 推理工具链 + 第三代英特尔®至强®可扩展处理器

为了能够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，Adlik 推理工具链结合第三代英特尔®至强®可扩展处理器可实现端到端的深度学习模型优化和部署的方案。

### 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给用户带来的价值，可以从三个方面来考量：

1. 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。
2. 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。
3. 简单方便的模型部署pipeline，缩短模型上线周期，更省部署时间；统一的模型推理和管理接口，更省模型迁移成本。

Adlik整体架构如图 1 所示：

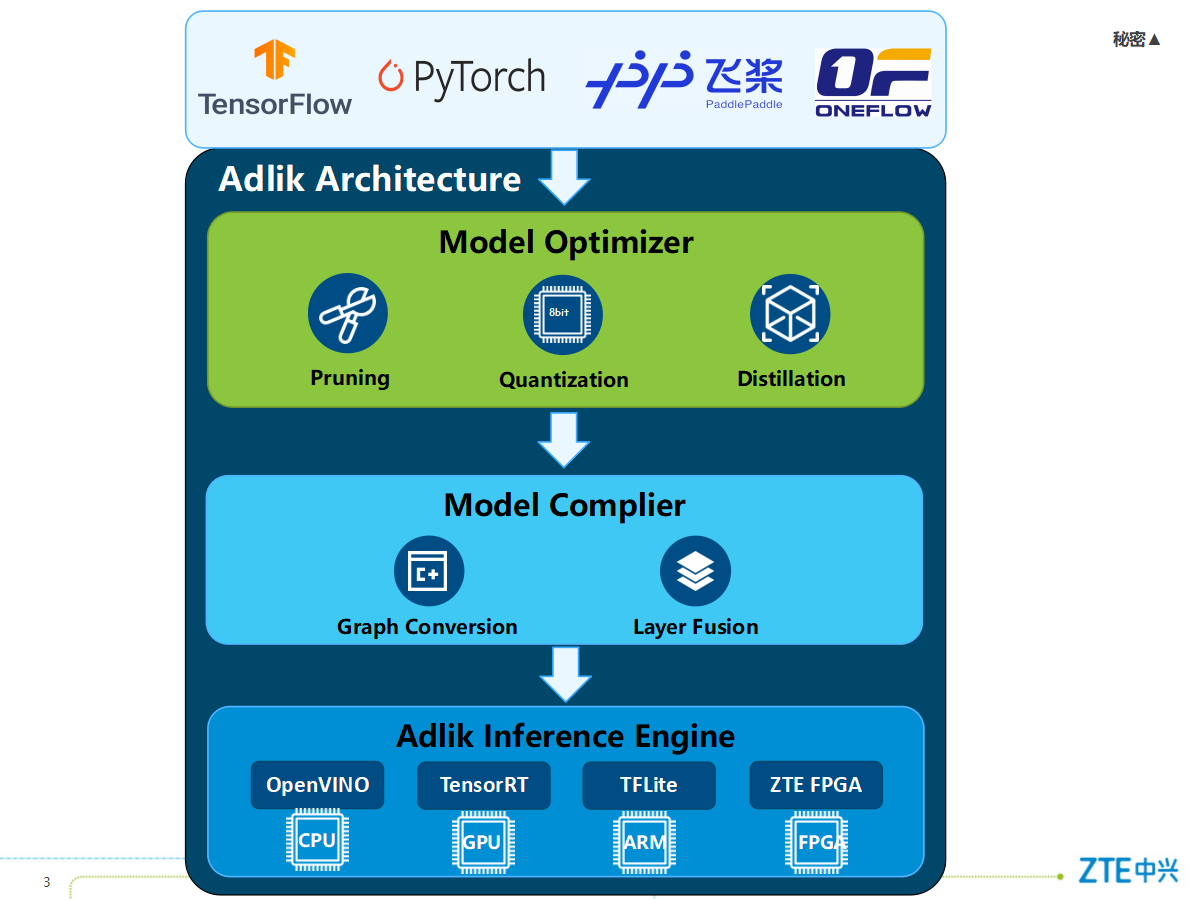


图 1. Adlik 整体结构

训练好的模型，通过Adlik 模型优化器处理，产生优化后的模型，然后通过模型编译器，完成模型格式转换，最终生成推理引擎支持的模型格式，部署到相应的硬件平台上。

#### Adlik模型优化器

Adlik模型优化器致力于在减小模型大小，加速模型推理的同时，提供高准确率的模型，旨在为开发者提供灵活高效的深度学习模型优化方案。为了能够在一般常规硬件上实现模型推理加速，Adlik模型优化器专注于结构化剪枝和INT8量化技术。Adlik模型优化器框架主要包括3个算法组件：剪枝、蒸馏和量化。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 IR模型（如 OpenVINO IR 模型）。

