亚马逊云科技 | NVIDIA

使用最合适的云服务和基础设施加速机器学习创新

轻松准备数据，构建、训练并部署机器学习应用程序

目录

利用机器学习开拓创新

得益于计算能力的进步、存储价格的下降以及云计算的普及，人工智能（AI）和机器学习（ML）已经进入主流应用。各行各业各种规模的企业，包括金融、零售、时尚、房地产、医护及其他行业中的企业，可以利用 AI 和机器学习（ML）实现广泛的业务益处。这包括获取有关客户的更深入的新洞察，确定网络威胁并响应，制定数据驱动型的更明智决策，以及改善招聘流程。[1]

由于这些益处，越来越多的企业投资于 AI 和机器学习（ML）。实际上，IDC 预测，到 2024 年，全球在 AI 领域的支出将达到 1100 亿美元。[2]

机器学习（ML）的使用日益普及的原因之一是它可以更深入地洞察数据。机器学习的工作原理是使用自然语言处理（NLP, Natural Language Processing）、计算机视觉和文档处理等计算算法，通过训练过程从现有数据中学习，然后通过推理过程做出有关新数据的决策。

目前最流行的一些算法包括：

- **自然语言处理（NLP）** – NLP 算法大规模地分析语言，能够理解上下文，解析语音，并近乎实时地进行翻译。这些算法可用于创建 Chatbot、垃圾邮件筛选程序、语音助理和社交媒体监控工具等机器学习（ML）应用程序。
- **计算机视觉** – 计算机视觉算法处理和分析视觉数据，以类似于人类的思维方式检测对象以及对图像分类，但速度和规模都远超人类。这些算法可用于提升工作场所安全性，实现数字身份验证，以及识别不当内容。
- **文档处理** – 文档处理算法可从文档中提取文本、字迹和数据，其功能超过了光学字符识别（OCR），可从表单和表格中识别、理解和提取数据。这些算法可用于从病历中提取信息以及自动处理财务文档。

[1] “成为 AI 驱动型企业。”德勤，2021 年
[2] “全球人工智能支出指南。”IDC，2021 年 8 月

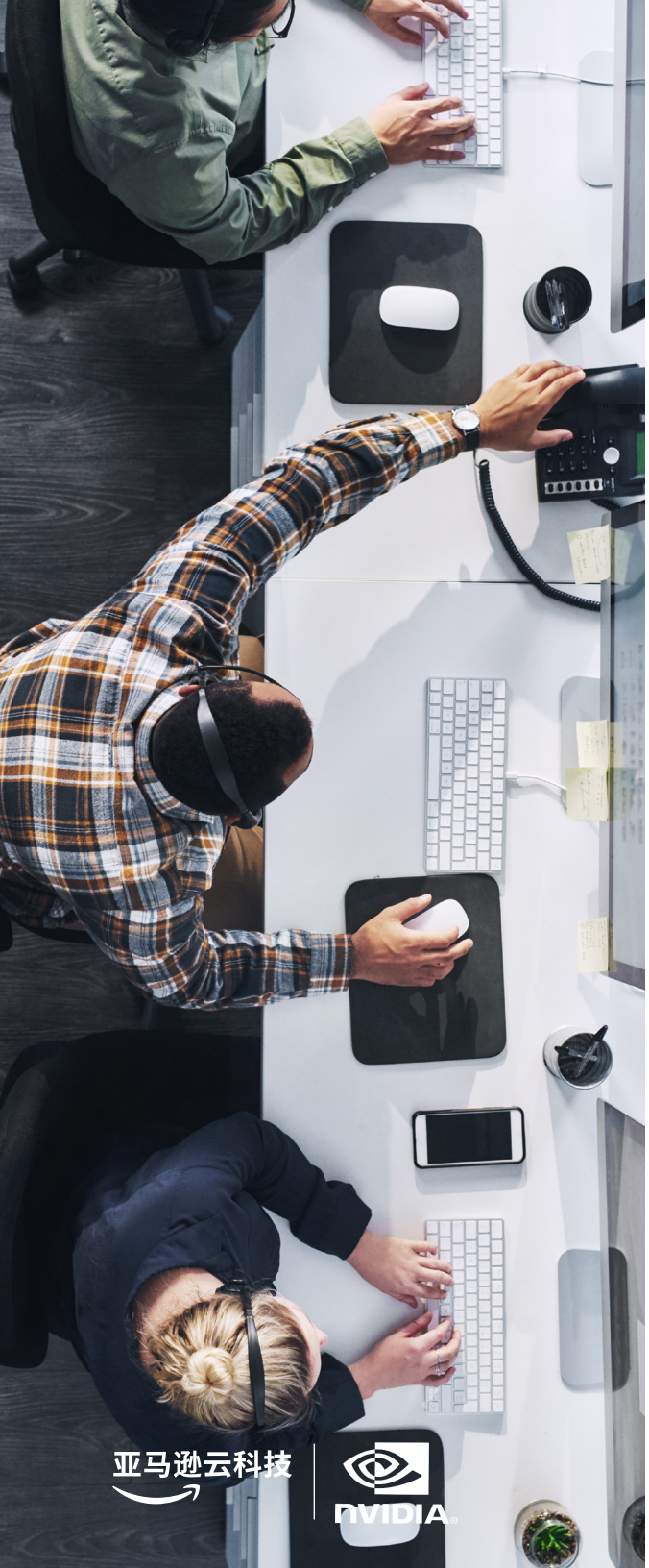
上述应用程序可带来巨大的潜在商业价值，但大规模快速实施同样提出了巨大的资源和基础设施需求。训练用于实现上述使用案例的机器学习模型需要大量的数据、成千上万的计算节点以及增强的节点间/节点内联网。

