案例研究

第二代英特尔® 至强® 可扩展处理器
OpenVINO™ 工具套件英特尔® 发行版
智能视频服务

优化深度学习推理效率，打造更智能视频服务

基于人工智能 (Artificial Intelligence, 以下简称 AI) 技术，为用户提供更为丰富多彩且能“投其所好”的在线视频服务，已成为众多网络视频服务商的共识。作为国内网络视频服务领域的领先企业，爱奇艺* 一直都在积极推动AI应用与视频服务的融合，在智能创作、智能生产、智能播放等全流程智能视频服务中已获取累累硕果。

随着视频 AI 应用的不断推陈出新及其在视频服务中戏份的逐渐加重，它们对爱奇艺基础设施也提出了新的挑战。为应对这些挑战，爱奇艺将 AI 与云计算结合，构建了创新的 Jarvis* 深度学习云平台，以满足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。

新平台的成功运行，很大程度上需要依靠更强的计算能力及其转化而来的更高深度学习效率。为帮助爱奇艺进一步优化深度学习云平台的服务效能，英特尔不仅以英特尔® 至强® 可扩展处理器为该平台输出了更强劲的算力，还基于英特尔® 架构处理器的技术特性，对平台的深度学习推理进行了大量的软、硬件调优，包括利用 OpenVINO™ 工具套件英特尔® 发行版 (以下简称 OpenVINO™ 工具套件) 执行的系统级优化。这些优化帮助爱奇艺大幅提升了在 AI 应用上的深度学习推理效率，并降低了平台总体拥有成本 (Total Cost of Ownership, TCO)，让 AI 在智能化的视频服务中展现出了更强的生产力。

“AI 技术的运用，既能让视频服务的创作、生产、分发和播放等环节变得更为高效和智能，也能给予用户更优的收视和互动体验。英特尔® 至强® 可扩展处理器、OpenVINO™ 工具套件不仅使我们的深度学习云平台获得了更强的算力，也使深度学习推理效率得到了显著的提升。”

张磊
研究员
爱奇艺

爱奇艺解决方案实现的优势：

- OpenVINO™ 工具套件的引入，助力爱奇艺 Jarvis 深度学习云平台有效提升其 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[1]；
- 英特尔® 至强® 可扩展处理器带来的计算性能增强，可进一步提升这些 AI 应用的推理效率。

在线观看视频已是今天互联网生活的重要组成部分之一，截止到 2018 年底，中国在线视频用户已达 6.12 亿，使用时长占比达 12.8%[2]。产业链的不断成熟，也驱使视频服务商引入更多新技术、新能力来提升服务效能、改善用户体验。作为具备产业影响力的高品质视频娱乐服务提供商，爱奇艺正积极引入更多 AI 能力，来构建全流程的在线视频智能服务体系。

图一 AI 在爱奇艺的应用

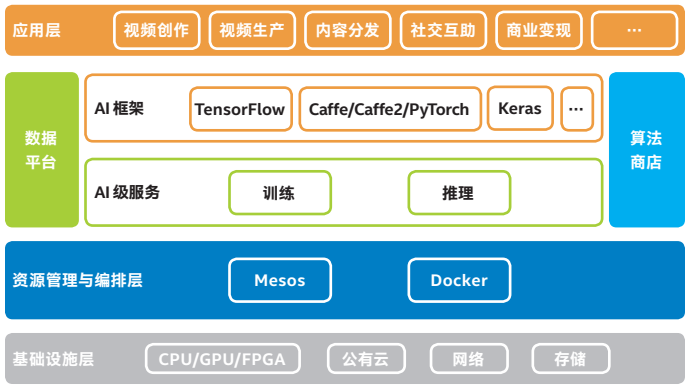
如图一所示，围绕视频的创作、生产、分发和变现等流程，爱奇艺利用 AI 能力将其逐一实施了智能化转型。以视频创作中的“选角色”为例，利用自然语言处理（Natural Language Processing，NLP），爱奇艺可以从角色信息和演员信息中抽取关键信息，并根据 AI 算法来判断二者的匹配度，选出合适的角色；再譬如在视频播放中，利用 AI 强化学习模型，视频平台可实现自适应码流（Adaptive Bit Stream，ABS）播放，改善视频收看体验。

