人工智能云服务（AI-AS-A-SERVICE）构建指南

如何借助人工智能和机器学习发展云服务业务

目录

简介

- 人工智能（AI）开始为所有类型的企业和组织带来真正的价值
- 云服务提供商（CSP）有机会帮助客户充分利用人工智能带来的优势，借此拓展自身的业务
- 要为构建引人注目的人工智能方案打好基础，云服务提供商需要关注以下方面：

  a. 改善基础设施的可扩展性；

  b. 为人工智能项目选择适当的框架；

  c. 确保存储系统构架合理，能够支持机器学习和深度学习工作负载；

  d. 提高运行机器学习和深度学习算法的平台的性能和速度；

人工智能和机器学习：拨开炒作的迷雾

关于“人工智能即将改变世界”的说法，我们已经听过无数次了。现在，随着人工智能的三要素（数据、计算和应用实例）日趋完备，这一预言可能即将成为现实。

1. 由于设备数量呈指数级增长，我们拥有的数据比以往任何时候都多。
2. 凭借如今的计算能力，我们已能够高速处理大量数据。
3. 在率先采用人工智能的应用案例和商业模式中，我们已经见证了人工智能变革性的力量。

数据和计算技术的逐步成熟，加快了机器学习和深度学习模型及神经网络的发展。训练和运行这些模型需要强大的处理能力和极高的数据存储量，但随着技术和现代计算模型（如云计算）的飞速发展，这一障碍正逐步消失。换句话说，越发经济且高效的计算、存储和数据技术是推动当今人工智能技术蓬勃发展的关键因素。这些因素使得一些人工智能应用案例取得了初步成功，所取得的成就和经验教训激发了一系列新的项目。例如，作为支撑自动驾驶技术不断发展的关键因素，人工智能正对汽车行业产生极其深远的影响：持续投资、原型设计和迭代更新让汽车行业取得了长足进展。再比如，高级分析技术给所有类型的组织带来了明显的优势。这两种应用案例展现了人工智能技术深不可测的潜力。

术语

人工智能

一个宽泛的术语，代表任何能够感知、推理、行动和适应的程序。

机器学习

能够根据数据构建模型并随着数据量的增加不断改进的学习算法。

深度学习

机器学习的一个子集：利用多层神经网络从大量数据中进行学习。

了解更多信息

图 1：了解人工智能、机器学习和深度学习。

现状：我们处于哪一阶段？

虽然我们逐渐开始了解人工智能带来的变革性力量，但目前只有部分大型组织在人工智能方面进行了大规模投资[1]。大多数中小型组织仍处于实施机器学习或深度学习技术的早期阶段，甚至尚未接纳此类技术[2]。

人工智能的蓬勃发展是有根有据的。除了让我们在无需人工干预的情况下实现流程自动化以外，现在人工智能还帮助我们实现以前无法实现的目标。例如，计算机可以比人类更快地筛选出包含肿瘤的医学扫描图像，比医学专家每分钟处理更多的图像。此外，人工智能系统还能对极大规模的数据进行观测，例如评估整个人群的健康数据，以预测和预防可能威胁生命的疾病 — 这种工作仅靠人力根本无法完成。在这个例子中，人工智能执行的是具有开创性的工作，这不是分担人类的工作，而是执行超出人类能力范围的任务。在未来几年里，很可能会涌现出更多这样的用例。人工智能不是孤立的技术事件，我们对人工智能的使用将随时间的推移不断发展和演进。

当前，开发人工智能系统的大多数企业和组织的侧重点是机器学习，即对机器进行训练，培养出制定决策（例如，学习如何识别图像或其他数据中的模式）所需的技能。而目前的主要增长领域是机器学习的一个子集 — 深度学习。深度学习让开发人员能够利用感测数据创建服务，为应用增加一个智能层，例如识别语音输入，让用户能够使用语音命令（如 Amazon Alexa*）。自动驾驶车辆是另一个高速发展的深度学习领域。让机器控制车辆的过程十分复杂，需要不断重复以下循环：

1. 感测周围世界的状态，包括其他车辆的位置和速度、当前路段的交通规则、行人和障碍物、行进方向等。
2. 通过规划和计算，确定最优的下一步行动。
3. 采取继续行进所需的行动。一旦采取行动，周围世界的状态就会由此发生改变，因此机器必须立即开始新一轮的感测过程（见图 2）。

图 2：要实现机器对车辆的控制，必须不断重复执行“感测、制定决策和采取行动”的过程。

我们欠缺什么？

人工智能系统在过去几年发展迅速，并在自动驾驶等领域取得了初步成功。但在其他领域，人工智能技术离大规模应用还有一定的差距。举例来说，我们还需要完善更精密的感测技术才能应对复杂和快速变化的环境。此外，我们还缺少用于训练人工智能系统的完整数据。除了技术本身，还有一些问题需要解决，例如，我们有时无法解释人工智能做出的决策和采取的行动。日后，这可能会成为一项法律要求，例如，当自动驾驶车辆出现交通事故时需要说明原因。

下一步该怎么走？

人工智能的潜力非常大，可能涉及到我们日常生活的所有领域。在许多方面，定义人工智能的范围就像试图界定传统软件的范围一样：有着无穷无尽的用途和可能性。从宏观角度看，人工智能开发人员目前的关注点在于模式识别，例如图像识别或语音识别。完善这些技术可为 Amazon Alexa、Microsoft Cortana* 和 Apple Siri* 等语音助理用例增加视觉搜索功能，例如在监控影像中搜寻犯罪嫌疑人。