为了满足这些要求，越来越多的企业将目光转向了云。云将数据、低成本存储、安全性和机器学习（ML）服务融为一体，并提供了高性能计算基础设施用于模型训练及部署。

Amazon Web Services 如何赋能机器学习（ML）

在 Amazon Web Services 上运行的机器学习（ML）超过其他云提供商，亚马逊云科技提供最广泛深入的服务产品组合，助力加速业务转型。从财富 500 强公司到初创公司，各种规模的企业越来越多地采用亚马逊云科技，因为它提供了高性能、低成本的基础设施服务与机器学习服务组合，并针对机器学习进行了优化。通过在云端运行其机器学习工作负载，客户可以按需访问基础设施和机器学习工具，这些资源可在数分钟内启动，从一个实例扩展到成千上万个实例，并且只需为所用资源付费。

我们来看一下目前使用机器学习取得成果的一些亚马逊云科技客户示例。



借助 Amazon Web Services Machine Learning 取得成功

数以万计的客户选择通过 Amazon Web Services Machine Learning 来实现广泛的业务成果。下面是几个例子：

- **美国国家橄榄球联盟**（NFL）与亚马逊云科技合作打造了 Next Gen Stats，这款程序可以结合新旧数据来提供比赛动态的洞察信息，从而与球迷互动。现在，直播员、合作伙伴、球迷等在比赛期间，可以在屏幕上实时使用 20 多种独有的 Next Gen Stats。最让人激动的统计数据使用了构建在 **Amazon SageMaker** 上的预测性机器学习（ML）模型，以前所未有的方式展示运动能力。
- **NerdWallet** 提供工具和咨询服务，让客户轻松管理财务。该公司非常依赖数据科学及机器学习（ML），为客户提供个性化的金融产品。NerdWallet 使用了多项亚马逊云科技服务，例如 Amazon SageMaker 和 Amazon Elastic Compute Cloud（Amazon EC2）P3 实例，以此来提升性能，并将数据科学家训练和迭代机器学习模型的时间从数个月缩短到几天。
- **Freddy's Frozen Custard & Steakburgers** 是一家连锁休闲快餐店，总部位于堪萨斯的维奇塔，该公司希望利用数据科学来找到一种更好的方法，用于评估其各餐厅的质量。借助由 Amazon SageMaker Autopilot 提供支持的 Domo AutoML，Freddy 的 IT 团队得以利用机器学习，在短短几周内就实现了商业价值，而在以前这需要几个月的时间。在准备好机器学习（ML）工具之后，该团队使用 5 倍大小的数据集实现更准确的预测，迅速加快了机器学习（ML）建模过程，并证明了他们对于员工配备的假设。

加快机器学习生命周期中的每一步

企业选择亚马逊云科技是因为我们以系统化的方式消除了机器学习生命周期中每一步的障碍。机器学习生命周期分为四个主要步骤：

1. 数据科学团队需要准备示例数据来训练模型。
2. 然后，他们需要选择用于构建模型的算法或框架。
3. 接下来，需要训练模型来做出预测并不断进行调整，以实现最高的准确性。
4. 最后，需要部署模型，即与其应用程序集成，并在生产环境中监控、扩展和管理模型。

