白皮书

数据中心
人工智能

为您的人工智能解决方案选择最佳基础设施策略

购买、构建、再利用或外包数据中心资源，实现图像识别、自然语言处理或预测性维护工作负载。

目录

简介

对于许多组织而言，问题不在于是否部署人工智能 (AI)，而是何时以及最重要的是如何部署人工智能？随着 IT 战略的重心从数据管理移向智能操作，企业越来越深刻地领略到人工智能在帮助人们解决问题、制定决策以及开展创新方面的意义。人工智能系统从复杂、结构化和非结构化的海量数据中获取信息，并将其转化为切实可行的洞察。

企业认识到，要想在竞争激烈的环境中保持持续增长，实施和使用人工智能将起到决定性的作用。这方面存在许多潜在机遇，包括：

- 一次性的定制机器学习解决方案，用于解决特定问题
- 普及化的解决方案，从进行预测到抢占先机，推动企业做出更明智的决策
- 有机会使用人工智能来推动创新、建立关联、识别新的发展机遇并转化为收入

要充分利用这些全新且激动人心的人工智能机遇，首先要考虑的问题之一就是选择合适的基础设施。人工智能解决方案常常需要新的硬件和软件，例如，在新数据变得可用时，利用可扩展处理或创建和微调模型，对数据源进行校正和标注。对于任何给定的人工智能解决方案，可以做出的选择包括：

- 再利用现有的硬件，以最低的成本提供人工智能解决方案
- 购买一次性人工智能解决方案，只用于满足特定应用案例的需求
- 构建一个更广泛的平台，支持多个人工智能解决方案的需求
- 将人工智能解决方案交付外包给第三方资源，包括公有云

人工智能和它所代表的机遇正处于快速演变时期，所以预算决策者可能不愿意在这方面投入太多。举例来说，如果企业内部缺乏相关的专业知识，可能对解决方案的交付造成危害，如果发生错误或延迟，则可能面临声誉受损的风险。而一旦对人工智能的作用产生怀疑，就会造成相当大的阻碍，导致人工智能无法发挥全部价值。

编写本指南的目的就是为了应对这些挑战，可帮助决策者选择合适的人工智能基础设施方案，以加快人工智能部署，在不产生额外成本或形成更长期问题的情况下获得相关经验。在以下各部分，我们将介绍每种选择的优缺点。

再利用现有硬件

准备采用人工智能的企业通常都希望利用数据中心的“空闲时段”来运行人工智能工作负载，或者基于单个“空闲”服务器、工作站节点或者小型集群来自行开发解决方案。这样做让您能够：

- 检验自己的想法，看看哪些可在您的组织中实施
- 研究硬件和软件选项
- 在真实场景中培养技能，积累经验
- 吸引业务部门参与，让他们了解人工智能的好处

对于再利用的硬件，具体的配置会因场景不同而有所变化：在下表中，我们给出了一种可能的配置，用作实施深度学习训练和测试的基础（使用英特尔® Caffe* 优化版）。

项目	型号/版本
硬件	
英特尔® 服务器系统	R1208WT
英特尔® 服务器主板	S2600WT
(2x) 英特尔® 至强® 可扩展处理器	英特尔® 至强® 金牌 6148 处理器
(6x) Crucial* 32GB LRDIMM DDR4	CT32G4LFD4266
(1x) 英特尔® 固态盘 1.2TB	S3520
软件	
CentOS Linux* 安装光盘（DVD）	7.3.1611
英特尔® Parallel Studio XE 集群版	2017.4
英特尔® Caffe* 发行版	MKL2017
英特尔® 机器学习扩展库（Linux* OS）	2017.1.016

再利用现有硬件来满足人工智能需求具有以下**优势**：

- 因为是利用现有的硬件资源，所以再利用解决方案的采购成本低，而且有望在最短时间内完成搭建。
- 单节点或小型集群解决方案的范围很小，因此研究工作能更好地侧重于紧凑环境，将精力放在人工智能本身上，而不是网络带宽或运营管理等问题上。
- 再利用方法提供了在基础设施中使用空闲处理时段的机会，因此进一步强化了人工智能“增强”现有功能的这一优势。