人工智能还可能在异常检测领域大放异彩，这在某些方面与模式识别恰好相反。基于人工智能的异常检测具有远超以往的巨大潜力，并且可能产生更深远的影响：例如，检测异常心跳有可能及早发现心脏骤停或其他健康问题；对正常的网络行为进行建模，以检测并标记犯罪分子的恶意网络行为。检测人员或事件的异常行为极其困难，因为根据定义，一开始关于异常的数据会非常少，甚至不存在。通过观测规模超出人类处理范围的数据，人工智能展现出为人类社会带来全新效益的强大威力和机遇。

云服务提供商能够从哪方面带来价值？

人工智能为云服务提供商提供了许多机会，包括改善服务和数据中心流程、为客户提供基于人工智能的全新服务，具体包括：

帮助客户克服人工智能挑战

70% 的企业希望在 2018 年实施人工智能[3]，但 91% 的企业预计会在采用人工智能的过程中遇到重大障碍[4]。换句话说，您的客户有明确的投资人工智能的意向，但他们中的许多人需要帮助，以理清所能实现的目标以及如何克服在采用人工智能过程中遇到的技术和文化障碍。好消息是，企业在采用人工智能的过程中最常见的障碍正是云服务提供商能够帮助客户克服的问题：缺乏适合的 IT 基础设施、专业人才及开发人工智能解决方案所需的专业知识。

协助客户组织数据

企业在开发人工智能方面通常需要技术支持，尤其是在清理和组织机器学习和深度学习算法所使用的数据方面。如果云服务提供商能够帮助客户解决这一难题，必能找到构建人工智能服务的有利途径。此外，这也有助于提高云服务提供商的信誉，成为值得信赖的顾问。

根据需要汇总客户的边缘设备和传感器数据

人工智能的许多用例（如自动驾驶）都涉及传感器收集的数据，并且在某些情况下，这些数据由边缘设备进行处理（边缘计算而非云计算）。如果无需将所有传感器数据传输到中央云进行集中处理，这样就能消除延迟，使人工智能系统更快地制定决策。对于自动驾驶等极其重视决策制定速度的应用来说，这一点至关重要。但是，保留制定决策所用的数据对于未来系统的发展也很重要。通过汇总传感器和决策制定的数据，云服务提供商能够奠定提供增值服务的基础，例如与其他数据源混合并应用大数据分析，从而为客户提供深刻洞察。

存储客户用于训练人工智能系统的数据

某些机器学习和深度学习算法需要大量的训练数据。许多组织无法以经济高效的方式存储如此多的数据，因此训练算法时需要花费很长时间。基础设施即服务（IaaS）提供商完全有能力占据这一市场，针对人工智能训练的需求提供量身定制的数据存储解决方案。

与客户合作创建所需的算法

您的客户可能希望在人工智能领域取得长足发展，他们大致明白自己想要什么，但还没有足够的能力或专业知识来构建自己的人工智能。提供人工智能开发解决方案（例如创建算法）也许是另一条可为云服务提供商带来利润的途径。Amazon、Google 等大型云服务提供商已经为客户和开发人员提供了许多算法和工具，因此这类产品并不是唯一的。但是，您的现有客户和潜在客户可能需要更多量身定制的解决方案，因此这对他们而言是一项极具吸引力的增值服务。要想抓住这个机会，就需要创建更贴近客户群需求的模型。

企业采用人工智能时面临的障碍

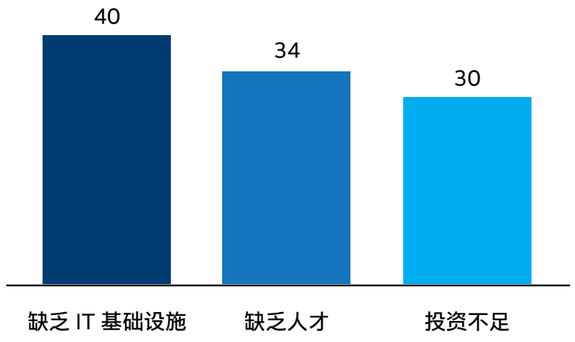

图 3：企业采用人工智能技术的重大障碍（来源：Teradata）[5]

云服务提供商能够从哪方面带来价值？

提供不断发展的人工智能云服务（AIaaS）

Google、Amazon、Microsoft、IBM 等超大型云服务提供商很快意识到了如何将他们在人工智能和机器学习上的投资转化为收入。在过去几年里，他们已将自己为改进服务而开发的技术转变为针对客户群的新产品。

但是，这些市场领导者采取了不同的方法来实现人工智能云服务（AIaaS）。Google 将机器学习技术包装在 Google 图片搜索（视觉识别）和 Google 翻译（语音识别）产品中进行转售；IBM 则采取了针对特定领域的方法，例如针对医疗服务和零售等行业采用 Watson 认知计算技术，以解决特定的行业痛点[6]。全球最大的云服务提供商纷纷采用多样化的方法转售人工智能产品，这表明其他云服务提供商也可以通过不同的途径为客户提供独特的人工智能云服务产品。

人工智能有可能帮助我们解决一些重大的社会和商业难题，这意味着云服务提供商可以深入钻研许多不同的领域。为了确定适合从哪个领域涉足人工智能云服务，可以思考以下问题：

- 您可以对目前业务中的机器学习、深度学习或其他人工智能部署进行研究，挑选出哪些可以做调整，从而为提供人工智能云服务产品奠定基础？
- 您的客户群可能最需要哪些类型的人工智能解决方案？例如，您的业务是否为那些从人工智能测试环境受益的开发人员提供服务？您是否有大量的医疗服务类客户需要简单易用的图像分类系统？有多少客户拥有呼叫中心并可能对人工智能语音识别系统感兴趣？
- 您的组织是否在人工智能的许多应用领域都拥有深厚的实力，例如预测性分析？您能否将它们包装成面向多个行业的服务？
- 与您的客户相似的组织目前是如何使用人工智能技术的？了解与您的客户相似的组织如何使用人工智能和机器学习技术，可帮助您为自己的客户确立发展路径。不管您为医疗服务、工业、金融服务、农业、汽车、应用开发、研究还是其他领域提供服务，都有大量的应用案例可帮助您确定从何处着手。

