案例研究

面向英特尔® 架构优化的TensorFlow
面向英特尔® 架构优化的Caffe
英特尔® MKL-DNN
英特尔® AVX-512
人工智能即服务
基础设施即服务

金山云云计算进阶：更优 AI 软硬件“打包出售”

金山云
WWW.KSYUN.COM

利用云服务加速人工智能（Artificial Intelligence，AI）应用的研发，已成为众多AI开发团队的共同选择。作为全球高品质云服务专家，金山云也致力于通过一系列高性能云服务器、云物理主机产品为AI工作负载提供更优 IaaS（Infrastructure as a Service，基础设施即服务）服务，助力用户在语音、图像、视频等诸多 AI 应用场景中占得先机。

为帮助用户更有效地提升 AI 研发效率，金山云与深度合作伙伴英特尔一起，不仅在其云实例中部署了英特尔® 至强® 可扩展处理器等先进硬件产品，还据此引入了包含面向英特尔® 架构优化的 TensorFlow、面向英特尔® 架构优化的 Caffe 等 AI 框架的镜像，并通过这种更优 AI 软硬件的“打包”，形成对 AI 工作负载有更优支持的 IaaS 能力，让用户在无需关心底层 AI 框架配置和调优等复杂性的同时，在基于英特尔® 至强® 可扩展处理器的云基础设施上一站式获取更优性能。

来自金山云的测试数据显示，多个优化后的 AI 框架在不同深度学习模型中实现的性能，都获得了数倍乃至数十倍的大幅提升。这充分表明，将更优 AI 软硬件“打包出售”的金山云增强型 IaaS 服务，可为不同应用场景下的 AI 研发提供强有力的性能助攻，加速其用户的AI研发进程。

“金山云一直致力于为 AI 研发提供第一流的云服务支撑，这不仅需要高性能的硬件平台作为基础，也依赖于针对软硬件的协同及深度优化。通过众多英特尔先进硬件产品的部署，以及面向英特尔® 架构优化的 TensorFlow、面向英特尔® 架构优化的 Caffe 等优化框架的引入，我们在帮助用户在获得增强型 IaaS 的同时，也大幅降低他们用在系统部署和优化上的投入，让他们能将更多的关注点投向AI业务本身。”

杨峰
金山云云计算研发总监
金山云

金山云实现的解决方案优势[1]：

- 通过向用户提供多种预打包的、面向英特尔® 架构优化的深度学习框架，使利用金山云实施AI研发的用户可以有效减少花在相关框架下载、部署及优化上的时间和精力，将更多资源投入到 AI 业务层面；
- 面向英特尔® 架构优化的 TensorFlow，助力基于英特尔® 架构处理器构建的金山云云实例，在多种深度神经网络模型上获得显著性能提升，提升幅度在 2.45 倍至 2.89 倍之间[1]；
- 基于英特尔® 架构处理器构建的金山云云实例，在引入面向英特尔® 架构优化的 Caffe 后，其性能在多种深度神经网络模型上获得大幅提升，在 ResNet50 模型中的增幅更达到近 30 倍[1]。

为 AI 提供高性能 IaaS 服务

持续演进的公有云服务正在 AI 研发中扮演更重要的角色，其对资源的灵活调配，以及高度的可扩展性能对 AI 研发所需的算力、算法和数据实施敏捷调度，提升效率。因此，越来越多的 AI 开发团队正选择以云服务为基础来开展 AI 研发与创新。

为向用户提供更具效能和性价比的 IaaS 服务，金山云正与英特尔展开深入合作，在其云实例（云服务器、云物理主机等）中持续引入英特尔® 至强® 可扩展处理器、英特尔® 傲腾™ 数据中心级固态盘、25GbE 英特尔® 以太网网络适配器等先进硬件产品与技术，作为构建高性能 IaaS 能力的基石。