再利用现有硬件来满足人工智能需求有以下**缺点**：

- 一次性的、最小化的解决方案并非始终能够与更广泛的解决方案或面向用户的工具轻松集成，因此限制了其适用性、适用范围和使用寿命。这种限制也可能导致出现多个技术“孤岛”。
- 除非现有的硬件直接符合需求，否则您会因为转化或调整不太适合的资源而增加额外开销。
- 如果管控不到位，测试配置可能变成现场使用的配置，并因此产生更多花费和风险，例如，如果您“借用”资源来测试一种需求，之后您需要归还所借的资源。

这个方法适合您吗？

根据上文所述，再利用现有硬件被视为一种有用的短期选项。有些组织使用的服务器注定是数据中心更新换代的一部分（这些通常都是整批采购，并在之后分阶段逐步安装）。您可以将这个选项当作说服预算负责人的方法，但是同样的，即使您的目标是开发一个短期解决方案，也要将长期计划谨记在心。

购买一次性解决方案

与我们沟通过的许多组织都在考虑购买定制解决方案这一选项，以满足特定的应用案例需求。实际情况通常是，当决策者确定某个明确的需求时，他们可能就不太会考虑一次性解决方案以外的选项，也就不会去考虑在整个企业内更有战略性地使用人工智能。出现这种情况有几个原因，尤其是构建一个广泛采用人工智能的业务案例要比针对特定需求来构建业务案例复杂得多。

图 1 显示了根据明确需求（在本例中为预测性维护）来提供服务的解决方案架构，因此可以当作定制的人工智能解决方案来进行购买。

本示例基于采用英特尔® 至强® 处理器的服务器构建，特别适合基于推理的人工智能模型使用。信息输入由广泛的数据源和传感器提供，来自于现有系统和终端（例如，制造设备、车辆或建筑数据）。起支撑作用的框架和软件能够提供训练和推理功能：英特尔致力于确保所有主要的深度学习框架和拓扑都能在英特尔® 架构上良好运行。它也会通过基于 Web 的 API，为库存管理软件和可视化技术工具等提供信息。

