正确实施 AI：

坚实的基础可保证包括您自己在内的所有人规划成功的 AI 之旅

IBM 的“数据-训练-推理”AI 模型如何为企业的长久成功保驾护航

概述

如果我们发现，数据科学背景不再是理解、解读和处理企业 AI 中最复杂方面的必要条件，会怎么样？

这些知识如何影响贵组织为保持竞争力所依赖的业务流程和应用？

就传统而言，企业 AI 一直都是分析专家的“专利”，用于深入理解如何构建和训练模型。但是，当 AI 计划扩大到整个企业范围时，一切都发生了改变。最明显的变化是采用基于价值的框架指导 AI 实施工作。这就是“数据-训练-推理”(DTI) AI 模型，也是本文所要介绍的主题。

在深入研究细节之前，大家必须清楚，DTI 模型并非线性工作流程。它是由三个阶段构成而且这三个阶段始终在不断互动的持续循环。由于该过程一直在持续进行，因此提取的洞察更加丰富、更有价值。

结果怎样呢？主要的利益相关方可以更迅速、更自信地做出更明智的决策。

数据-训练-推理 (DTI) AI 模型

第 1 阶
数据

AI 专家指出，数据阶段是这三个阶段中最耗时的一个。准备现有数据以供采集消化（用更通用的术语说，就是将数据加载到 AI 模型进行训练）的工作量非常大。

俗话说，垃圾进垃圾出。在 AI 领域更加印证了这句话。如果所使用数据的质量、真实性甚至数量存在问题，那么据此创建的 AI 模型也就存在问题，只能为企业带来存疑的结果。如果没有扎实的数据基础，AI 项目就是空中楼阁，甚至在开始之前就已“误入歧途”。

“如果数据的质量、真实性甚至数量存在问题，那么据此创建的 AI 模型也就存在问题，只能带来存疑的结果。”

您知道吗？

遍布世界的海量数据，加之快速增强的服务器基础架构性能，引发了最近十年的 AI 革命。

用于 AI 训练的数据来自各种不同的数据源。这可能包括我们熟悉而且易于理解的来源，例如来自现有企业数据仓库的过往销售数据或客户编号。但也可能来自实时来源，例如网络边缘的物联网设备或其他互联网数据流（例如 Twitter）。

有关数据的 4 个事实

首先，数据来自许多不同的来源。它们可能具有未经处理的格式，例如文本、图像、声音或原始数值。数据科学团队需要花费大量时间来收集数据，然后将它们整理为合适的格式，以便加载到环境中，供框架使用。这些步骤很重要。分析和识别数据集的具体特性对于模型而言非常重要，这会对推动业务价值的结果产生影响。

第二，许多数据可能起积极作用，也可能起消极作用。企业数据通常零散的分布于多个地方。准确的科学结论通常是多个不同来源数据共同形成的结果，但重复会带来意想不到的负面结果。

第三，数据的时效性至关重要。基础条件可能在不断迅速变化，因此需要向 AI 模型提供最新数据，以继续推动业务价值的增长。过时的数据会影响模型的价值。模型的“新鲜度”取决于投入生产环境的数据的“新鲜度”，因此需要不断采集新的数据。这表明了此步骤本身以及整个“数据-训练-推理”模型的循环性质。为了保持相关性，必须实施相应的计划，不断更新用于训练模型的底层数据集。

最后，要确保海量数据在数据中心内顺畅流动，需要有独特的计算基础架构。如果没有合适的基础架构来执行此项任务，AI 工作流程的吞吐量和性能都会严重下降，根本无法进行要求最苛刻的步骤 — 训练。

我们可以总结出哪些要点呢？干净、相关和新鲜的数据是挖掘宝贵洞察的关键所在。所有这些工作都需要在首次开始训练工作负载之前完成。

第 2 阶
训练

大多数非数据科学家在听到 AI 工作负载时，就会想到 AI 工作流程中的训练阶段。但这个观点并不完全正确。训练是人工智能施展“魔法”的地方 — 将数据转变为 AI 模型。