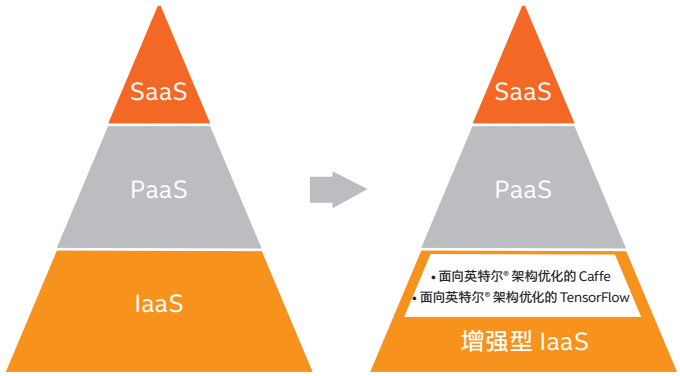
以金山云采用的英特尔® 至强® 铂金 8168 处理器为例，其具备优化的微架构，集成多达 24 个内核并支持 48 个线程，在计算密集型的AI推理工作负载中，具有更出色的计算性能和可扩展性。同时其提供的英特尔® 高级矢量扩展 512（英特尔® AVX-512），可同时处理 16 个单精浮点数，与上一代英特尔® 高级矢量扩展 2（英特尔® AVX 2）相比，能通过更多的融合乘加（Fuse Multiple Add，FMA）单元令单精浮点数处理能力加倍，这一特性在应对高密度的AI矢量计算需求时，有着非常明显的优势。

同时，英特尔® 处理器平台也在通过持续不断的更新换代，来强化这些应用优势，随着全新第二代英特尔® 至强® 可扩展处理器的到来，其集成的英特尔® 深度学习加速技术也将与其他硬件技术与产品相配合，用于提升金山云云实例整体架构的性能与可扩展性，有望在语音、图像、视频等诸多 AI 应用场景中，为 AI 开发团队添油加力。

针对英特尔® 架构的 AI 框架优化

但硬件性能的强化是否能带来 AI 工作效率的同比例提升？来自金山云的探索表明，在选择高性能硬件设施之余，AI 开发团队通常还需要推进复杂的深度学习框架的安装、部署和调优，来确保效率最优化。这不仅会消耗宝贵的时间，如果优化效果不佳还会带来严重的资源浪费。例如用户申请了拥有 24 个 vCPU（虚拟处理器核心）的云主机实例，但由于原生 AI 框架在并行处理能力上的不足，可能只有 50% 的处理器核心得到了充分利用，造成用户不得不申请更多的云实例来满足需求，从而降低了系统运行效率，并提升了总拥有成本（Total Cost of Ownership，TCO）。

为了破解这个难题，金山云正与英特尔携手，为用户提供面向英特尔® 架构优化的 AI 镜像，镜像中集成了多种优化的深度学习框架。这一面向 AI 研发的增强型 IaaS 云服务，让用户可以免去下载、安装、配置和调优这些框架的繁琐步骤，在基于英特尔® 至强® 可扩展处理器的云平台上，一站式获得更优运行性能。



图一 面向 AI 研发的增强型 IaaS 云服务

以面向英特尔® 架构优化的 TensorFlow 为例，首先，它对面向深度学习的英特尔® 核心数学函数库（英特尔® MKL-DNN）中的深度学习原语（DNN Primitive）提供了良好的支持，这包括了 2D 卷积、内积/矩阵乘法、批量归一化、ReLU 激活以及多维转置等一系列常用 AI 计算过程。当面向英特尔® 架构优化的 TensorFlow 在英特尔® 至强® 可扩展处理器平台上部署时，将会优先使用英特尔® MKL-DNN 提供的原语，来帮助用户快速搭建所需的功能模块。

同时，面向英特尔® 架构优化的 TensorFlow 也以代码重构的方式，将深度学习计算中所需的大量计算过程，例如卷积、矩阵乘法等进行矢量化，并交由英特尔® AVX-512 来执行处理，最大程度地挖掘英特尔® AVX-512 具备的矢量计算优势。另外，面向英特尔® 架构优化的 TensorFlow 还可以对空闲处理器核心进行调度，从而将英特尔® 至强® 可扩展处理器的多核心优势进一步放大。