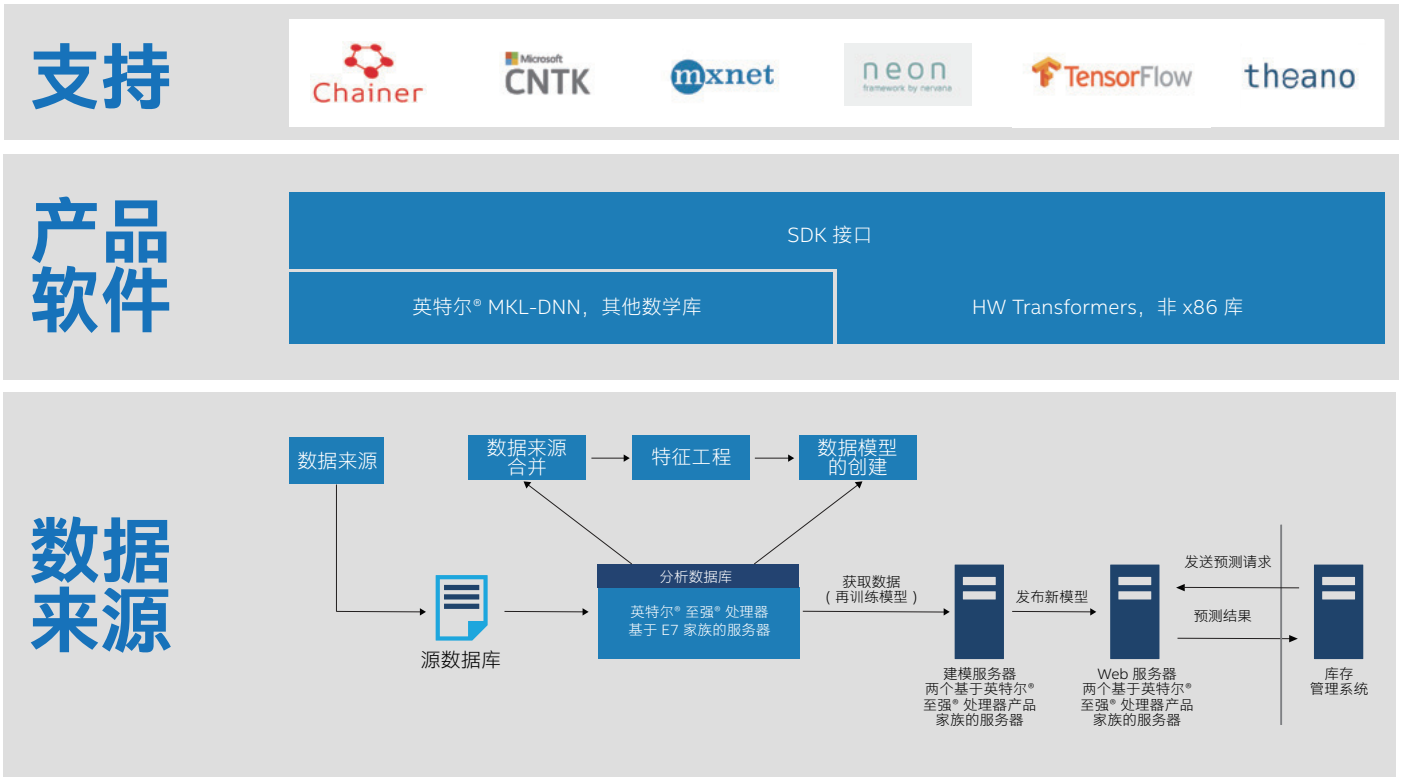

图 1. 适用于高性能人工智能应用案例（如预测性维护）的硬件和软件架构

为特定的应用案例购买人工智能解决方案有以下**优势**：

- 比起更加一般化的解决方案，现成的方法能够更快的部署和采用，因为只需要支持相关的业务部门和利益相关者团体的需求。
- 涵盖范围更小意味着更容易学习有关人工智能的技能和专业知识，因为较小的团队可以集中精力优化单个解决方案，之后再考虑更广泛的优化和扩展问题。
- 对于特定的应用案例，一次性解决方案可能是最佳选项，因为相比于那些为多个共享资源的场景设计的解决方案，一次性解决方案能为特定应用案例提供更高的效率和性能。
- 关于硬件成本，一次性解决方案的价格可能比用于支持更广泛应用案例的架构的价格更便宜。

为特定的应用案例购买人工智能解决方案有以下**缺点**：

- 如果解决方案的特定元素不随相关场景需求的变化而变化，则选择的解决方案可能会过时。
- 一次性方法可能导致多个人工智能“孤岛”，解决方案之间彼此孤立，因此需要并行开发和管理。
- 就单个解决方案而论，价格可能更便宜，但是按整体计算，则可能更昂贵，例如，当组织需要重复开发技术架构时便是如此。

这个方法适合您吗？

总体而言，购买定制人工智能解决方案的优势在于速度快和有针对性，但劣势在于成本可能更高，且有可能造成不同场景间出现人工智能“孤岛”，为此，需要在这两者之间进行权衡。当组织在首次尝试人工智能并取得成效之后，决定大范围采用人工智能时，就可能出现上述情况。

要确定此方法是否适合您，不妨考虑一下，继此场景之后，贵组织有多大可能会更广泛地采用人工智能。如果您已经部署一次性解决方案，或者将来可能部署，或许您应该考虑一下构建更广泛平台的优势。但是，如果您尚在测试原先的想法，那么再利用现有硬件，或者外包部署也许是更好的选择。

构建更广泛的平台

对于那些具有更多人工智能经验或想要响应多个业务领域需求的组织，可能会考虑采用更广泛的基础设施解决方案，以支持更普遍的人工智能工作负载。此方法与 IT 行业如今盛行的新兴“平台”

架构类似，也就是说，这种方法提供高度可扩展、可作为单池管理的基础设施层，可在服务器计算、存储和网络等方面使用虚拟化技术和软件定义技术。

在人工智能方面，此平台可与多种开源和商业软件包配合使用，经过配置后可满足单个工作负载的需求。图 2 显示这种架构如何基于以下各层为特定场景（在本例中是面部识别）提供支持：

- 硬件：包含计算、输入/输出（I/O）和辅助处理节点，以及可扩展的存储和网络连接。设备和系统之间的通信以超高速度的骨干为基础，如英特尔® Omni-Path 高带宽网络（英特尔® OP Fabric）。
- 软件：由操作系统和虚拟化层构成，人工智能特定的模块库可在其上运行，实现算法处理和分析、数据管理和 I/O，以及获取和传输数据源及虚拟化分析结果。
- 流程：其中涉及人工智能应用的“业务逻辑”，利用库模块来提供诸如面部识别等功能。此流程层考虑到了学习算法的训练及结果的评估/推理。