如果不深入了解机器学习、深度学习和人工智能背后的理论，就可能将训练归结为一个迭代过程 — 采用前一个步骤的数据创建模型。这些模型基于现实世界中的类似数据，对未来进行预测。由于图形处理单元 (GPU) 的出现，直到最近 10 年才有可能以这种方式解决问题。

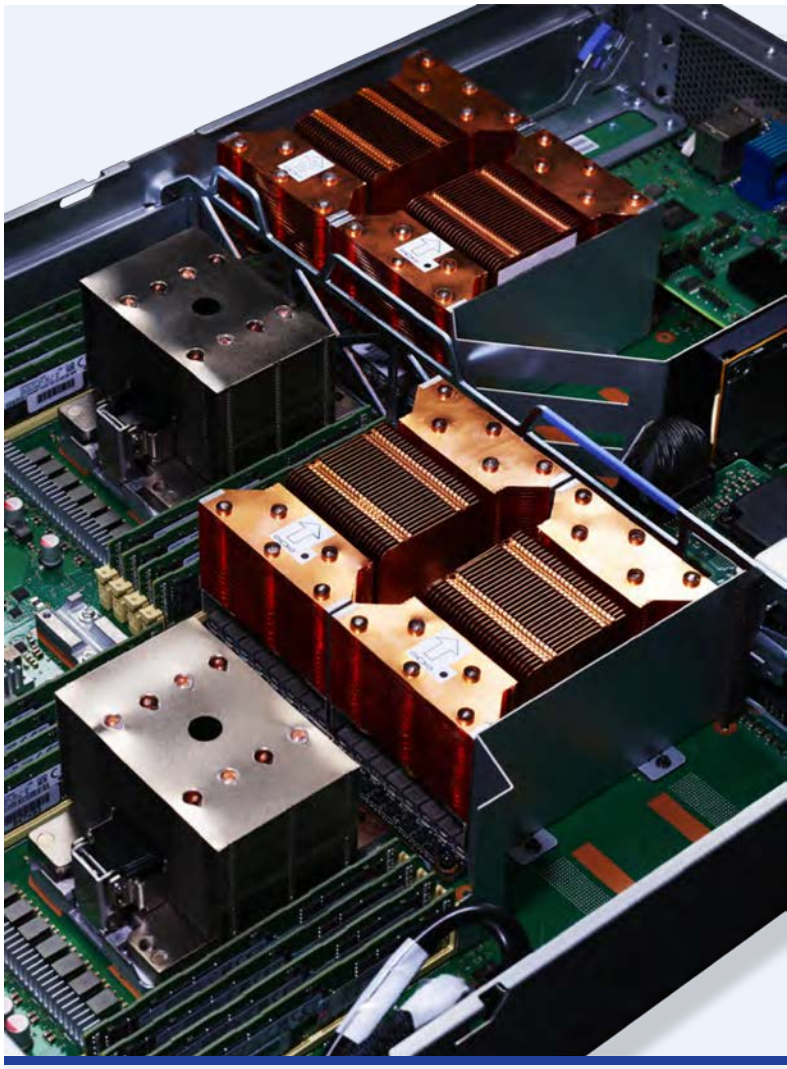
我们能够同时利用中央处理器 (CPU) 和这些新 GPU 的服务器（称为“加速”服务器）。随着传统上以 CPU 为中心的数据中心所承载的 AI 工作负载不断增长，因此需要使用“加速”服务器来扩充数据中心的计算能力。功能更强大的计算也会产生更高的资源成本。因此，正确分配这些昂贵的资源至关重要。分配不当可能会迅速搞砸 AI 项目。

“训练是人工智能发挥魔力的地方，数据将变成 AI 模型。”

