亚马逊云科技 | NVIDIA

巧用机器学习开拓创新

借助应用了机器学习的产品、服务和客户体验，更好地满足客户需求

引言

机器学习不再是全球科技巨擘和数据科学专家的专利，现已成为主流技术。幸赖于云技术，机器学习广泛应用的障碍正在迅速消失。云将数据、低成本存储、安全性和机器学习服务，以及高性能、经济高效的基于 CPU 和 GPU 的计算实例，统统融为一体，这对于机器学习的成功至关重要。云还提供了一种随用随付的成本模型，可进一步帮助客户控制成本。

目前最前沿的复杂深度学习模型由灵感来自人脑功能的多层深度神经网络组成，因此迫切需要更加强大的计算资源。这些高级模型需要安全、可扩展且经济高效的 CPU 和功能强大的 GPU，以及 GB 或 TB 级存储。

有了云，您可以选择完全托管式服务，自动管理您的基础设施，从而为您省去软硬件维护的麻烦，或者您也可以选择自行管理机器学习生命周期，从以更加亲力亲为的方式定制基础设施的规模和能力中受益。

无论您最终做何选择，只要有了云，您便无需在前期为一切可能的选项一一投资。资源是按需提供的，且始终保持在最新状态，随时可为您提供专用的机器学习工具、计算、存储、联网和最新的基础设施创新成果。

借助机器学习更好地满足客户需求

有几个常见的机器学习使用案例可帮助客户实现业务转型

智能文档处理

简介：

各组织通常都有多种类型的文档，比如发票、患者信息表、贷款申请和合同等，这些文档包含申请人姓名、实体（地址或品牌）或患者健康史等数据，而这些数据对组织的业务又至关重要，需要进行处理。借助于使用文本处理算法构建的机器学习模型，智能文档处理可从数百万个文档中提取文本，可以阐明数据的情感或数据之间的关系，并增加了一道验证、纠正或增强机器学习结果的人工工序，以求提高准确性和合规性。

用途：

智能文档处理从数字文档中提取数据，以执行诸如处理贷款申请、分析客户情绪、确定患者治疗方式，或者从发票中筛选出不合规的购买行为等任务。

成果：

文档处理基于机器学习，因此数据准确性更高，数据处理速度更快。它还可以提高客户满意度，提供更准确的信息，并帮助各公司更快、更恰当地响应请求。

智能文档处理提高了员工的工作效率，让员工可以将更多的时间花在业务关键型任务上，并减少为了获得洞察而需要人工浏览各种文档以及执行手动数据输入所耗费的时间。自动化文档工作流降低了数据提取和分析的复杂性，使企业能够减少用于这些劳动密集型操作的预算和资源。

客户成功案例

Mayo Clinic 的一项研究表明，近 90% 的美国放射科医生都在满负荷或超负荷工作，而 **Rad AI** 有望帮助减轻他们的工作负担。该公司训练机器学习模型，替放射科医生阅读详细文档并自动总结结果，医生用这些结果来诊断患者的病症，然后制定出治疗方案。Rad AI 选择将其文档摘要应用程序从基于 GPU 的 Amazon Elastic Compute Cloud（Amazon EC2）旧实例迁移到由 NVIDIA A100 Tensor Core GPU 提供支持的 **Amazon EC2 P4d** 最新实例。Rad AI 通过在 Amazon EC2 P4d 实例上部署其应用程序，将机器学 习推理时间缩短了 60%，从而 能够向放射科医生提供更快、更准 确的报告，进而改善患者的疗效。

确的报告，进而改善患者的疗效。**汤森路透社**是世界上最值得信赖的资讯提供商之一，其专家团队汇集信息、创新和洞察，为全球各地的企业揭示复杂情况。在法律、税务、新闻等其它领域，汤森路透社拥有 150 多年以来人工加以注释的各类数据。2018 年，该公司选择了 **Amazon SageMaker** 来加快其研发工作。汤森路透社自部署 Amazon SageMaker 以来，能够充分利用诸如即时回答生成、长文本摘要以及完全交互式、对话式问答等高级功能。这些功能使汤森路透社能够构建全面的辅助 AI 系统，引导用户找到满足所有信息需求的最佳解决方案。

Rad AI 通过在 Amazon EC2 P4d 实例上部署其应用程序，将机器学习推理时间缩短了 60%，从而能够向放射科医生提供更快、更准确的报告，进而改善患者的疗效。

计算机视觉

简介：

借助计算机视觉（CV），机器能够识别图像中的人物、地点和事物，准确度达到或超过人类水平，并且速度和效率要高得多。CV 自动从单个图像或图像序列中提取、分析、分类并理解有用信息。图像数据可以采用多种形式，例如单张图像、视频序列、来自多个摄像机的视图或三维数据等。

用途：

一些公司使用 CV 来检测媒体和娱乐节目中的不当内容，帮助提高品牌声誉和安全性。例如，在医护领域，CV 应用程序用于分析医学图像，并识别需要额外分析的图像，从而改善患者治疗。在制造领域，许多公司正在使用 CV 自动检测装配线上的组件是否存在缺陷。