构建更广泛的平台来满足人工智能需求有以下**优势**：

- 基于平台的方法能够提供统一的配置点和唯一的部署目标。因此，能够通过降低管理费用，进一步削减日常使用人工智能的成本。
- 从技能和经验角度来看，虽然平台可能比单个解决方案更复杂，但它有利于构建专业知识。
- 从组织角度来看，平台可由单个团队而非多个团队管理，从而加强组织内部及与各业务部门之间的沟通。

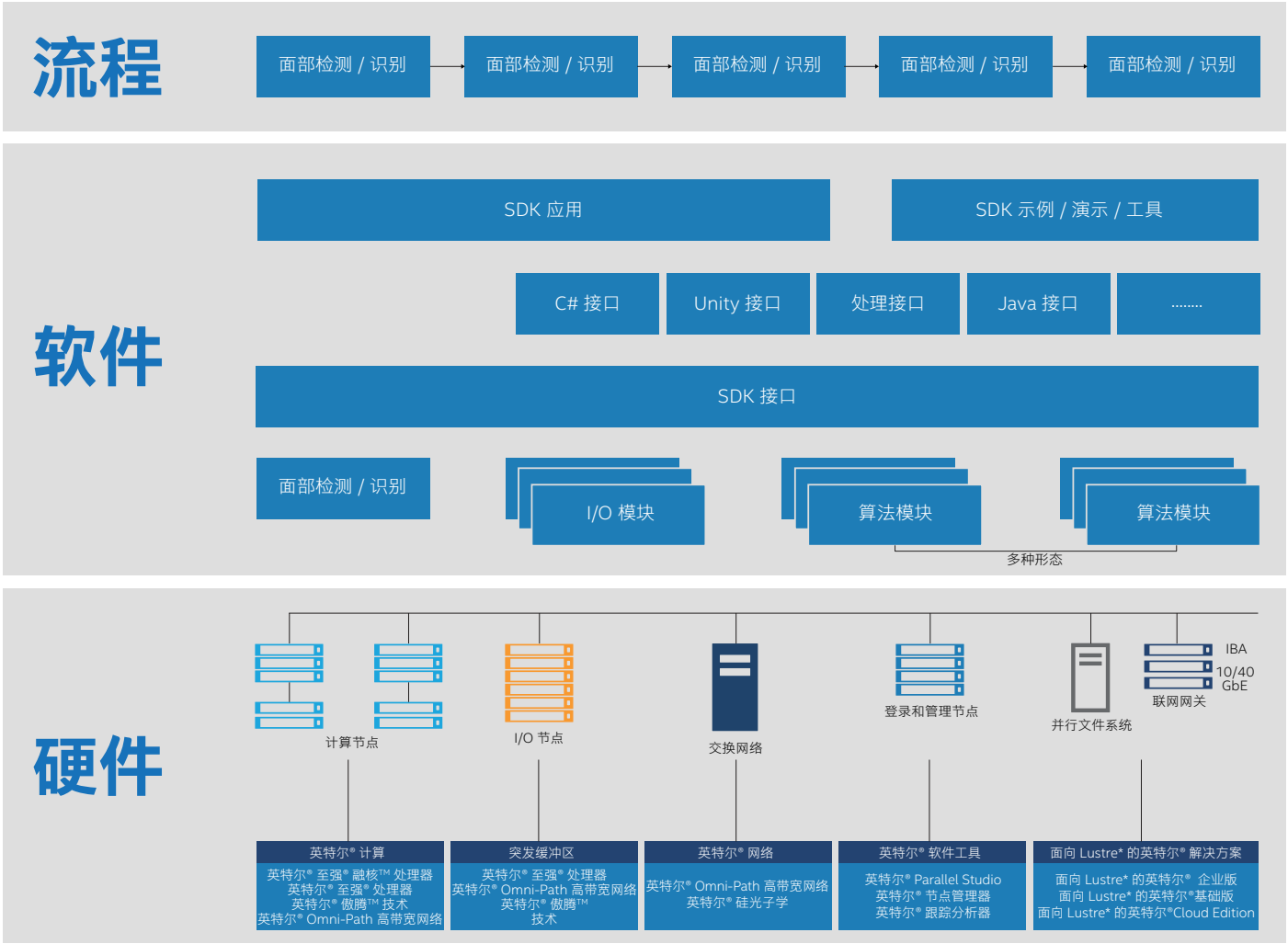

图 2. 适用于广泛的人工智能应用案例的硬件和软件架构

构建更广泛的平台来满足人工智能需求有以下**缺点**：

- 比起一次性解决方案，首次构建人工智能平台的过程可能看起来更加复杂，成本也更加高昂（请注意，并非一定是这样。虽然构建的平台可以面向更加广泛的用途，但最初可以部署为较小版本）。
- 如果内部有人员掌握相应的技能，则构建更广泛的平台时将从中获益，尤其是首次配置时。对于组织来说，这可能是一项挑战，可能增加部署风险。
- 如果架构不太合适，那么创建更广泛的平台也意味着可能面临更大的风险。例如，相比于实际需求，所构建的平台规模可能太小，或者太大。

这个方法适合您吗？

对于想要扩大人工智能使用范围的组织，构建更广泛的平台很有意义。对于尚在早期实施阶段的组织而言，平台的优势显而易见，但是，如果缺乏技能且/或缺乏将人工智能变成日常业务工作的一部分的意愿，那么构建更广泛的平台可能还为时尚早。

要确定此方法是否适合您的企业，您可以跨业务部门测试对人工智能的需求，例如，开展概念验证研究（基于现有或外包基础设施），借此获得经验和认同。如果结果显示，您对人工智能的需求仍然有限，那么您可以考虑采用一次性解决方案。如需进一步了解在服务器计算、存储和网络方面对人工智能平台的要求，请查看我们关于创建概念验证的白皮书。

外包解决方案的交付工作

无论处于哪个人工智能实施阶段，组织都可能想要使用第三方资源（包含基于公有云的选项）和技能，以提供全栈解决方案或利用现有资源。外包人工智能解决方案的具体元素可能包括：

- **基础设施硬件**：按使用付费的基础设施即服务（包括 CPU 和固态盘），可实现按需分配
- **人工智能专用软件：**有些提供商现在提供人工智能函数库，包括语音和图像识别