对速度（和准确性）的要求

即使在理想情况下，哪怕是训练一个模型也可能需要数天、数周甚至数月的时间。此外，在部署到生产环境之前，模型通常需要训练五到六次。加快训练模型的速度具有不可估量的价值，但必须兼顾速度和准确性，这样才能在面市速度和推动实现关键价值之间取得平衡。

训练流程中，最耗时的任务之一就是设置和重新设置模型的超参数。超参数是数据科学家在训练开始之前为模型选择的值。现代模型可能有数百个这样的值。即便使用样本数据集运行模型，迭代式的设置和重新设置过程也可能需要数小时。通过自动并行地搜索超参数，可以为数据科学家节省数周甚至数月的时间，并可缩短获得结果以及获得准确结果的时间。



“加快训练模型的速度具有不可估量的价值，但必须兼顾速度和准确性，这样才能在面市速度和推动实现关键价值之间取得平衡。”

您知道吗?

如果数据科学家无法保证模型参数取得初步成功，那么可能会浪费数天、数周甚至数月的时间。

训练可视化之类的工具使数据科学家可以查看训练进度，并在训练结果不理想时发出警报。数据团队可在训练开始后的几小时内，停下工作以便对参数进行重新调整，然后重新开始工作，而不是等看到糟糕效果后进行改进。

“时间的矛盾”

与传统代码不同，如果不用“新鲜”数据对 AI 模型进行再训练，模型会随着时间的推移而偏离基础数据。因此，必须持续对现有模型进行再训练，以确保相关性和有用性。但是，还必须能够快速确立新模型，并将其推广到生产环境中。

这就将 IT 领导推到了有关资源分配的时间矛盾的风口浪尖。一些工作负载或租户比另一些更重要。或者说，它们要遵守更严格的服务级别协议 (SLA)。这会反映在资源调度层中。更迅速、更高效的资源调度安排有助于更快地实现与业务相关的模型准确性。

如果不这样，混乱会随之而来。若每个数据科学家或项目都限制在一台机器中，最终会形成计算孤岛。用户被困于一台机器，无法伸展。当用户不使用该机器时，资源也就浪费了。

公平共享和基于优先级的调度安排，有助于在多个训练作业之间动态共享 GPU 资源。这样，还可以在不中断任何作业的情况下，对 GPU 进行优先使用和回收。这有助于提高成本不菲的数据科学团队的生产力；避免 GPU 资源受阻或缺少的情况。并且，由于可在工作负载之间灵活地调度 GPU，因此可以持续保持高水平的资源利用率，最大程度利用这些昂贵的资源。

务必时刻记住训练资源的高价值性质，以及在这个阶段取得成功所付出的必要努力，最终会得到回报：形成有助于提升企业价值的模型。但是，模型的完成并不代表大功告成。接下来还必须进行部署。