亚马逊云科技可为机器学习（ML）工作流的每一步提供理想的基础设施。您可以自定义计算、联网和存储等基础设施，以适应性能和预算要求。您可以在广泛的高性能、经济高效和可扩展的基础设施中进行选择。

使用此基础设施的最简单快捷的方式是借助 **Amazon SageMaker**，这是一种完全托管式服务，将数据标注、数据准备、特征工程、统计数据偏差检测、AutoML、训练、优化、托管、可解释性、监控和工作流等广泛的功能融为一体。

客户还可以使用 **Amazon Deep Learning Containers**（Amazon DL Containers）。Amazon DL Containers 是预安装了深度学习框架的 Docker 镜像，让您可以跳过复杂的构建流程并从头开始优化环境，帮助您轻松快速地部署自定义环境。此外，**Amazon Deep Learning AMI** 提供预先配置的环境，向机器学习从业人员和研究人员提供所需的基础设施及工具，加快云端任意规模的深度学习速度，从而快速构建深度学习应用程序。您还可以使用常见机器学习框架（例如 Apache MXNet、PyTorch 和 TensorFlow）以及工作流服务和库（例如 Chainer、Gluon、Horovod 和 Keras），快速启动 Amazon EC2 实例并构建、训练和部署模型。

现在，您已经基本了解机器学习（ML）开发过程的工作方式以及亚马逊云科技所能提供的帮助，接下来我们更详细地深入探索这四个阶段。

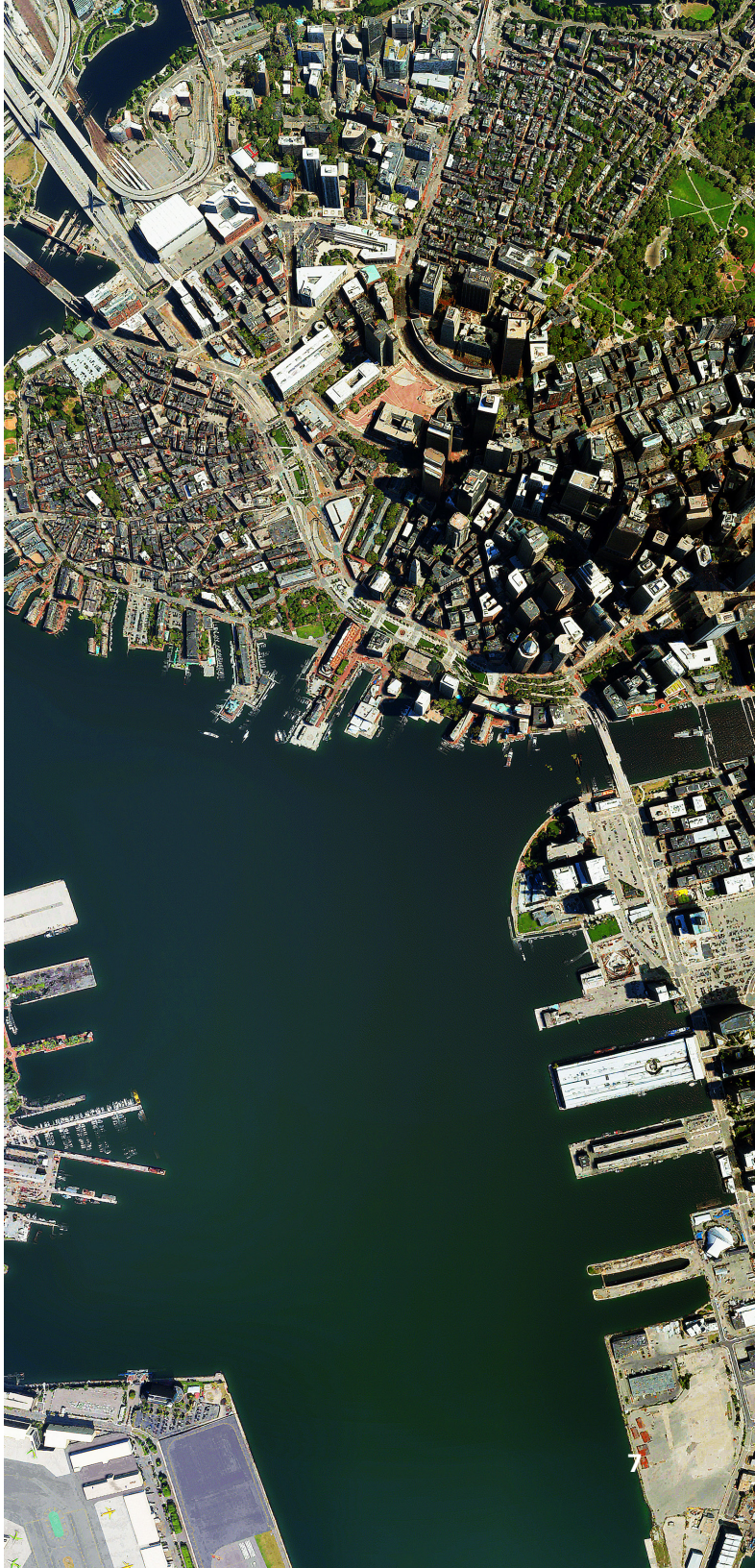
轻松快速地准备数据

面临的挑战

数据是机器学习发展的推动力。但是，即使制定了正确的数据策略，对数据进行管理仍然是机器学习模型构建过程中最耗时、最困难的工作。许多客户表示，他们花了约 80% 的时间在数据准备任务上，例如数据收集、清理和标注。

数据分为两种：结构化数据和非结构化数据。结构化数据是指高度组织化的定量数据，可由机器学习轻松解释。但是，结构化数据仅占数据总量的一小部分。非结构化数据是指定性数据，包括图像、手写笔记和地理空间数据。虽然它非常有价值，但在机器学习中更难使用。机器学习的大多数洞察都源自非结构化数据，但许多现有的数据管理工具往往无法执行非结构化数据分析。例如，医生需要分析来自 X 射线、MRI 和纸质处方的信息。

造成复杂性进一步加剧的是，大部分机器学习工程团队需要编写代码来执行机器学习所需的常见数据准备任务，或者与由其他企业管理的独立的提取、转换、加载（ETL, Extract, Transform, Load）框架集成。