技术考量因素

一旦您确定了实现人工智能的业务途径，下一步就要调整支持人工智能工作负载所需的基础设施。机器学习和深度学习过程通常会对处理、存储和功耗提出极高的要求，但大多数用例不需要全新的基础设施。确定如何融合新技术与现有系统，以尽可能低的总体拥有成本（TCO）实现最佳的人工智能成果是迈向成功的第一步。您需要特别注意哪些基础设施性能因素呢？请看下面列出的几点优先考虑事项。

1. 可扩展性

不管您如何实现云上人工智能服务创新，要想说服客户投资购买您提供的人工智能解决方案，就必须给出有竞争力的价格。也就是说，您必须以尽可能经济高效的方式提供人工智能解决方案，尽可能多地利用您现有的基础设施。部署专门处理人工智能工作负载的新系统可能需要付出高昂的资本和功耗费用。

好消息是，大多数基于英特尔® 架构的现有虚拟化基础设施都能支持常见的机器学习和人工智能应用负载。您只需要最大限度发挥现有基础设施的可扩展性，确保它能够支持您希望运行的机器学习算法。在这方面，英特尔® 至强® 可扩展处理器能够助您一臂之力，与上一代技术相比，该处理器能够将人工智能/深度学习的训练速度提高 2.2 倍[7]，最高可将总体拥有成本降低 65%[8]。更新您的基础设施，最大限度提高可扩展性，同时降低总体拥有成本，这将为您提供经济高效的人工智能工作负载打下良好的基础。

2. 框架

开始构建人工智能解决方案之前，根据您的应用案例、目的和内部技能组合选择最佳的框架至关重要（见表 1）。一些框架更适合机器学习任务（如视觉分类），其他的则适合特定的深度学习任务。有些框架提供了大量的预训练模型，降低了入门难度；另一些则可能具有更陡峭的学习曲线。您的开发团队可能更喜欢那些基于他们熟悉的语言或平台的特定工具。

无论如何，最重要的是，英特尔已针对英特尔® 至强® 技术对表 1 中的每种框架进行了优化，使其性能在某些情况下实现了数量级的提升[9]。在英特尔® MKL-DNN [10,11] 的辅助下，使用英特尔® 至强® 处理器可实现 100 多倍的性能提升。例如，在英特尔® 至强® 处理器 E5-2666 v3（c4.8xlarge AWS* EC2* 实例）上使用 Apache MXNet 的 AlexNet*、GoogleNet* v1、ResNet-50* 和 GoogleNet v3 在所有可用 CPU 内核上的推理吞吐量分别提高了 111 倍、109 倍、66 倍和 92 倍[10,11]。图 1 和图 2 比较了分别使用英特尔® 至强® 处理器 E5-2699 v4 和英特尔® 可扩展铂金 8180 处理器及英特尔® MKL-DNN 运行 TensorFlow 和面向英特尔® 架构优化的 Caffe 时的训练和推理吞吐量。随着进一步的优化，这些框架及其他框架的性能预计会继续提升。

技术考量因素

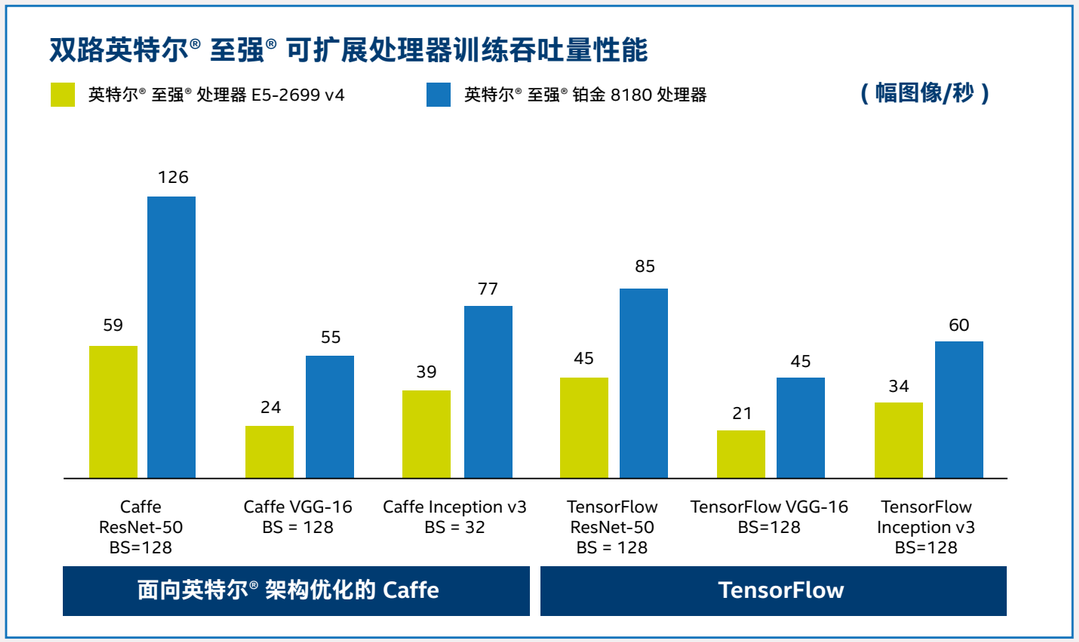

图 4：使用 ResNet-50、VGG-16 和 Inception-v3 及各种 mini-batch sizeds (BS) 时，在英特尔® 至强® 处理器 v4 (原代号 Broadwell) h (浅蓝色) 和英特尔® 至强® 可扩展处理器 (原代号 Skylake) j (深蓝色) 上运行面向英特尔® 架构优化的 Caffe 和 TensorFlow 的训练 (Training) 吞吐量。从采用 AVX-2 指令集的英特尔® 至强® 处理器 v4 开始，英特尔® MKL-DNN 可提供显著的性能提升；在引入了 AVX-512 指令集的英特尔® 至强® 可扩展处理器上，英特尔® MKL-DNN 让性能大幅飙升。