随着 AI 技术在爱奇艺视频服务中戏份的不断增加，它们对既有基础设施也提出了更多需求。首先，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；其次，多样化的 AI 模型和框架，需要基础设施提供更优的支持。此外，对于部署在云端、内容分发网络（Content Delivery Network，CDN）以及客户端等不同环境中的 AI 应用，如何有效调配计算资源提升它们的效率，也是爱奇艺关注的核心需求。

爱奇艺决定通过构建基于云的深度学习平台 Jarvis 来应对以上挑战，它通过与英特尔开展深入技术合作，针对英特尔® 架构进行了全面软、硬件优化，最终有效提升了 AI 应用的效率。

解析爱奇艺深度学习云平台

爱奇艺用于满足 AI 应用需求的 Jarvis 深度学习云平台自下而上分为四层。如图二所示，其基于英特尔® 架构的高性能计算、网络、存储产品所组成的硬件平台，构成了整个云平台坚实的基础设施层。基础设施层之上，是由 Apache Mesos* 和Docker* 等构成的资源管理及编排层，能够实现底层资源的统一管理、调度和审计。同时，平台提供的容器化运行环境，允许使用者按需申请、释放资源，能够有效提升平台的资源利用率。



图二 爱奇艺 Jarvis 深度学习云平台的架构图解

资源管理及编排层之上是 Jarvis 平台的核心能力层，它以数据平台、算法商店以及 AI 能力（AI 训练及推理）等模块，提供了一站式 AI 服务能力平台。数据平台可以通过数据抓取、众包采集以及导入公开数据集等方式，汇聚来自爱奇艺内外的海量数据，并提供智能标注、众包标注等多种数据标注方式，用于对文本、图片、视频和音频等结构化/非结构化数据实施标注。算法商店则为平台使用者提供了各类 AI 算法、网络及模型的展示和交流，并提供了各类算法的效果评估能力来供使用者取舍。

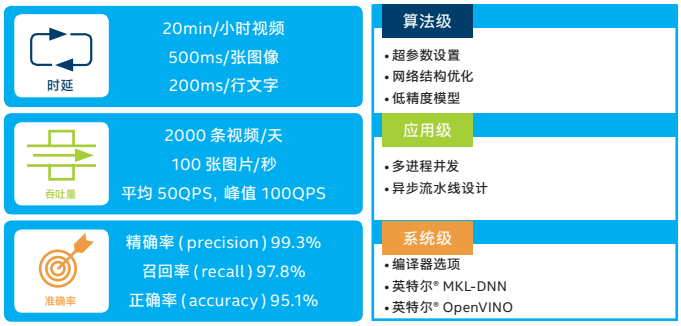
对于关键的 AI 模块，Jarvis 平台一方面对 Caffe*、Tensorflow*、Torch*、Keras*、MxNet* 等主流深度学习框架提供了支持；另一方面，平台的训练模块也为 AI 应用提供了基于底层计算资源池的分布式训练环境，而推理模块则提供了丰富的 AI 模型及服务，让使用者得以快速、便捷地部署高性能的推理服务。在平台的最上层——应用层，则部署了一系列视频 AI 应用，例如视频创作、内容分发、商业变现等，供爱奇艺内外部用户使用。

通过 Jarvis 深度学习云平台，爱奇艺形成了完善的 AI 应用部署流程。首先，来自商业智能（Business Intelligence，BI）平台、大数据分析和各种 AI 应用需求的数据源将被汇总到数据系统，使用者