亚马逊云科技 | NVIDIA

解决方案

Amazon SageMaker 可帮助处理结构化数据和非结构化数据。Amazon SageMaker Ground Truth Plus 可帮助客户轻松创建高质量的训练数据集，而无需构建标注应用程序，也无需管理标注员工。Amazon SageMaker Ground Truth Plus 既可帮助将数据标注成本降低多达 40%，又可帮助满足您的数据安全、隐私和合规性要求。您只需上载数据，Amazon SageMaker Ground Truth Plus 就会创建数据标注工作流并加以管理。

对于结构化数据，**Amazon SageMaker Data Wrangler** 采用无代码的可视化界面，大大简化了结构化数据的准备工作。Amazon SageMaker Data Wrangler 包含 300 多种内置的数据转换功能，使您能够快速标准化、转换和组合各个特征而无需编写任何代码。借助 Amazon SageMaker Data Wrangler 的可视化模板，您能够在 Amazon SageMaker Studio 这款针对机器学习（ML）的首个完全集成开发环境（IDE）中进行查看，从而快速预览和检测这些转换是否按预期完成。

准备好数据之后，您能够使用 Amazon SageMaker Pipelines 构建完全自动化的机器学习（ML）工作流，并将其保存在 Amazon SageMaker 特征存放区中供重用。

> ***“借助 Amazon SageMaker Data Wrangler，我们现在可以通过交互式方法，高效地选择、清理、探索和理解数据，使得数据科学团队能够创建特征工程管道，轻松地扩展到包括数亿行的数据集…[并且] 更快地将我们的机器学习工作流投入使用。”***[3]
>
> **Caleb Wilkinson，首席数据科学家，INVISTA**

[3] Amazon SageMaker Data Wrangler

跨多个框架构建准确的模型

面临的挑战

在有了训练数据之后，您需要选择学习风格满足需求的机器学习（ML）算法。这个过程会比较困难，因为可供选择的算法有数十种。Apache MXNet、PyTorch 和 TensorFlow 等机器学习框架使开发更轻松，不过这些框架通常最适合特定算法。这经常会导致需要跨多种算法和框架进行管理和构建，这个过程会很复杂，容易出错并且需要占用大量资源。