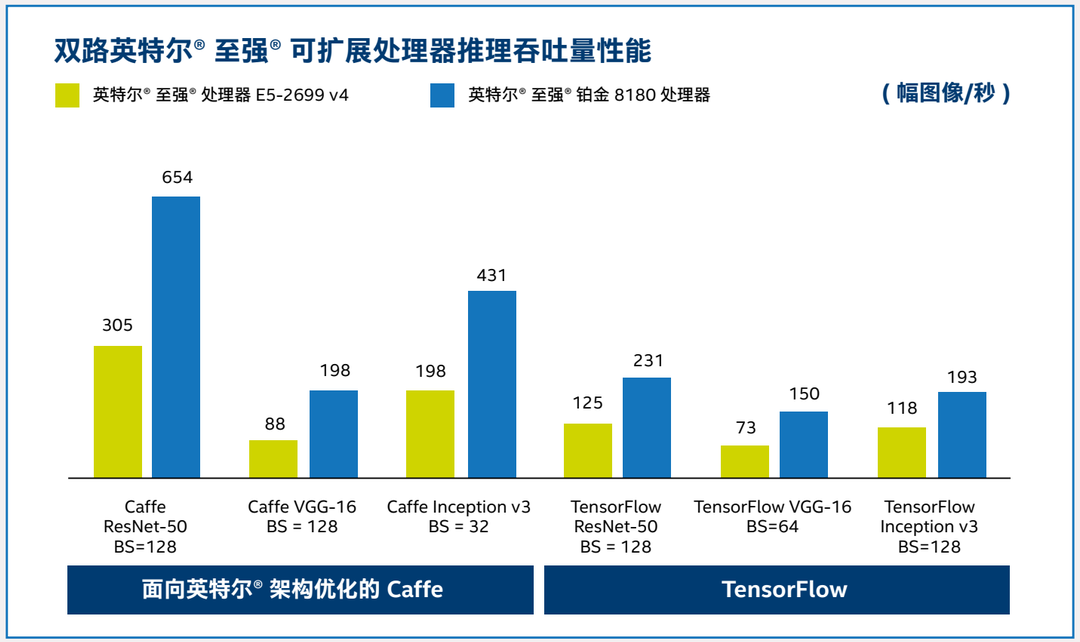

图 5：使用 ResNet-50、VGG-16 和 Inception-v3 及各种 mini-batch sizes (BS) 时，在英特尔® 至强® 处理器 v4 (原代号 Broadwell) h (浅蓝色) 和英特尔® 至强® 可扩展处理器 (原代号 Skylake) j (深蓝色) 上运行面向英特尔® 架构优化的 Caffe 和 TensorFlow 的推理 (Inference) 吞吐量。使用 fp32 精度计算推理内容。降低精度可提高性能。

通过使用针对英特尔® 至强® 技术进行了优化的框架，您可以从英特尔在性能改进方面的大量投资中受益，从一开始就在人工智能项目中取得更好的成果。通过升级到新一代的英特尔® 架构来不断更新基础设施，针对新型处理器技术优化所带来的性能改进会让您受益匪浅。

框架	用例	编程语言	优点	缺点
TensorFlow*	使用数据流图进行计算，实现可扩展的机器学习。 支持卷积神经网络和递归神经网络	Python C++	• 文档齐全 • 使用计算图抽象 • 通过 TensorBoard 提供可视化功能，简化程序的理解、调试和优化过程	• 速度可能较慢 • 预训练模型较少 • 不是完全开源的
Theano*	数值计算库。 允许用户基于数组（arrays）和张量（tensors）定义、优化和评估数学表达式。	Python	• 针对 CPU 和 GPU 进行了优化 • 可高效处理数字任务 • 更高层次的分拆框架 • 提供大量示例代码和教程	• 功能不及其他库多 • 难以使用 — 错误信息含义模糊 • 2017 年 11 月后没有什么重大进展
Caffe*	快速、开放的深度学习框架。可高效构建用于图像分类的卷积神经网络（CNN）。	C++	• 无需编写代码即可训练模型 • 高性能 • 提供 Python 和 MATLAB 绑定	• 不适用于递归神经网络 • 新架构性能低下 • 各方面都不出色
Caffe2*	Facebook 创建的开源深度学习框架。	Python C++	• 专为表达式、高速度和模块化而构建 • 能够训练大型机器学习模型并在移动设备上提供人工智能	• 与其他一些框架相比支持力度不足，服务和功能也较少
Torch*	科学和数字运算。深度学习研究。递归神经网络和卷积神经网络建模。	C	• 易于使用的模块化前端 • 快捷、高效 • 大型生态系统，包含许多由社区推动的软件包，提供许多预训练模型	• 文档质量不高 • 可立即使用的即插即用代码较少 • 基于 Lua
Microsoft CNTK*	计算网络、学习算法和模型描述	C++	• 灵活且相对较快 • 允许分布式训练 • 支持 C++、C#、Python 和 Java • 支持 GPU	• 用网络描述语言（Network Description Language (NDL)：一种新型语言）实现 • 缺少可视化功能
Apache Spark* MLlib	可扩展的机器学习库。 分类、回归、聚类。 处理海量数据。	Scala	• 可在 Java、Scala、Python 和 R 中使用 • 快速处理海量数据	• 入门很难 • 即插即用功能仅适用于 Hadoop
MXnet*	具有自动并行化功能的现代深度学习框架。 可帮助新手设计、训练和复用深度神经网络。	Python	• 易于在 AWS 云上部署大规模深度学习 • 预定义的 CloudFormation 模板 • 支持从 Tensor Flow 迁移 • 支持众多编程语言	• 用户群较小 • 调试比较困难
Neon*	专为现代深度神经网络（包括递归神经网络和卷积神经网络，以及长期/短期记忆模型和深度自动编码器）设计的深度学习框架		• 易于使用 • 可在 AlexNet、Visual Geometry Group（VGG）和 GoogLeNet 上扩展 • 与英特尔 GPU 内核库紧密集成 • 快速	• 无法配置为使用多个 CPU/GPU 线程