Adlik模型优化器支持多节点、多GPU剪枝和调优方案，支持通道剪枝，和filter剪枝，能够有效降低模型参数量和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，该方法只需用户指定网络类型（如ResNet-50 等）和限制条件（如FLOPs，Latency），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。自动剪枝方法的流程如图 2 所示。

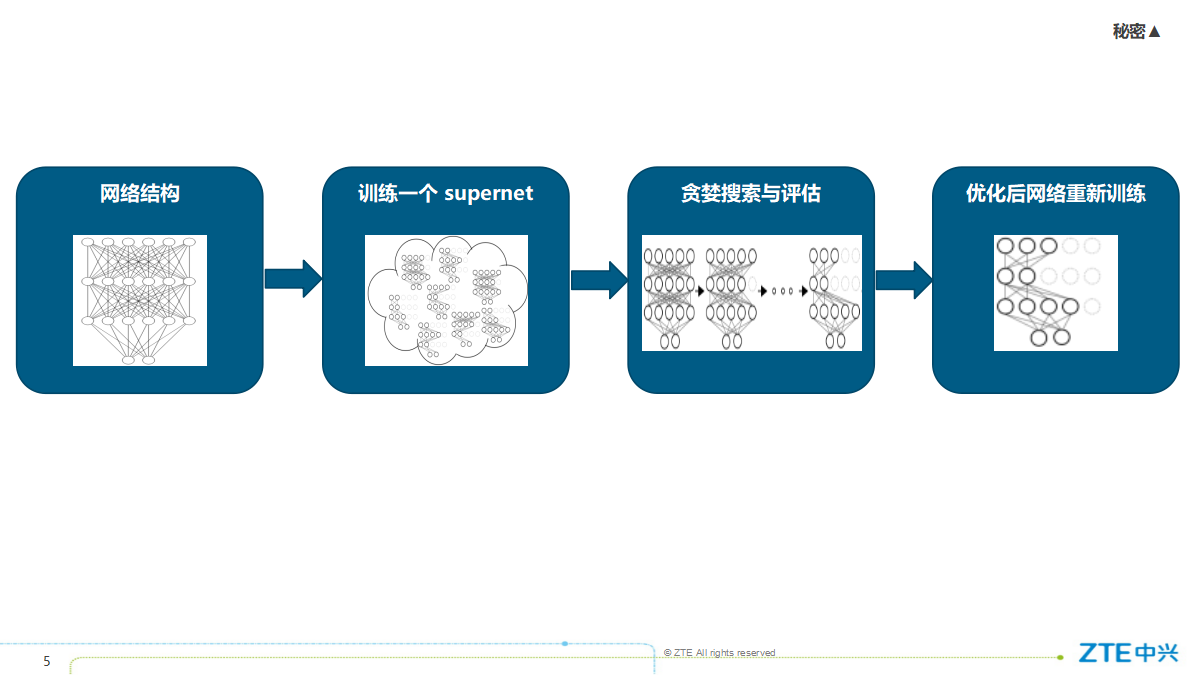


图 2. 自动剪枝的算法流程

一般地，利用已训练完成的大模型去指导小模型训练，使得小模型具有与大模型相当的性能，小模型相对于大模型参数量大幅降低，从而可以实现模型压缩与加速。小模型可以通过剪枝的方式获取。剪枝模型的FLOPs、参数量以及模型大小都将变小，但模型的精度也相应降低了。同时，当我们将 FP32的模型量化为 INT8时，模型的精度也会有一定的损失。为了弥补剪枝和量化导致的模型精度损失，我们采用知识蒸馏的方法，用适当的老师模型（大模型）去蒸馏学生模型（小模型），使得学生模型具有与老师模型相当或更好的性能。

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中（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，我们主要专注于基于输出响应的蒸馏方法。目前，针对单阶段目标检测模型 (如 YOLO 系列) 还没有较为有效的知识蒸馏方法。因此，我们重点研究了较为有效的 YOLO 系列模型的知识蒸馏方法。

**OpenVINO 量化工具和推理引擎**

OpenVINO™ 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推理引擎运行时。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。

OpenVINO 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包括缺省量化Default Quantization (DQ)和精度感知量化Accuracy-aware Quantization (AAQ)[3]。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（例如，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。