- **数据管理**：基于服务的平台可提供高度可扩展的基础，用于数据校正和交付

内部工程师可与第三方提供商合作，提供全部或部分外包的解决方案架构。

将解决方案的交付工作外包出去有以下**优势**：

- 功能“现成”可用，可最大程度减少部署和配置问题
- 对于刚开始采用人工智能的企业，预先开发的选项可帮助他们减少技能和资源开销
- 按使用付费服务的成本是可控的，对于难以确定所需资源的企业而言非常合适
- 在引入新解决方案之前，可使用外部服务来增强和测试新解决方案
- 组织可利用第三方的技能和知识让自己受益

将解决方案的交付工作外包出去有以下**缺点**：

- 由于要管理与外包商之间的关系，因此可能会增加成本和降低效率，对业务部门尤其如此，这种情况会增大推动创新的难度。
- 由此而来的基础设施架构可能带来数据瓶颈，具体取决于数据来源，例如，组织可能需要将数据从内部系统上传到云端。
- 如果解决方案的交付工作被外包给第三方，那么要积累内部经验和技能（尤其是与解决方案架构和数据科学有关的经验和技能）将会更加困难。这可能丧失获取宝贵知识的机会。
- 比起运行内部系统，使用外包资源的成本可能随时间增加。

面向英特尔® 至强® 处理器的人工智能软件优化

为了让数据科学家和开发人员能够在工作中使用他们喜欢的框架，英特尔对适用于许多广受欢迎的人工智能框架的深度学习库实施了优化，其中包括 Theano* 和 TensorFlow*。

面向深度神经网络的英特尔® 数学核心函数库（英特尔® MKL-DNN）在这些框架的底层运行，是一款新的加速器。它采用特定的数学函数来支持深度学习，并在配备英特尔® 高级矢量扩展 2（英特尔® AVX-2）和英特尔® 高级矢量扩展 512（英特尔® AVX-512）指令的 x86 机器上实施优化。作为开源项目，它将持续关注所有主流框架的新趋势。

另外，英特尔® BigDL 是面向 Spark* 的分布式深度学习库，可直接在现有的 Spark 或 Apache Hadoop* 集群上运行。它能将预先训练的 Torch* 模型载入 Spark 框架，并能有效地进行横向扩展以针对大数据级别实施数据分析。

这个方法适合您吗?