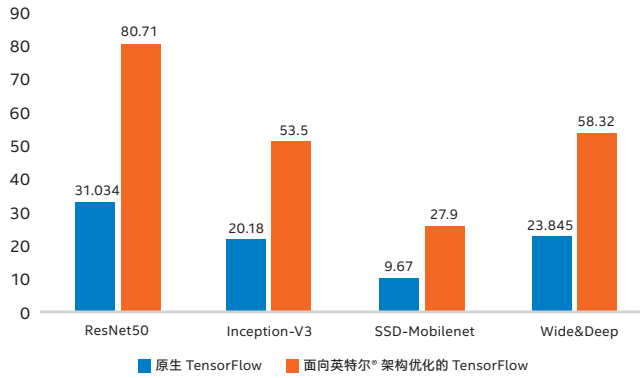
同样，面向英特尔® 架构优化的 Caffe 也能充分利用英特尔® MKL-DNN 来加速 AI 工作负载中的各计算过程，例如使用英特尔® MKL-DNN 中的高度矢量化和线程化的构建模块，在 C 和 C++ 接

口中快速实现卷积神经网络模型，并融入层融合等技术方案，来进一步提升 AI 推理性能。

优化前后的性能对比

为验证在引入面向英特尔® 架构优化的 AI 镜像后，云主机实例所能获得的性能提升，金山云与英特尔一同，针对深度学习中常用的 ResNet50（一种残差网络）、ResNeXt50（一种升级残差网络）、Inception-V3（一种卷积神经网络）、SSD-Mobilenet（一种目标检测网络）以及 Wide&Deep（一种经典的推荐算法模型，基于 movielens-1M 数据集）网络模型，在金山云通用型 N3 实例[2]上进行了一系列测试。这些神经网络模型目前正被广泛地应用于图像分割、内容推荐等常用 AI 场景。

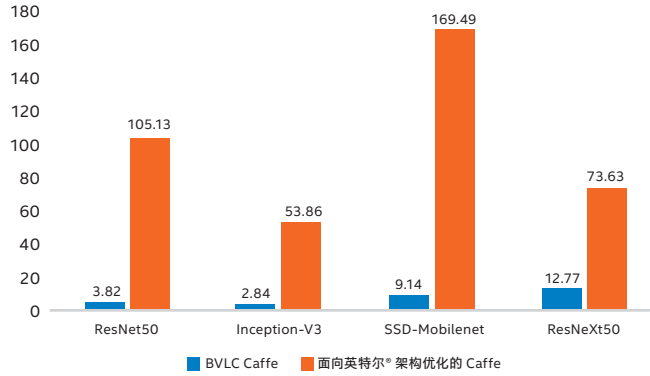
首先是面向英特尔® 架构优化的 TensorFlow 与原生 TensorFlow 在ResNet50、Inception-V3、SSD-Mobilenet 和 Wide&Deep 四种深度神经网络中实施 AI 推理的性能对比。其中 ResNet50、Inception3 和 SSD-Mobilenet 中的 Batch_Size 设为 1，而 Wide&Deep 中的 Batch_Size 设为 256。



图二 基于原生 TensorFlow 与面向英特尔® 架构优化的 TensorFlow，在不同深度神经网络的 AI 推理性能对比

对比测试结果如图二所示，在采用面向英特尔® 架构优化的 TensorFlow 后，与原生 TensorFlow 相比，云实例在不同深度神经网络中，AI 推理性能均获得了不同程度的提升。其中在 SSD-Mobilenet 网络中，性能提升达 2.89 倍左右。

而在另一组测试中，双方比较了面向英特尔® 架构优化的 Caffe 与 BLVC Caffe 在 ResNet50、Inception-V3、SSD-Mobilenet 和 ResNeXt50 四种神经网络中实施前向传播的性能，四种神经网络中的 Batch_Size 均设为 1。



图三 基于原生 Caffe，与面向英特尔® 架构优化的 Caffe 在不同深度神经网络的 AI 前向传播测试性能对比

对比测试结果如图三所示，通过测试数据可以看出，BVLC Caffe 在各个深度神经网络中的性能表现均差强人意，而与其相比，面向英特尔® 架构优化的 Caffe 在不同深度神经网络中的 AI 前向传播性能均获得了数倍乃至数十倍的大幅提升。在 ResNet50 网络中，提升幅度更是达到了惊人的 27.5 倍左右。

