拨开深度学习部署迷雾，还看 OpenVINO™推理引擎

摘要：OpenVINO™工具套件分发版附带英特尔深度学习推理引擎支持开发人员创建与部署深度学习模型，以及实现快读集成。

《沉默的真相》、《隐秘的角落》，这些热门的悬疑推理大戏，都来自爱奇艺的“迷雾剧场”。然而，很多人可能不知道：“迷雾”成功的背后，离不开另一种推理——人工智能深度学习。爱奇艺的 Jarvis 深度学习云平台，在业务层面服务于全流程智能视频服务，包括智能创作、智能生产、智能播放等等，在技术层面，它支撑智能视频服务在业务弹性扩展、资源统一调度和主流深度学习框架支持等方面的要求。

在深度学习的具体使用过程中，爱奇艺需要应对 AI 基础设施层面的两个主要挑战：
1、AI 应用的爆发式增长，需要基础设施能够提供快速、便捷的部署能力；
2、多样化的 AI 模型和框架，需要基础设施提供更优的支持。

利用 OpenVINO™ 工具套件英特尔发行版，爱奇艺大幅提升了 AI 应用的深度学习推理效率，不同应用的优化加速能力可达数倍至数十倍。

深度学习应用的基础设施部署是普遍存在的难题。知名云服务提供商 Rackspace 调查了 1870 名全世界各行各业的 IT 主管，在实施 AI 面对的诸多困难中，“缺少支持 AI 的技术基础设施”排名第二。这正是英特尔的 OpenVINO™ 工具套件和深度学习部署工具套件的用武之地，特别是英特尔深度学习推理引擎的部署，使用独立于硬件的统一 API，使得部署深度学习解决方案更加容易上手。

一、转换、推理、集成，离不开这些最佳实践

在部署英特尔深度学习推理引擎的过程中，共分三个步骤：

- **转换：**

将经过训练的模型从特定框架（例如 Caffe 或 TensorFlow）转换成独立于框架的中间表示（IR）格式。

- **模型推理/执行：**

经过转换后，推理引擎使用 IR 来执行推理。

- **集成至产品：**

模型推理通过示例验证后，将推理引擎代码集成到实际应用或管道中。

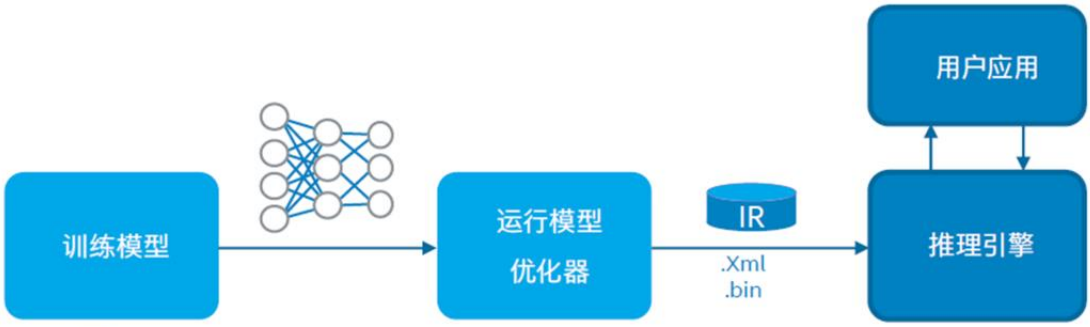


图 1.