客户成功案例

牛津大学的**花园、图书馆和博物馆（GLAM, Gardens, Libraries & Museums）**收藏了 2100 万件藏品，这些都是世界上特别重要的文物和标本。利用 Amazon SageMaker，牛津大学目前正在使用由基于英伟达 GPU 的 **Amazon EC2 P3 实例**提供支持的 CV 来构建增强版图像识别系统，并加快对其大量硬币收藏进行编目的过程。以前，分析一枚硬币需要志愿者 10 分钟到几个小时的时间，今后，一旦构建完成图像识别系统，预计分析过程只需几分钟。[1]

Aerobotics 是一家农业科技公司，业务遍及全球 18 个国家/地区，总部位于南非开普敦。该公司的使命是提供智能工具来回馈世界。为达成这一使命，该公司在其 Aeroview 平台上为农民提供富有实用价值的数据和洞察，以便农民们可以在生长季节的适当时间进行必要的介入。公司的主要数据来源是无人机航拍图像：捕捉果园中树木和水果的视觉和多光谱图像。

Aerobotics 使用 Amazon SageMaker 改进他们的 Tree Insights 产品，该产品提供每棵树的重要参数的测量值（如树冠面积和健康状况等），还能提供枯死和缺失树木的位置。农民利用这些信息进行精确的干预，如修理灌溉线、以不同速率施肥，以及订购新的树木补种等。

以前，分析一枚硬币需要志愿者 10 分钟到几个小时的时间，今后，一旦构建完成图像识别系统，预计分析过程只需几分钟。[1]

[1] “牛津大学推出行业领先的图像识别机器学习原型，旨在增强钱币学的数字化”，亚马逊云科技，2021 年

个性化推荐

简介：

如今，消费者在考虑、购买和使用产品和服务时，都希望通过数字化渠道获得实时的、经策管的体验。机器学习算法可用于根据个人喜好和行为，跨渠道扩展并创造个性化的客户体验。

用途：

企业可以构建能够提供广泛的个性化体验的应用程序，包括特定的产品推荐、个性化的产品重新排名和定制型直接营销。通过基于机器学习的个性化，企业可以超越僵化的静态规则，使用推荐系统，向客户提供高度个性化的建议。

成果：

机器学习可以帮助企业提供高度个性化的体验，从而提高客户参与度、转化率、收入和利润率，并在数字世界中创造差异化。

客户成功案例

跨国电子商务公司 **Zalando** 决定在云端将其机器学习工作负载标准化，进而改善客户体验、提高工作效率，同时推动公司业务发展。借助 Amazon SageMaker，Zalando 得以更好地开展宣传活动、生成个性化服装推荐并提供更具吸引力的客户体验，同时将工程师和数据科学家的工作效率提高 20%。[2]

NerdWallet 是一家个人理财初创公司，通过提供适当的工具和建议，帮助客户轻松偿还债务、选择最佳金融产品和服务并且实现主要的人生目标，例如购房或储蓄退休金等。该公司非常依赖数据科学及机器学习为客户提供个性化的金融产品推荐。以前，NerdWallet 会以列表的形式为客户推荐他们可能喜欢的信用卡，但无法预测被客户接受的可能性。借助 Amazon SageMaker，该公司可以更有效地为客户匹配合适的金融产品。

通过结合使用 Amazon SageMaker 和配备了 NVIDIA V100 Tensor Core GPU 的 Amazon EC2 P3 实例，NerdWallet 还提高了灵活性和性能，可将数据科学家训练机器学习模型所需的时间从几个月缩短到几天。

借助 Amazon SageMaker，Zalando 得以更好地开展宣传活动、生成个性化服装推荐并提供更具吸引力的客户体验，同时将工程师和数据科学家的工作效率提高 20%。[2]

[2] “亚马逊云科技客户成功案例：Zalando”，亚马逊云科技

使用场景不断增多

除了这些常见使用场景之外，我们还观察到，
有一些新兴使用场景正在迅速涌现

自主操作系统保护行人和驾驶人员

自主操作系统使用许多不同的机器学习模型来感知环境，并在没有人工干预的情况下运行。自主操作系统依靠传感器、执行器、复杂算法、机器学习系统和强大的处理器来快速执行软件。

机器人就是自主操作系统的一个例子。作为尖端技术的提供者，Amazon Robotics 早就知道，如果能使用人工智能（AI, Artificial Intelligence）和机器学习来自动处理货品配送流程的关键方面，将意味着有可能为公司带来巨额收益，因此在 2017 年，公司投入了若干团队来实现这一点。随着该公司迭代机器学习项目，他们转向了 Amazon Web Services 和 Amazon SageMaker，后者是一项托管式服务，可帮助数据科学家和开发人员快速准备、构建、训练并部署高质量的机器学习模型。这样一来，Amazon Robotics 团队从维护和管理 GPU 的艰巨任务中解放了出来，同时仍能使用大量 GPU 跨多个区域大规模运行推理。截至 2021 年 1 月，该解决方案为公司节省了近 50% 的机器学习推理成本，并使生产力提高了 20%，同时节省了比例差不多的总体成本。