构建模型同样需要大量实验和迭代。大多数团队使用 Jupyter Notebooks 方便多个团队共同构建模型和共享工作。不幸的是，随着开发的模型越来越多，共享工作和扩展变得日益困难。

解决方案

如果您希望使用预构建的算法和完全托管式服务来构建高效、准确且功能强大的机器学习模型，Amazon SageMaker 就是适合的解决方案。Amazon SageMaker 包括十多种预先构建的算法，能够部署在您选择的框架上。借助 **Amazon SageMaker Studio**，您能够在单个可视化界面中构建模型，这样可以将数据科学团队的工作效率提升 10 倍。[4]

当您训练模型时，Amazon SageMaker Studio 能够为您提供完整的访问、控制和可见性。您能够快速上载数据、创建新笔记本和调整机器学习（ML）实验。所有机器学习（ML）开发活动，包括笔记本、实验管理、自动创建和调试模型以及模型和数据偏移检测，均可在 Amazon SageMaker Studio 中进行。

Amazon SageMaker Studio Notebooks 管理计算实例，以查看、运行或共享笔记本。底层的计算资源是完全弹性的，因此您能够轻松地启用或关闭可用资源。更改在后台自动发生，不会中断您的工作。您也可以通过几次单击，与其他人共享笔记本。其他人会得到保存在同一个地方、完全一样的笔记本。

如果您准备使用自动机器学习（AutoML）来构建模型，**Amazon SageMaker Autopilot** 能够根据您的数据，自动构建最佳的机器学习模型。您也可以使用 **Amazon SageMaker JumpStart** 快速轻松地向市场推出机器学习应用程序。

加快 150 多种开源模型的部署速度，包括来自流行的 Model Zoo 的可一键单击部署的机器学习模型和算法。只需几次单击即可开始使用，并可轻松地使用预构建解决方案将机器学习（ML）应用程序推向市场，这些解决方案预配置了发布到生产环境时需要的全部亚马逊云科技服务，包括 Amazon CloudFormation 模板和参考架构。

[4] Amazon SageMaker Studio 产品页面，访问时间：2021 年 5 月 11 日
[5] Barr, Jeff。“亚马逊云科技 re:Invent 2020 – Swami Sivasubramanian 的机器学习主题演讲。”亚马逊云科技，2020 年 12 月
[6] Amazon SageMaker 客户

以更低的成本更快地训练模型

面临的挑战

构建模型后，数据科学团队会根据其数据集训练模型，以便模型能够对新数据进行准确预测。训练是一个迭代过程，随着时间的推移，需要重新训练和调整模型，以适应新数据或模型偏差。随着深度学习的普及，模型正变得越来越复杂。模型复杂性每两年加倍一次，许多领先模型现在包含一万亿个参数。训练和调整这些大型模型需要进行大量计算，而且有时成本非常高。

在我们不断推升机器学习（ML）模型的性能和容量上限时，模型训练所需的时间和成本也只会随之增长。这种不断增长的资源消耗会妨碍企业充分利用机器学习（ML）所带来的优势，减慢创新速度，并会使得高管不愿意支持公司对机器学习（ML）进行投资。

“与上一代 GPU 实例相比，在使用 Amazon EC2 P4d 实例时，无需对现有代码进行任何修改，我们就能够将对象识别的训练时间缩短 40%。”[7]

Jack Yan，工程部高级总监，TRI-AD

解决方案

亚马逊云科技提供了采用 CPU、GPU 和自定义加速器的广泛 Amazon EC2 实例供选择，以适合您的机器学习（ML）用例要求。对于具有高内存和网络性能要求的对话 AI 和视频标记等使用案例，Amazon EC2 P4d 和 Amazon EC2 P3 实例是理想之选。

Amazon EC2 P4d 实例是亚马逊云科技平台上性能最高的深度学习训练实例。与上一代 Amazon EC2 P3 实例相比，在训练机器学习（ML）模型时，其性能提升了 2.5 倍并且成本降低了 60%，因而让您能够以极高的效率训练最复杂的多节点机器学习（ML）模型。

Amazon EC2 P3 实例提供了高性能和经济高效的深度学习训练，适用于训练中型到大型模型以及单节点分布式训练使用案例。

Amazon EC2 G5 实例由 A10G GPU 提供支持，训练成本比 Amazon EC2 P3 实例低 15%。它们是一种成本高效的解决方案，用于训练针对自然语言处理、计算机视觉和推荐引擎使用案例的中等复杂程度的单节点机器学习模型。