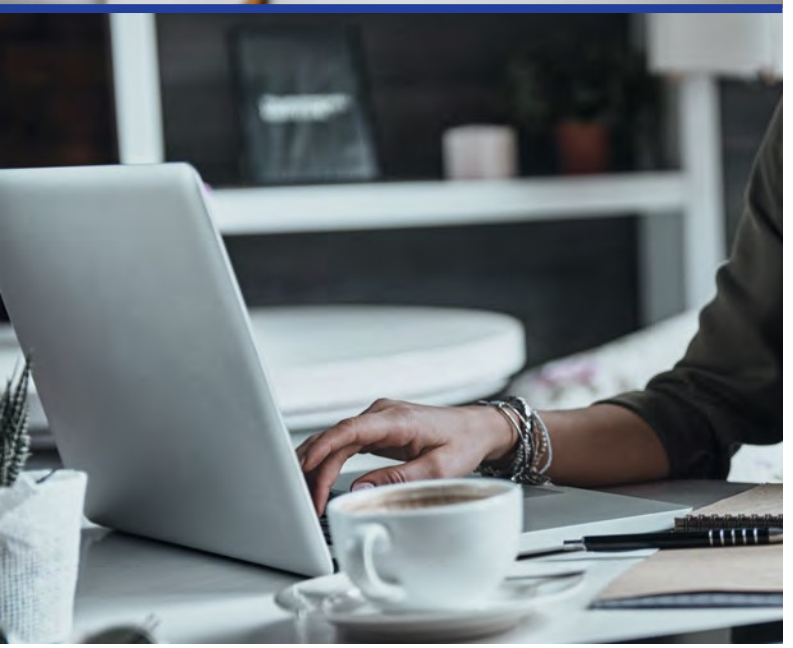
第 3 阶
推理

在 AI 生命周期中，部署到生产环境是从模型中获得洞察的阶段。这个阶段称为推理。（一些人也称其为得分。）这是深度学习发挥价值的地方，也是检验可解释性和公平性指标等 AI 高级概念的阶段。

推理阶段实际上是所有先前阶段的汇总。如果数据质量不理想，或训练不准确，则推理会受到影响。而如果没有正确的推理，等于前功尽弃。

这个阶段所面临的挑战与训练阶段不同。训练可能经历多个循环，需要花费数天或数周的时间；而推理通常是个不到一秒的过程，需要快速、准确的洞察力。

“推理阶段实际上是所有先前阶段的汇总。而如果没有正确的推理，等于前功尽弃。”



推理的实际示例

比如一个付款处理器，它训练了一个模型，用于检测平台上消费者交易中发生的欺诈行为。在处理支付时，客户不会容忍付款过程的长时间延迟。因此，处理器必须对“从 AI 模型获得洞察”放置亚秒级的 SLA，以确保每笔交易都能顺畅进行，并确保客户体验不会受到欺诈支票的影响。

为了应对这些挑战，底层硬件也必须有所不同。训练是在数据中心集中进行的；而推理通常是在边缘进行，比如智能手机等设备或接近边缘的设备。

零售商店中运行的小型服务器机架就是接近边缘场景的一个例子。更具体地说，该服务器在商店中生产有关客户交易或视频回放的实时洞察。

通过适当的资源调度和分配，就可以更快地从模型中获得洞察。通过无缝扩展策略，可以迅速在本地或边缘向上扩展推理能力，以应对需求的增长。与训练阶段相似，如果能够在公共资源池中弹性分配推理任务，则有助于满足严格的 SLA 要求。

如 DTI 框架中的循环箭头所示，在现实世界中收集的数据通过推理回送到工作流程的数据阶段。随着不断应用更深入、更新鲜的基础数据，这个循环动作有助于不断提高模型的准确性。因此，循环周而复始。

整合模型

在 AI 工作流程的每个阶段，需要组合合适的人员、流程和基础架构（包括硬件和软件），才能获得成功。这些都是强大基础的重要组成部分，是在整个企业中部署 AI 的关键。

IBM 提供的 AI 基础架构可以适应不断变化的业务优先级，帮助贵组织实现 AI 之旅中每个阶段的目标。IBM Power Systems 提供行业领先的企业 AI 基础架构，专门服务于机器学习、深度学习和推理。您由此可以：

- 为组织补充新思想和新能力。
- 大规模显著增强对业务决策的信心。
- 借助与企业一起成长的解决方案，充分发挥人员、流程和基础架构的价值。
- 凭借业界最高的数据吞吐量和 IBM 的先进研究成果，更快地找到有意义的结果，保持在 AI 技术的最前沿。

在具备公认安全性的 Power Systems 上获得所有这些优势，无缝集成由 IBM 保护的多种开源框架。

正确实施 AI 可以带来丰厚的红利。现在您已经知道了如何持续规划成功 AI 之旅，那么只剩下一个问题：

您准备好开始了吗？

请访问：

ibm.biz/EnterpriseAI

美国出品
2019 年 7 月

74027174CNZH-00