可以通过 WEB 界面、命令行或者 API 接口将数据加入任务进行训练，而通过训练获取或更新的模型则会被纳入 AI 模型库；最后，使用者可以在算法商店中选择合适的 AI 算法实施推理，并最终以 HTTP 等方式反馈给上层应用或请求。

基于英特尔® 架构的软硬件优化

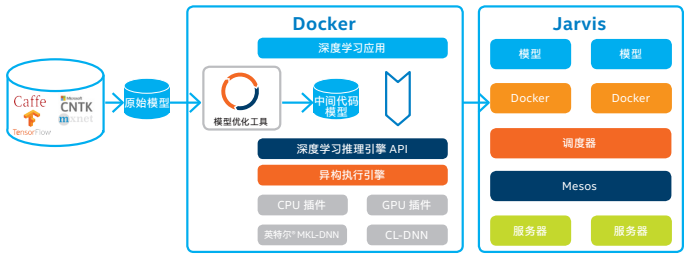
提升深度学习推理效率，是爱奇艺 Jarvis 深度学习云平台增强视频服务生产力的关键能力之一。在英特尔帮助下，爱奇艺基于英特尔® 架构处理器对云平台的深度学习推理能力进行了充分的优化。



图三 爱奇艺制定的推理性能优化指标和优化方案

如图三所示，爱奇艺首先确定了响应时延、吞吐量和模型准确性三个维度的性能指标，并制定出系统级、应用级和算法级三个层级的优化方案。其中，算法级优化侧重于优化深度学习模型本身，使用诸如超参数设置、网络结构剪切和量化等方法，来减小模型的大小和计算强度，进而加速推理过程。在应用级优化上，则是通过改进特定应用程序和服务的流水线和并发性来提升推理效率。通常意义上的深度学习服务不仅包括推理，还包括数据预处理、后处理和网络请求响应，良好的并发设计可有效提升这些应用在服务器上的端到端性能。在系统级优化上，通过引入 SIMD（Single Instruction Multiple Data）指令集、OpenMP 多线程库及英特尔® MKL/MKL-DNN 数学库等优化方法，充分加速整个平台的计算能力，从而全面提升平台的效率。

爱奇艺在系统级优化的基础上，还为 Jarvis 平台引入了来自英特尔的 AI 工具套件——OpenVINO™。



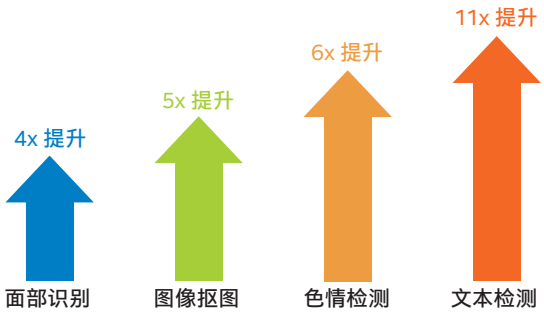
图四 基于 OpenVINO™ 工具套件的 Jarvis 平台推理优化过程

如图四所示，OpenVINO™ 工具套件首先会使用一个模型优化器（Model Optimizer Tool）将原生深度学习模型进行转换和优化，并生成 IR（Intermediate Representation），IR 含有优化后的网络拓扑结构、模型参数以及模型变量，推理引擎（Inference Engine）会读取 IR，执行推理过程。

作为计算机视觉和 AI 技术有机融合的成果，OpenVINO™ 工具套件可以为 Jarvis 平台加速部署在不同计算平台（包括英特尔® 处理器、FPGA 和 VPU）之上的模型推理效率。它包括深度学习推理加速工具包以及计算机视觉工具包，可对 TensorFlow、MXNet、Caffe 等深度学习框架提供良好支持。