亚马逊云科技不仅提供了性能最高、经济效益最佳的基础设施，而且还积极地投资于基础设施，以跟上客户不断变化的需求。亚马逊云科技还在不断引入新的芯片和实例来降低训练成本。**Amazon EC2 DL1** 实例采用英特尔旗下的 Habana Labs 公司提供的 Gaudi 加速器，这些实例专门针对训练深度学习模型而设计。这些实例的性价比相比目前基于 GPU 的实例高 40%，非常适合自然语言处理和计算机视觉用例。

[7] Amazon EC2 P4d 实例

亚马逊云科技 | NVIDIA

解决方案

要了解针对机器学习（ML）训练和优化进行了优化的亚马逊云科技基础设施选项对比，请参考以下图表。

实例类型	每实例的最大芯片数	硬件类型	网络带宽	存储	额外的功能
Amazon EC2 P4d	8 个 GPU A-100	英伟达	400 Gbps EFA，GPU-Direct RDMA	8 TB NVMe	可以部署于由 4000 多个 GPU、高速网络和高吞吐量低延迟存储组成的 Amazon EC2 UltraClusters
Amazon EC2 P3	8 个 GPU Tesla V100	英伟达	100 Gbps，EFA	1.8 TB NVMe	支持所有主流机器学习（ML）框架
Amazon EC2 DL1	8 个 Gaudi 加速器	Habana Labs，英特尔	400 Gbps，ENA	8 TB NVMe	**Habana SynapseAI SDK**
Amazon EC2 G5	8 个英伟达 A10G Tensor Core GPU	英伟达	100 Gbps	7.6 NV/Me	支持所有主要框架和英伟达库

Amazon SageMaker 使用内置工具来管理和跟踪训练实验、自动选择最佳超参数、调试训练任务以及监控 GPU、CPU 和网络带宽等系统资源的利用率，从而减少训练和调整机器学习模型的时间和成本。Amazon SageMaker 会自动根据您的训练任务要求纵向扩展或缩减基础设施，从一个 GPU 到数千个 GPU，或从 TB 级存储到 PB 级存储。此外，由于您只需为使用的资源付费，因此您可以更有效地管理训练成本。

要更快地训练深度学习（DL, Deep Learning）模型，您可以使用 Amazon SageMaker Training Compiler，通过图表和内核级优化来更有效地利用 GPU，从而将模型训练过程加快多达 50%。此外，您可以通过几行代码将数据并行性或模型并行性添加到训练脚本中，Amazon SageMaker 分布式训练库会自动跨 GPU 实例拆分模型和训练数据集，帮助您更快地完成分布式训练。

快速且经济高效地部署模型

面临的挑战

在您对模型进行了训练和优化，使之达到所需的准确度和精度级别后，就能够将模型放到生产环境中进行预测。这称为机器学习（ML）的预测或推理步骤。

如果一个模型需要数百毫秒才能生成文本翻译、对图像应用筛选条件或者生成产品建议，就会让用户觉得应用程序反应迟钝或者难于使用，导致用户流失。通过加快推理速度，您能够减少整体应用程序延迟，提供顺畅的体验。

开发和运行机器学习应用程序时，高达 90% 的基础设施成本花在了推理上，这就使得对高性能、低成本的机器学习推理基础设施的需求非常关键。[8]

[8] Amazon EC2 Inf1 实例，访问时间：2021 年 5 月 11 日

解决方案

针对机器学习（ML）推理，亚马逊云科技提供了广泛的高性能、经济高效且易于使用的实例。对于计算机视觉和 NLP 等高度复杂的模型，最佳选择是 **Amazon EC2 Inf1** 实例。Amazon EC2 Inf1 实例由亚马逊云科技全新构建，采用 Amazon Inferentia，相比采用 GPU 的 Amazon EC2 实例，其每实例的成本降低多达 70%，吞吐量提升高达 2.3 倍。

考虑到模型、框架或操作员支持等原因，希望继续使用英伟达生态系统进行推理的客户能够利用 **Amazon EC2 G5** 实例进行高性能的推理。

如果您在寻求通过利用英特尔 AVX-512 向量神经网络指令的模型进行推理，**Amazon EC2 C5** 实例可帮助加快卷积等典型的机器学习运算，并自动提升广泛的深度学习工作负载的推理性能。