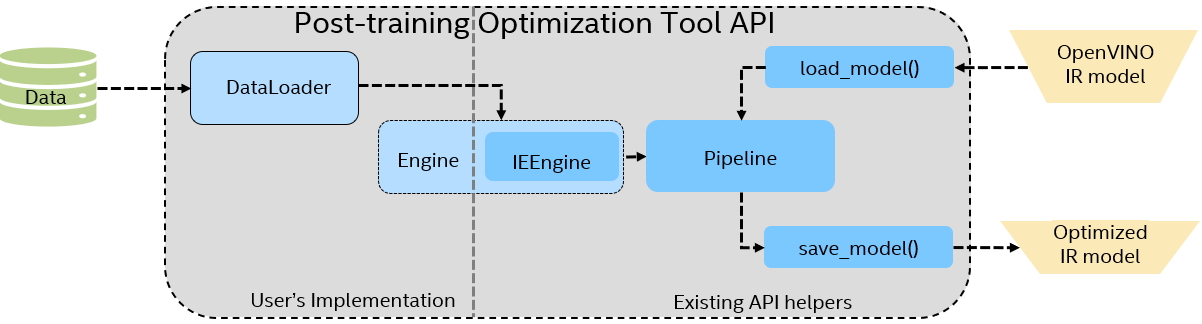


图3. 基于POT API进行INT8量化的通用流程

### 第三代英特尔®至强®可扩展处理器

第三代英特尔® 至强® 可扩展处理器和上一代相比，它带来了 1.46 倍的性能提升[1]，支持 2.66 倍的内存容量[2]并且有一系列特性来支持各种复杂的工作负载，有助于推动经济高效、灵活且可扩展的边缘计算架构，为 AI、数据分析等关键任务提供增强的每节点性能。

英特尔® 深度学习加速这项专门打造的内置加速技术，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。

* 借助所有第三代智能英特尔® 至强® 可扩展处理器上提供的 int8，矢量神经网络指令 (VNNI) 能够通过最大限度地利用计算资源、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。
* 借助部分第三代智能英特尔®至强®可扩展处理器上提供的 bfloat16，业界首次实现 16 位脑浮点 (blfoat16) 的 x86 支持，通过英特尔® 深度学习加速带来增强的人工智能推理和训练性能。

## 应用示例

Adlik 模型优化器的自动剪枝和蒸馏功能结合 OpenVINO 的量化工具和推理引擎，部署在第三代英特尔®至强®可扩展处理器上可实现模型推理的性能的显著提升。此处，选择在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，测试其在该方案下模型推理性能的提升。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® Platinum 8378 处理器，服务器详细配置如表 1 所示。吞吐量使用 OpenVINO 的benchmark 测试工具，测试使用 38 cpu cores。

表 1. 测试配置

|  |  |
| --- | --- |
|  | **ZTE 5300 G4X** |
| CPU | Intel(R) Xeon(R) Platinum 8378C CPU @ 2.80GHz \* 2 |
| CPU cores | 38 \* 2 |
| 内存 | Micron DDR4 64G 2933 MT/s \* 16 |
| 硬盘 | Micron\_5300\_MTFD 3.5TB \* 2 |
| 网卡 | Intel Corporation I210 Gigabit Network Connection \* 2 |

图4. 图像分类 ResNet50模型优化测试结果

从图 4 的测试结果表明，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。INT8量化后的模型的吞吐量比未量化模型提升了 2.96 倍。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。

图5. 目标检测 YOLOv5m模型优化测试结果

从图5 的测试结果表明，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。优化后的 YOLOv5m模型吞吐量比原始模型提升了 3.39 倍。

以上不同场景下的两个典型的深度学习模型的测试结果表明，经优化后的模型其突出的推理性能得益于 Adlik 推理工具链高效易用的模型优化器和对 OpenVINO 推理引擎工具的支持，以及第三代英特尔®至强®可扩展处理器采用 了能够极大提升 INT8 模型的推理性能的VNNI 英特尔深度学习加速 (Intel®DL Boost) 。

同时，该端到端的深度学习模型优化和部署方案，已在中兴通讯和电信运营商的相关业务场景中广泛使用，优化后的模型极大地提升企业 CPU 推理服务器资源利用率，为企业的节能减排也做出了一定的贡献。

## 未来技术展望

[1] 请访问：<https://edc.intel.com/content/www/cn/zh/products/performance/benchmarks/3rd-generation-intel-xeon-scalable-processors/> ，并使用相应的性能编号[#] 来访问完整的系统配置和性能详情：代际性能平均提升高达 1.46 倍 [125]。

[2] 第三代英特尔® 至强® 铂金 8380 处理器：8 通道，3200 MT/s（2 DPC）。与之相比的第二代英特尔® 至强® 铂金 8280 处理器：6 通道，2666 MT/S（2 DPC）。

[3] 精度感知量化算法 <https://docs.openvino.ai/latest/accuracy_aware_README.html>