表 1. 选择人工智能框架时的考量因素[12]

技术考量因素

3. 存储和大数据获取

机器学习和深度学习算法需要很高的容量来存储训练所需的数据，因此在开始人工智能之旅前，您的存储系统是一个需要特别关注的方面。不管您将机器学习作为业务流程的一部分还是打算为客户运行算法，基本上肯定都要处理不同来源的海量数据：需要汇总和处理结构化数据及非结构化数据，并且将流数据与现有数据汇集在一起。

根据您的业务所侧重的人工智能用例，您可能需要保留呈指数级增长的数据量 — 毕竟在分析工作完成前，我们不清楚哪些才是关键信息。要为您的特定人工智能用例构建正确的存储环境：

- 花些时间确定您的人工智能用例的存储需求。与所有人工智能项目利益相关者合作，了解以下几点：您的机器学习算法需要哪些训练数据？数据来自哪里？如何汇总数据？需要创建数据湖吗？需要以同样快的速度访问所有这些存储吗？
- 在现有基础设施中确定可构成解决方案一部分的存储资源，并确定有效提供人工智能工作负载所需的新存储的类型和容量。
- 弄清为了以最低的总体拥有成本最高效地运行机器学习算法，是否需要对您的存储资源进行分层以及如何分层。如果需要对存储资源分层，则需要调查自动分层方法，以减轻机器学习可能带来的存储管理负担。

4. 性能与速度

快速、高效的处理对所有人工智能实例而言极其重要，但根据具体活动，可以采用多种方法满足这些需求。京东 JD.com 通过切换到 BigDL，在英特尔® 至强® 集群上使用 Apache Spark* 进行分布式深度学习，将图像检测和提取解决方案的运行速度提高了三倍以上。[13]

为了实现人工智能所需的速度和性能，关键是要确保您的处理引擎能够快速访问存储和联网接口。I/O 可能是数据中心内的关键瓶颈，特别是在信息持续流动的人工智能实例中。确保您的 I/O 硬件能够满足您要运行的人工智能工作负载的需求。

京东借助英特尔® 至强® 处理器加快图像检测速度

京东 JD.com 将其先前基于 GPU 的图像检测和提取解决方案升级为 BigDL，以便在英特尔® 至强® 集群上使用 Apache Spark 进行分布式深度学习。与之前采用 GPU 的解决方案相比，此次迁移将系统性能提高了约 3.83 倍[1]。

了解更多信息

1 平台：双路英特尔® 至强® 铂金 8180 CPU @ 2.50GHz（28 核），禁用超线程（HT），禁用睿频，通过 intel_pstate 驱动程序将扩展调节器设置为“performance”，384GB DDR4-2666 ECC RAM。CentOS Linux 版本 7.3.1611 (Core)，Linux 内核 3.10.0-514.10.2.el7.x86_64。固态盘：英特尔® 固态盘 DC S3700 系列（800GB，2.5 英寸 SATA 6Gb/秒，25 纳米，MLC）。

美团在英特尔的帮助下推出人工智能云服务

与许多云服务提供商一样，美团使用 GPU 运行图像分类和人工智能工作负载。通过切换到基于英特尔® 至强融核™ 处理器的服务器，实现数据中心扩展，美团提高了性能并降低了硬件成本。英特尔与美团已密切合作多年，指导该公司完成迁移，包括共同开发软件并提供支持，以优化其性能。**了解更多信息**

中国银联利用神经网络改善风险控制

中国银联是一家国际金融机构，从事银行服务和支付系统服务，每年在移动、在线和社交媒体等新兴渠道中处理高达 200 亿次支付。中国银联已从基于规则的模型转移到基于 Apache Spark* 并由英特尔技术提供支持的神经网络模型，从而借助人工智能提高了风险控制的准确性。**了解更多信息**

技术考量因素

5. 硬件同质性

更快的计算无疑让人工智能和机器学习工作负载受益匪浅，例如将一部分处理任务分流到现场可编程门阵列（FPGA）或 GPU 上。[14]FPGA 通常用于实现神经网络，因为它们可以在同一设备的计算、逻辑和存储器资源中处理不同的算法。通过将运算硬编码运行到硬件中，可实现更快的性能[15]。

但是，加速技术可能会大幅削弱计算的同质性，加重云服务提供商在管理、维护和保障整个系统安全方面的工作量。确保硬件具有同质性，而不必为加速器分别维护单独的集群，可大幅简化人工智能工作负载的运行并带来诸多优势，例如简化管理、程序修补、修复和更换工作，实现更直接的安全控制。

6. 新的人工智能技术

人工智能领域的技术正在飞速发展，应对机器学习和深度学习技术不断带来的新挑战的硬件和软件也层出不穷。去年，英特尔宣布将推出英特尔® Nervana™ 神经网络处理器（英特尔® Nervana™ NNP），旨在满足人工智能用例对速度、数值并行性和可扩展性的需求。该架构采用全新的设计，专为神经网络构建，支持所有深度学习原语，同时确保核心组件尽可能高效。英特尔® Nervana™ NNP 等技术可帮助投资人工智能的组织提高人工智能系统的功能和性能，同时简化部署工作。时刻关注人工智能技术领域的最新趋势可确保您快速改进人工智能项目和产品，加快产品上市速度，更快地为客户部署新服务，帮助您在激烈的竞争中脱颖而出。