以视频播放时的弹幕为例，为了让弹幕信息不干扰正常视频播放，Jarvis 平台通过基于 Deeplab v3+* 深度学习模型的 AI 应用，让弹幕信息隐藏到视频中的主要对象之后。Deeplab v3+ 模型是基于深度卷积网络的语义图像分割模型，它可以通过对单个视频帧上的图像抠图来实现该功能。与传统的计算机视觉算法相比，此模型可适应各种复杂的纹理和场景，例如在前景和背景颜色相似的情况下，提供更准确的结果和更便捷的部署能力。

来自爱奇艺的测试数据显示，OpenVINO™ 工具套件的引入，帮助 Jarvis 平台将实时弹幕显示的推理速度提升达 5 倍左右。爱奇艺 Jarvis 平台上的其他深度学习模型，也验证了 OpenVINO™ 工具套件带来的加速效果。如图五所示，面部识别应用的效率也提升达 4 倍左右，涉黄内容检测的效率提升达 6 倍左右，而在文本检测应用中，推理性能在优化后，更是提升达 11 倍之多[3]。



图五 OpenVINO™ 工具套件提升 Jarvis 平台推理效率

值得一提的是，拥有更强算力的英特尔® 至强® 可扩展处理器，能让基于 OpenVINO™ 的 AI 应用性能进一步提升。新款的处理器往往集成有更多核心，而平台在执行离线推理作业时，推理吞吐量会随着处理器核心数量的增加而线性增加；另一方面，新款处理器中更优化的指令集，例如英特尔® 至强® 可扩展处理器内置的英特尔® AVX-512，也能提供更为强劲的性能加速。来自爱奇艺的对比数据显示，在同样使用 OpenVINO™ 工具套件的情况下，使用英特尔® 至强® 金牌 6148 处理器，与使用英特尔® 至强® E5-2650 v4 处理器相比，用户可获得额外翻倍的性能加速[4]。随着今年发布的、更新的第二代英特尔® 至强® 可扩展处理器在爱奇艺 Jarvis 云平台上的逐渐部署，其性能表现势必会更上一层楼。

展望

现在，以 OpenVINO™ 工具套件为代表的各类优化方法与工具，已在爱奇艺 Jarvis 深度学习云平台上的十余个 AI 应用中获得应用，并在数千个核心中进行了部署[5]。来自爱奇艺 Jarvis 平台一线使用者的反馈表明，这些优化方法已帮助诸多视频 AI 应用有效提升了性能，让视频服务生产力获得了大幅提升。

面向未来，爱奇艺还将与英特尔继续携手，进一步优化其深度学习效率，还计划为其 Jarvis 深度学习云平台添加更多异构计算资源来加速特定任务。同时，在服务灵活性，调度优化和参数自动选择方面，双方也计划更充分地利用计算资源，并使深度学习推理服务获得更为灵活的部署能力。

[1, 3, 4, 5] 数据援引自英特尔Blog: https://software.intel.com/en-us/articles/optimization-practice-of-deep-learning-inference-deployment-on-intel-processors，以及 2018 英特尔人工智能大会上，爱奇艺发言内容：《爱奇艺 AI 视频云架构与优化实践》。所有相关性能数据均源自爱奇艺进行的内部测试，所欲了解更多细节，请联络爱奇艺。

[2] 该数据援引自 CNNIC（中国互联网络信息中心）所发布的第 43 次《中国互联网络发展状况统计报告》。

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。

性能测试中使用的软件和工作负荷可能仅在英特尔微处理器上进行了性能优化。诸如 SYSmark 和 MobileMark 等测试均系基于特定计算机系统、硬件、软件、操作系统及功能。上述任何要素的变动都有可能导致测试结果的变化。请参考其他信息及性能测试（包括结合其他产品使用时的运行性能）以对目标产品进行全面评估。更多信息，详见 www.intel.com/benchmarks。

性能测试结果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。

英特尔技术特性和优势取决于系统配置，并可能需要支持的硬件、软件或服务得以激活。产品性能会基于系统配置有所变化。没有任何产品或组件是绝对安全的。更多信息请从原始设备制造商或零售商处获得，或请见 intel.com。