其中每个步骤都有一些最佳实践，可以优化性能。要想知道优化的效果如何，有些必备的准备工作，特别是要想清楚需要收集哪些性能数据，具体包括：

- 明确哪些是需要测量的操作，避免包含一次性的操作，推理引擎之外的操作也要单独跟踪。
- 测量延迟与吞吐量。
- 比较原生/框架代码的性能。

做到上述几点，就能获得可信的性能数据。

接下来，让我们看看完成前述三个步骤的部分相关最佳实践。

1、转换

- 针对目标硬件，使用对应插件完成优化；
- 注意 CPU 的吞吐量模式；
- 面向 GPU 和 FPGA 部署时，需要注意相关参数和细节，具体请参考技术文档。

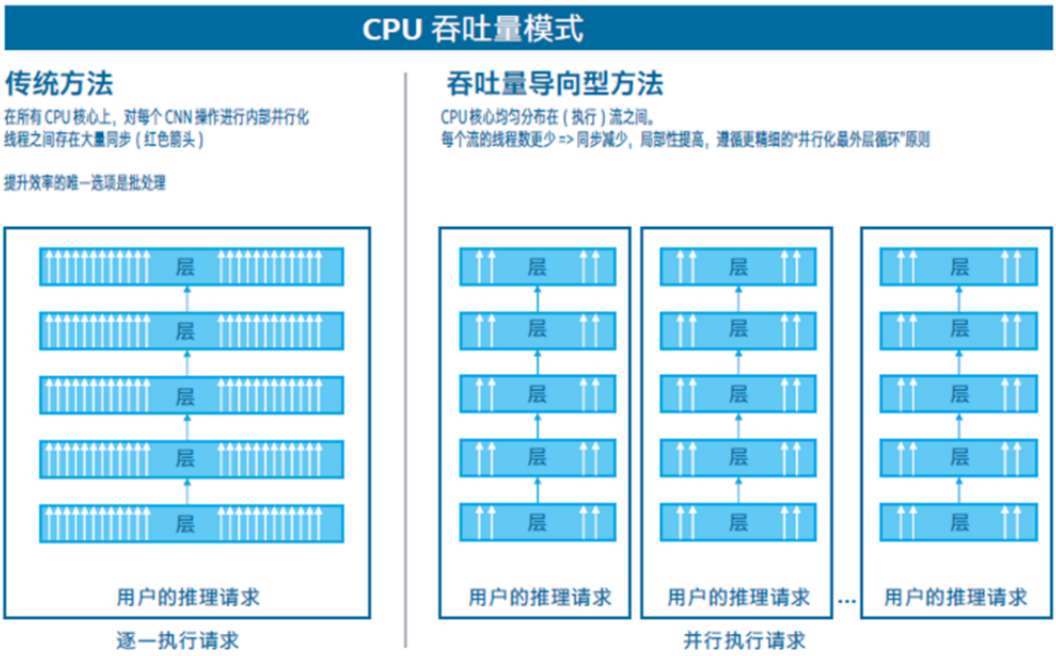
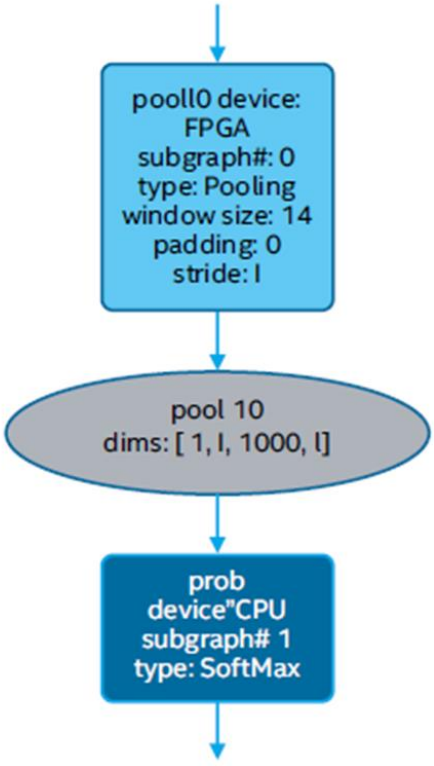


图 2.

2、模型推理/执行

- 在异构模式下的模型推理和执行，要用加速器计算推理网络中负载最繁重的部分；
- 要在不同硬件器件上运行网络分支，这样可以更高效地使用所有的可计算器件；
- FPGA 的异构场景需要注意相关参数；
- 异构执行的结果可通过设置参数辅助可视化分析。



3、产品集成，即将推理引擎插入应用

- 针对 NUMA 架构处理器，需要完成特定参数设定，从而实现最佳性能；
- 应使用推理引擎加速图像预处理和转换；
- 如果要在推理引擎和媒体/图形 API 之间共享数据，一般采用基于共享系统内存的方法；
- 如果应用同时执行多个推理请求，需要考虑特定性能因素，详情请参考技术文档；
- 使用推理引擎异步 API 可提高应用的整体帧速率；
- 使用英特尔 VTune 放大器工具，能够帮助理解分析性能数据。

二、革命性的 AI，无处不在

2019 年，德勤发布了《全球人工智能发展白皮书》，其中的调研指出：大多数采用人工智能的企业相信，AI 将在未来 3 年内彻底改变他们的组织和行业。

从高质量医学成像，到库存自动化货架检查，从人工智能交通控制，到多媒体动画制作，工业和制造业、医疗保健、实验室和生命科学、零售、安全和安保，OpenVINO™ 在所有这些行业中都在大显身手，以高吞吐量、高效率提升推理应用程序的表现。随着 AI 的无处不在，一些机械的、重复的工作将由 AI 帮我们完成，也许我们就能省下来更多时间、更多精力去看烧脑大戏，让我们的大脑在另一种形式的推理中拨开迷雾，现出暖阳。

如欲了解更多 OpenVINO 开发资料，请扫描下方二维码，我们会把最新资料及时推送给您。