结论和后续行动

人工智能领域为云服务提供商提供了很多机会，但能否抓住这些机会取决于云服务提供商是否部署了正确的基础设施来支持这些新的工作负载类型。对云服务提供商来说，根据现有的专业知识构建自己的人工智能功能和服务，从而取得客户信任并推动早期成功可能十分重要。如果云服务提供商有意通过人工智能拓展业务，可采取以下后续行动：

研究

收集有关客户及竞争对手如何运用人工智能技术的创意和最佳实践 — 您可以从案例研究开始，并深入阅读英特尔在人工智能领域的投入和面向开发人员的英特尔® AI Academy。

组建自己的团队

确保所有业务职能部门的关键利益相关者都参与到您的人工智能项目中。要让人工智能项目取得成功，产品/服务开发、业务负责人、IT、开发人员、运营和领导层都需要付出相应的努力。

评估现有 IT 系统

使用上面的“技术考量因素”部分作为确定基础设施和流程适用性的起点，发现可以升级或改进的环节。

适用于云服务提供商的其他资源

访问 intel.cn/csp

英特尔在人工智能领域的投入

英特尔® 人工智能产品组合针对各种业务需求提供了广泛适用和灵活的解决方案。

图 6：英特尔® 人工智能产品组合

深入阅读

- 美团基于 IA 构建人工智能：从改善客户体验到打造差异化云服务
- 英特尔® AI Academy：帮助开发人员构建人工智能解决方案
- 面向人工智能专家的英特尔资源
- 面向云服务提供商的英特尔资源

1 McKinsey (2017) Artificial Intelligence: The Next Digital Frontier?（麦肯锡（2017 年）人工智能: 下一个数字前沿？）https://www.mckinsey.com/~/media/McKinsey/Industries/Advanced%20Electronics/Our%20Insights/How%20artificial%20intelligence%20can%20deliver%20real%20value%20to%20companies/MGI-Artificial-Intelligence-Discussion-paper.ashx

2 McKinsey (2017) Artificial Intelligence: The Next Digital Frontier?（麦肯锡（2017 年）人工智能: 下一个数字前沿？）https://www.mckinsey.com/~/media/McKinsey/Industries/Advanced%20Electronics/Our%20Insights/How%20artificial%20intelligence%20can%20deliver%20real%20value%20to%20companies/MGI-Artificial-Intelligence-Discussion-paper.ashx

3 Forbes (2017) '10 Predictions for AI, Big Data and Analytics in 2018'（福布斯（2017 年）“关于 2018 年人工智能、大数据和数据分析的十大预测”）

4 Teradata (2017) 'State of Artificial Intelligence for Enterprises'（Teradata（2017 年）“企业对人工智能的接受程度”）

5 Teradata (2017) State of Artificial Intelligence for Enterprises（Teradata（2017 年）“企业对人工智能的接受程度”）, http://assets.teradata.com/resourceCenter/downloads/AnalystReports/Teradata_Report_AI.pdf

6 Fast Company (2017) 'How Amazon, Google, Microsoft, And IBM Sell AI As A Service（Fast Company（2017 年）“Amazon、Google、Microsoft 和 IBM 如何销售人工智能云服务产品）, https://www.fastcompany.com/40474593/how-amazon-google-microsoft-and-ibm-sell-ai-as-a-service

7 英特尔® 至强® 铂金 8180 处理器与英特尔® 至强® 处理器 E5-2699 v4 比较。注: 与采用 BVLC-Caffe 平台的英特尔® 至强® 处理器 E5-2699 v3 相比, 使用优化后的框架和英特尔® MKL 库, 性能在过去两年提升 113 倍, 配置为: 双路英特尔® 至强® 铂金 8180 处理器 CPU @ 2.50GHz（28 核）, 禁用超线程 (HT), 禁用睿频, 通过 intel_pstate 驱动程序将扩展调节器设定为“performance”, 384GB DDR4-2666 ECC RAM。CentOS* Linux 版本 7.3.1611 (Core), Linux 内核 3.10.0-514.10.2.el7.x86_64。固态盘: 英特尔® 固态盘 DC S3700 系列（800GB, 2.5 英寸 SATA 6Gb/秒, 25 纳米, MLC）。性能评测标准基于: 环境变量: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=56, CPU 频率设置为 cpupower frequency-set -d 2.5G -u 3.8G -g performance

深度学习框架: Caffe*:（http://github.com/intel/caffe/）, 修订版 f96b759f71b2281835f690af267158b82b150b5c。推理能力的测量基于“caffe time --forward_only”命令, 训练能力的测量基于“caffe time”命令。对于“ConvNet”拓扑, 使用虚拟数据集。对于其他拓扑, 数据在本地存储, 并且在训练之前在内存中缓存。拓扑规范来源于 https://github.com/intel/caffe/tree/master/models/intel_optimized_models（GoogLeNet、AlexNet 和 ResNet-50）、https://github.com/intel/caffe/tree/master/models/default_vgg_19（VGG-19）和 https://github.com/soumith/convnet-benchmarks/tree/master/caffe/imagenet_winners（ConvNet 基准测试; 更新文件以使用最新的 Caffe prototxt 格式, 但功能等效）。英特尔 C++ 编译器版本 17.0.2 20170213, 英特尔® MKL 小型库版本 2018.0.20170425。使用“numactl -l”运行 Caffe。