结语

通过向用户提供高性能硬件基础设施，以及包含多种优化深度学习框架的AI镜像，并将这两者的能力“打包”输出给用户，金山云的增强型 IaaS 可为 AI 开发团队提供性能更强、技术方案更完善、扩展性更好的一站式解决方案，助他们将资源更多投向应用研发、投向业务本身，在提升开发效率的同时降低 TCO。

着眼未来，金山云与英特尔还将围绕云服务如何进一步推动 AI 研发效率开展更多也更深入的技术合作，特别是随着第二代英特尔® 至强® 可扩展处理器、英特尔® 傲腾™ 数据中心级持久内存等全新硬件产品在金山云的逐步部署，双方也将针对如何利用英特尔® 深度学习加速技术、高密度内存云实例对 AI 的应用价值推进更多研究、探索以及产品化的工作。

Tips

面向英特尔® 架构优化的 TensorFlow: TensorFlow 是广泛应用于 AI 各领域的深度学习开源框架, 其可对计算机视觉、语音识别、自然语言处理等工作负载提供良好支持。为提升原生 TensorFlow 在英特尔® 架构处理器平台上的运行性能, 英特尔与合作伙伴一起, 进行了大量优化工作。包括: 对英特尔® AVX-512 指令集的更有效利用; 提高处理器核心利用率来实现性能提升; 在给定的层或操作上实施并行化, 或在层间实施并行化; 平衡使用预取、缓存模块化技术, 以及改进空间和时间局部性的数据格式等。在实施以上优化后, 面向英特尔® 架构优化的 TensorFlow 在性能上, 相较原生 TensorFlow 有了巨大的提升。

面向英特尔® 架构优化的Caffe: Caffe 是由伯克利视觉和学习中心 (Berkeley Vision and Learning Center, BVLC) 和社区贡献者开发的深度学习框架, 其内置了大量预训练模型, 不仅可为 AI 应用提供强大的视觉、语音和多媒体支持, 还支持使用 OpenCV (一种广泛使用的计算机视觉库) 为移动设备增加计算机视觉功能。面向英特尔® 架构优化的Caffe不仅继承了 BVLC Caffe 的全部优点, 还提供了针对英特尔® 架构优化的功能和多节点分布式训练和评分。其可通过代码矢量化来高效使用处理器资源, 从而改进函数调用性能、降低算法复杂程度, 并减少计算次数。同时, 该版本还引入大量针对处理器和系统的代码优化、以及 OpenMP 的代码并行化技术, 使其与 BVLC Caffe 相比, 性能获得大幅提升。

更多详情, 请参考链接:

https://github.com/IntelAI/models

https://github.com/intel/caffe

https://github.com/tensorflow/tensorflow#community-supported-builds

https://www.intel.ai/tensorflow-optimizations-intel-xeon-scalable-processor/

https://software.intel.com/en-us/articles/intel-optimization-for-tensorflow-installation-guide

https://software.intel.com/en-us/mkl/documentation/view-all

https://www.intel.com/content/www/us/en/architecture-and-technology/avx-512-overview.html

[1] 解决方案优势所涉及数据均引用本文“优化前后的性能对比”部分所涉及的测试, 详见该部分描述。

[2] 金山云通用型 N3 实例采用英特尔® 至强® 铂金 8168 处理器, 工作频率为 2.70GHz, 采用 DDR4 DRAM内存, 测试中的面向英特尔® 架构优化的 TensorFlow 版本为 r1.12, 面向英特尔® 架构优化的 Caffe 版本为 1.15, 本文测试数据均采用 24 个虚拟处理器核心的云主机实例, 详情请参阅金山云官网: https://marketplace.ksyun.com/products/10291

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。

描述的成本降低情景均旨在在特定情况和配置中举例说明特定英特尔产品如何影响未来成本并提供成本节约。情况均不同。英特尔不保证任何成本或成本降低。

英特尔、Intel、至强、傲腾、MKL-DNN、AVX-512 是英特尔公司在美国和或其他国家的商标。英特尔商标或商标及品牌名称资料库的全部名单请见intel.com上的商标。