使用以下图表来比较针对机器学习（ML）推理进行了优化的亚马逊云科技基础设施选项。

实例类型	每实例的最大芯片数	硬件类型	网络带宽	存储	额外的功能
Amazon EC2 Inf1	16 个 Amazon Inferentia 芯片	Amazon Inferentia	100 Gbps	19 Gbps 的 EBS 带宽	**Amazon Neuron SDK**，该软件支持所有主流机器学习框架，只需最少的代码更改即可将模型迁移到 Amazon EC2 Inf1 实例上
Amazon EC2 G5	8 个英伟达 A10G Tensor Core GPU	英伟达	100 Gbps	7.6 NV/Me	支持所有主要框架和英伟达库
Amazon EC2 C5	96 个 vCPU	Intel AVX	25 Gbps	4 个 900 NVMe SSD	在 Nitro 上构建

解决方案

Amazon SageMaker 提供了一系列广泛的机器学习基础设施和模型部署选项，可帮助满足您的实时需求或分批需求。部署模型后，Amazon SageMaker 会创建持久性终端节点以集成到应用程序中，从而进行机器学习预测。该软件支持从低延迟（几毫秒）和高吞吐量（每秒数十万个推理请求）到 NLP 等使用案例的长时间运行推理的完整推理要求范围。无论您是自带模型和容器还是使用亚马逊云科技提供的模型和容器，都可以使用 Amazon SageMaker 实施 MLOps 最佳实践，减轻大规模管理机器学习模型的运营负担。

对于具有间歇性和不可预测的使用模式的使用案例，Amazon SageMaker Serverless Inference 可让您以按使用付费的定价来部署机器学习模型，而无需担心服务器或集群。在部署模型时，只需选择无服务器选项，Amazon SageMaker 就会根据推理请求的数量自动预置、扩展和关闭计算容量，这样一来，您便无需管理复杂的扩展策略和预先预测流量需求。

Amazon SageMaker Inference Recommender 可帮助您选择最佳的可用计算实例和配置来部署机器学习模型，从而实现最佳的推理性能和成本。SageMaker Inference Recommender 会自动选择计算实例类型、实例计数、容器参数和模型优化以进行推理，从而最大限度地提高性能并降低成本。

Amazon SageMaker 模型部署功能与 MLOps 功能原生集成，其中包括 **Amazon SageMaker Pipelines**（工作流自动化和编排）、SageMaker Projects（用于机器学习的 CI/CD）、**Amazon SageMaker 特征存放区**（功能管理）、Amazon SageMaker Model Registry（模型和构件目录，用于跟踪沿袭并支持自动化审批工作流）、**Amazon SageMaker Clarify**（偏差检测）和 **Amazon SageMaker Model Monitor**（模型和概念偏差检测）。因此，无论您是部署一个模型还是数万个模型，Amazon SageMaker 都有助于减少部署、扩展和管理机器学习模型的运营开销，并更快地将模型投入生产。

> ***Autodesk 正在使用 [亚马逊云科技] 的 Inferentia，推进我们的人工智能虚拟助手 Autodesk Virtual Agent（AVA）的认知技术。通过试部署 Inferentia，我们在 [Amazon EC2] G4dn 上的 NLU（自然语言理解）模型的吞吐量提升了 4.9 倍，并且我们有意在基于 Inferentia 的 [Amazon EC2] Inf1 实例上运行更多工作负载。”***[9]
>
> **Binghui Ouyang**，高级数据科学家，Autodesk

[9] Amazon EC2 Inf1 实例

构建强大平台，赋能机器学习

选择合适的服务和基础设施能够极大地提升机器学习（ML）工作负载的性能，帮助您更快地准备用于机器学习（ML）的数据，构建可靠的先进模型，快速完成大规模的训练，并以强大且经济高效的方式部署它们。不论您是要将大量开发工作分载到完全托管式服务，从头开始创建模型，还是处于这两个阶段之间的任何位置，合适的服务和基础设施都能帮助您更快地完成机器学习（ML）项目，实现更好的结果。

亚马逊云科技提供了高性能、低成本的机器学习服务与基础设施服务的理想组合，并针对机器学习（ML）进行了优化。通过在云端运行机器学习（ML）工作负载，您能够按需访问基础设施和机器学习（ML）工具，这些资源能够在数分钟内启动实例，从一个实例扩展到成千上万个实例，并且您只需为所用资源付费。

开始使用机器学习 ›

亚马逊云科技 | NVIDIA