平台配置: 双路英特尔® 至强® CPU E5-2697 v2（2.70GHz, 12 核）, 启用超线程 (HT), 启用睿频, 通过 intel_pstate 驱动程序将扩展调节器设定为“performance”, 256GB DDR3-1600 ECC RAM。CentOS Linux 版本 7.3.1611 (Core), Linux 内核 3.10.0-514.21.1.el7.x86_64。固态盘: 英特尔® 固态盘 520 系列, 240GB, 2.5 英寸 SATA 6Gb/秒, 25 纳米, MLC。

性能的测量基于: 环境变量: KMP_AFFINITY='granularity=fine, compact,1,0', OMP_NUM_THREADS=24, CPU 频率设置为 cpupower frequency-set -d 2.7G -u 3.5G -g performance

深度学习框架: Caffe*:（http://github.com/intel/caffe/）, 修订版 b0ef3236528a2c7d2988f249d347d5fdae831236。推理能力的测量基于“caffe time --forward_only”命令, 训练能力的测量基于“caffe time”命令。对于“ConvNet”拓扑, 使用虚拟数据集。对于其他拓扑, 数据在本地存储, 并且在训练之前在内存中缓存。拓扑规范来源于 https://github.com/intel/caffe/tree/master/models/intel_optimized_models（GoogLeNet、AlexNet 和 ResNet-50）、https://github.com/intel/caffe/tree/master/models/default_vgg_19（VGG-19）和 https://github.com/soumith/convnet-benchmarks/tree/master/caffe/imagenet_winners（ConvNet 基准测试; 更新文件以使用最新的 Caffe prototxt 格式, 但功能等效）。GCC 4.8.5, 英特尔® MKL 小型库版本 2017.0.2.20170110。

8 4 年估算总体拥有成本（TCO）最高降低 65% 的示例: 根据使用 VMware ESXi* 虚拟整合工作负载的等效机架性能, 对比已安装的 20 台搭载英特尔® 至强® 处理器 E5-2690（前身为“Sandy Bridge-EP”）的双路服务器, 并使用 Guest OS RHEL* 6.4 运行 VMware ESXi* 6.0 Ga; 总成本达 919,362 美元。与之相比, 5 个全新英特尔® 至强® 奔腾 8180 (Skylake) 使用 64 位 Guest OS RHEL* 6 运行 VMware ESXi* 6.0 U3 GA 的总成本为 320,879 美元（包括基本购置成本）。假设服务器价格基于当前 OEM 为双路服务器（搭载英特尔® 至强® 处理器 E5-2690 v4）和四路服务器（搭载 2 个 E7-8890 v4）发布的零售价格, 并且可能会根据所提供系统的实际价格发生变化。

9 Intel (2017) TensorFlow* Optimizations on Modern Intel® Architecture（英特尔（2017 年）— 英特尔® 现代架构上的 TensorFlow* 优化）, https://software.intel.com/zh-cn/articles/tensorflow-optimizations-on-modern-intel-architecture

10 平台: 双路英特尔® 至强® 铂金 8180 CPU @ 2.50GHz (28 核), 禁用超线程 (HT), 禁用睿频, 通过 intel_pstate 驱动程序将扩展调节器设置为“performance”, 384GB DDR4-2666 ECC RAM。CentOS Linux 版本 7.3.1611 (Core), Linux 内核 3.10.0-514.10.2.el7.x86_64。固态盘: 英特尔® 固态盘 DC S3700 系列 (800GB, 2.5 英寸 SATA 6Gb/秒, 25 纳米, MLC)。

性能评测标准基于: 环境变量: KMP_AFFINITY='granularity=fine, compact‘, OMP_NUM_THREADS=56, CPU 频率设置为 cpupower frequency-set -d 2.5G -u 3.8G -g performance

Caffe: (http://github.com/intel/caffe/), 修订版 f96b759f71b2281835f690af267158b82b150b5c。推理能力的测量基于“caffe time --forward_only”命令, 训练能力的测量基于“caffe time”命令。对于“ConvNet”拓扑, 使用虚拟数据集。对于其他拓扑, 数据在本地存储, 并且在训练之前在内存中缓存。拓扑规范来源于 https://github.com/intel/caffe/tree/master/models/intel_optimized_models (GoogLeNet、AlexNet 和 ResNet-50)、https://github.com/intel/caffe/tree/master/models/default_vgg_19 (VGG-19) 和 https://github.com/soumith/convnet-benchmarks/tree/master/caffe/imagenet_winners (ConvNet 基准测试; 更新文件以使用最新的 Caffe prototxt 格式, 但功能等效)。英特尔 C++ 编译器版本 17.0.2 20170213, 英特尔® MKL 小型库版本 2018.0.20170425。使用“numactl -l”运行 Caffe。

TensorFlow: (https://github.com/tensorflow/tensorflow), 提交ID 207203253b6f8ea5e938a512798429f91d5b4e7e。使用虚拟数据进行三种 convnet 基准测试 (alexnet、googlenetv1、vgg, https://github.com/soumith/convnet-benchmarks/tree/master/tensorflow) 得到的性能数据。GCC 4.8.5; 英特尔® MKL 小型库版本 2018.0.20170425; 对于 alexnet、vgg 基准测试, 互操作并行性线程设置为 1, 对于 googlenet 基准测试, 互操作并行性线程设置为 2; 内部操作并行性线程设置为 56; 使用的数据格式为 NCHW; 对于 googlenet 和 vgg 基准测试, KMP_BLOCKTIME 设置为 1, 对于 alexnet 基准测试, KMP_BLOCKTIME 设置为 30。推理能力的测量基于“--caffe time -forward_only -engine MKL2017”选项, 训练能力的测量基于“--forward_backward_only”选项。