外包人工智能解决方案的入门成本较低，且基于云的方案非常适合用于开展实验和实施短期研究，这是其优势。但是，劣势在于需要付出长期成本、难以积累内部技能和经验，以及难以大规模实施。例如，如果场景的数据密集度非常高（比如从制造系统中获取信息或从零售环境中获取信息），那么使用内部资源可能更有意义。

确定最佳选项

如上述例子所示，并不存在“一体适用”的人工智能解决方案，每种解决方案都需仔细考虑以下因素:

- 类型、规模和业务模式
- 需求和应用范围
- 内部基础设施的可用性
- IT 和业务部门的技能、专业知识和经验
- 对人工智能的重视和投入程度
- 内部或外部来源的数据可用性
- 短期与长期能力规划

其中许多因素取决于组织所处的人工智能阶段。那些刚开始了解人工智能优势的组织可能会认为再利用现有硬件，或者利用云服务是快速实现价值的最简便方式。所处阶段更深入一些的组织可能考虑购买一次性人工智能解决方案，用于实现特定目标，而那些从长期考虑的组织可能认为更广泛的人工智能平台最为合适，处理大量内部数据时尤其如此。

如图 3 所示，这种进阶也与组织内部的技能和经验、用户信任度以及整体人工智能投资回报有关：我们的白皮书 The AI Readiness Model（人工智能就绪模型）更深入地探讨了这些主题。虽然更广泛的平台能在长期时间内提供最大优势，但对于尚在积累技能和建立信任度的组织而言，这种方法还是难以消化。

无论选择哪种选项，开始时最重要的一点就是了解人工智能将用于应对哪些场景。我们建议组织在购买或实施部署前，先咨询同行或专家。关于技能和知识，则还是尽早开始累积为好，例如，通过英特尔® AI Academy 开始累积。

考虑运行人工智能概念验证？请查看英特尔的“成功概念验证解析”。

再利用现有硬件	外包解决方案的交付工作	购买一次性解决方案	构建更广泛的平台
适用于以下情况: • 想要研究或测试自己的想法 • 希望获得内部人员的认同	适用于以下情况: • 希望以更低成本入门 • 主要使用外部数据源	适用于以下情况: • 希望快速部署解决方案 • 只准备部署有限的人工智能	适用于以下情况: • 具备丰富的人工智能使用经验 • 计划在多种场景下使用人工智能

图 3. 业务价值、用户信任度和总体投资回报率随成熟度增加而增加

了解更多信息

欲详细了解英特尔® 人工智能产品组合及其如何为您的人工智能之旅提供支持，请访问：www.intel.cn/ai。

有关英特尔的性能优化型机器学习和深度学习库及框架，请访问此处：https://software.intel.com/zh-cn/ai-academy

参考和资源

英特尔® AI Academy：https://software.intel.com/zh-cn/ai-academy

有关可解释人工智能的挑战和机遇，请访问：https://ai.intel.com/the-challenges-and-opportunities-of-explainable-ai/

零售业的未来在于人工智能：https://ai.intel.com/future-retail-artificial-intelligence

英特尔® AI Academy – 学习基础知识：https://software.intel.com/zh-cn/ai-academy/basics

Loihi – 英特尔的全新自学芯片承诺加快人工智能的步伐：https://newsroom.intel.cn/editorials/intels-new-self-learning-chip-promises-accelerate-artificial-intelligence/

人工智能伙伴关系，英特尔人工智能事业部数据科学部主任Yinyin Liu：https://ai.intel.com/partnership-on-ai/

英特尔® 实感™ SDK 2016 R2 文档，SDK 架构：https://software.intel.com/sites/landingpage/realsense/camera-sdk/v1.1/documentation/html/index.html?doc_essential_programming_guide.html

IT@Intel: AI Optimizes Intel's Business Processes: An Audit Case Study（IT@英特尔：人工智能优化了英特尔的业务流程：审计案例分析），白皮书，2017 年 11 月：https://www.intel.cn/content/www/cn/zh/homepage.html

使用英特尔® Caffe 优化版在单节点英特尔® 至强® 可扩展处理器系统上实施深度学习训练和测试（英特尔® AI Academy）：https://software.intel.com/zh-cn/articles/deep-learning-training-and-testing-on-a-single-node-intel-xeon-scalable-processor-system

0318/RD/CAT/PDF