MxNet: (https://github.com/dmlc/mxnet/), 修订版 5efd91a71f36fea483e882b0358c8d46b5a7aa20。使用虚拟数据。推理能力使用“benchmark_score.py”测得, 训练能力使用修改版“benchmark_score.py”测得, 后者也运行后向传播。拓扑规范来源于 https://github.com/dmlc/mxnet/tree/master/example/image-classification/symbols。GCC 4.8.5, 英特尔® MKL 小型库版本 2018.0.20170425。

Neon: 内部版本。使用虚拟数据。main.py 脚本用于 mkl 模式下的基准测试。使用的 ICC 版本: 17.0.3 20170404, 英特尔® MKL 小型库版本 2018.0.20170425。最新版本的 Neon 结果: https://www.intelnervana.com/neon-v2-3-0-significant-performance-boost-for-deep-speech-2-and-vgg-models/

11 平台配置: 双路英特尔® 至强® CPU E5-2699 v3 @ 2.30GHz (18 核), 启用超线程 (HT), 禁用睿频, 通过 intel_pstate 驱动程序将扩展调节器设定为“performance”, 256GB DDR4-2133 ECC RAM。CentOS Linux 版本 7.3.1611 (Core), Linux 内核 3.10.0-514.el7.x86_64。操作系统驱动器: Seagate* Enterprise ST2000NX0253 2 TB 2.5 英寸内置硬盘。

性能的测量基于: 环境变量: KMP_AFFINITY='granularity=fine, compact,1,0‘, OMP_NUM_THREADS=36, CPU 频率设置为 cpupower frequency-set -d 2.3G -u 2.3G -g performance

英特尔 Caffe: (http://github.com/intel/caffe/), 修订版 b0ef3236528a2c7d2988f249d347d5fdae831236。推理能力的测量基于“caffe time --forward_only”命令, 训练能力的测量基于“caffe time”命令。对于“ConvNet”拓扑, 使用虚拟数据集。对于其他拓扑, 数据在本地存储, 并且在训练之前在内存中缓存。拓扑规范来源于 https://github.com/intel/caffe/tree/master/models/intel_optimized_models (GoogLeNet、AlexNet 和 ResNet-50)、https://github.com/intel/caffe/tree/master/models/default_vgg_19 (VGG-19) 和 https://github.com/soumith/convnet-benchmarks/tree/master/caffe/imagenet_winners (ConvNet 基准测试; 更新文件以使用最新的 Caffe prototxt 格式, 但功能等效)。GCC 4.8.5, MKLML 版本 2017.0.2.20170110。

BVLC-Caffe: https://github.com/BVLC/caffe, 推理和训练能力测量基于“caffe time”命令。对于“ConvNet”拓扑, 使用虚拟数据集。对于其他拓扑, 数据在本地存储, 并且在训练之前在内存中缓存。BVLC Caffe (http://github.com/BVLC/caffe), 修订版 91b09280f5233cafc62954c98ce8bc4c204e7475 (提交日期 5/14/2017)。BLAS: atlas 版本 3.10.1。

12 来源: DZone (2018) 10 Best Frameworks and Libraries for AI (DZone (2018 年) 十佳人工智能框架和库) https://dzone.com/articles/progressive-tools10-best-frameworks-and-libraries, Towards Data Science (2017) A Survey of Deep Learning Frameworks (Towards Data Science (2017 年) 深度学习框架调查) https://towardsdatascience.com/a-survey-of-deep-learning-frameworks-43b88b11af34, Bosch Research and Technology Center (2016) Comparative Study of Deep Learning Software Frameworks (博世研究与技术中心 (2016 年) 深度学习软件框架比较研究) https://arxiv.org/pdf/1511.06435.pdf, Microway (2016) Deep Learning Frameworks: A Survey of TensorFlow, Torch, Theano, Caffe, Neon and IBM< Machine Learning Stack (Microway (2016 年) 深度学习框架: 一项关于 TensorFlow、Torch、Theano、Caffe、Neon 和 IBM《机器学习堆栈》的调查) https://www.microway.com/hpc-tech-tips/deep-learning-frameworks-survey-tensorflow-torch-theano-caffe-neon-ibm-machine-learning-stack/, Liu & Zang (2017) Caffe2 vs TensorFlow: Which is the better deep learning framework? (Liu & Zang (2017 年) Caffe2 与 TensorFlow: 谁是更好的深度学习框架?) http://cs242.stanford.edu/assets/projects/2017/liubaige-xzang.pdf

13 GPU: 20 * NVIDIA Tesla* K40。CPU: 英特尔® 至强® 处理器 E5-2650 v4 @ 2.20GHz, 1200 个逻辑内核 (每台服务器有 24 个物理内核), 启用英特尔® 超线程技术 (英特尔® HT 技术), 并且在 Yet Another Resource Negotiator (YARN) 中配置为支持 50 个逻辑内核。https://software.intel.com/en-us/articles/building-large-scale-image-feature-extraction-with-bigdl-at-jdcom

14 IDC (2017 年) — IDC 称, 在企业寻求管理认知工作负载的当下, 加速计算得到飞速发展, https://www.idc.com/getdoc.jsp?containerId=prUS42909117

15 英特尔 FPGA, https://www.altera.com/solutions/industry/computer-and-storage/applications/machine-learning/machine-learning.html

此处提供的所有信息可在不通知的情况下随时发生变更。关于英特尔最新的产品规格和路线图, 请联系您的英特尔代表。

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。

基准性能测试在实施近期针对“Spectre”和“Meltdown”漏洞的软件补丁和固件更新之前发布。实施更新后, 这些结果可能不再适用于您的设备或系统。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。没有计算机系统是绝对安全的。请查询您的系统原始设备制造商或零售商, 或访问 www.intel.cn 了解更多信息。

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。没有计算机系统是绝对安全的。请查询您的系统原始设备制造商或零售商, 或访问 www.intel.cn 了解更多信息。

0318/CAT/XX/PDF