通过预测型医护拯救生命

目前，由于患者数据不足，或者没有足够的时间来分析并关联大型患者数据集，医疗人员对严重疾病的及时诊断常常被延误。为了帮助应对这一挑战，**CloudMedx** 正在开发机器学习模型，这些机器学习模型能够理解不同的疾病、症状和药物之间的相互关系，并能够帮助预测疾病进展和患者出现并发症的可能性。[3]

利用机器学习改善废水管理

对于公共卫生以及保护宝贵的水资源而言，废水管理至关重要。**Opseyes** 利用 AI 开发了第一个用于废水处理厂的快速显微镜测试。这种测试能有便于用户立即检查工厂状况，迅速消除污染威胁，从而避免停机。[4]

机器学习让音乐家如虎添翼

Sunhouse 是一家由音乐家和技术专家创立的 AI 初创公司，该公司正在创建一个基于机器学习的系统，使鼓手能够将架子鼓变成用于巡演、创作和即兴表演的整个生产套件。Sunhouse 的技术采用定制的鼓传感器和 AI 支持的声学映射，使鼓成为一种富有表现力的工具，能够使用样本、效果和 MIDI 进行创作和表演。Sensory Percussion 使鼓手能够用鼓槌控制电子设备，为打造创意和创作音乐开辟了一条全新的道路。Sunhouse 的解决方案自建立以来，已被纽约爵士乐界的现场表演者和当今许多著名鼓手广泛使用，其中包括 Marcus Gilmore 和 Wilco 的 Glenn Kotche。[5]

[3] Darrow, B.，“此 AI 软件 {CloudMedx} 旨在征服心脏病”，CloudMedx，2016 年
[4] Caufield, B.，“AI 的力量：初创公司 Opseyes 即时分析废水”，英伟达博客，2021 年
[5] Finkle, L.，“鼓声响起：AI 初创公司 Sunhouse 的创始人 Tlacael Esparza 找到了自己的节奏”，英伟达播客，2021 年

来自亚马逊云科技和英伟达的解决方案

亚马逊云科技和英伟达合作已超过 10 年，不断为客户提供强大、经济高效且灵活的基于 GPU 的解决方案。**基于英伟达 GPU 的 Amazon EC2 实例**提供了快速准确地训练机器学习和深度学习模型所需的高性能、成本优化型基础设施。借助由 NVIDIA A100 Tensor Core GPU 支持的最新 **Amazon EC2 P4d 实例**，开发人员能够将训练模型的时间从几天缩短到几分钟。您能够以高效率和优化的成本，运行最复杂的多节点训练。这使您能够快速试验、训练并调整模型，从而加速创新。

它们是第一批在云中支持 400 Gbps 实例联网的实例。与上一代 Amazon EC2 P3 和 Amazon EC2 P3dn 实例相比，Amazon EC2 P4d 实例使深度学习模型的性能平均得以提高 2.5 倍。

Amazon EC2 P4d 实例还部署在称为 EC2 UltraClusters 的超大规模集群中，这种集群由云中具有最高性能的计算、联网和存储资源组成。每个 EC2 UltraCluster 都是全球最强大的超级计算机之一，使您能够运行最复杂的多节点机器学习模型训练。在 EC2 UltraClusters 中，您可以根据机器学习项目需求，从几个 NVIDIA A100 Tensor Core GPU 快速扩展到数千个。

新的 **Amazon EC2 G5** 实例是最新一代基于 NVIDIA GPU 的实例，可用于一系列广泛的机器学习使用案例。与 Amazon EC2 G4dn 实例相比，该实例的机器学习推理性能提高了 3 倍，机器学习训练的性能提高了 3.3 倍。

客户可以使用 G5 实例获得高性能且成本高效的基础设施，来训练和部署用于自然语言处理、计算机视觉和推荐器引擎使用案例的更大、更复杂的模型。

G5 实例配备了多达 8 个 NVIDIA A10G Tensor Core GPU 和第二代 AMD EPYC 处理器。它们还支持多达 192 个 vCPU、高达 100 Gbps 的网络带宽和高达 7.6 TB 的本地 NVMe SSD 存储空间。

立即借助 Amazon SageMaker 这一利器开始利用强大的机器学习基础设施

要从亚马逊云科技的强大机器学习基础设施中获益，最简单快捷的方法就是部署 **Amazon SageMaker**。这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基本功能融为一体。

Amazon EC2 上提供的英伟达 GPU 可以帮助开发人员显著加快在 Amazon SageMaker 中开发机器学习算法的各个阶段（包括模型训练和推理），从而降低构建和部署机器学习应用程序的总体成本。

为实现这些优势，英伟达提供了 **NGC 目录**，这是 GPU 优化库、用于计算机视觉的预训练 AI 模型、对话式 AI 和推荐器、应用程序框架和推理服务解决方案（例如 **Triton Inference Server**）的全面集合，用于简化机器学习模型在 CPU 和 GPU 上的部署。NGC 目录中的软件不断优化，使开发人员能够轻松利用最新的英伟达创新。软件每月都会发行更新版，使用户能够享受最新功能和性能改进。

英伟达 NGC 目录可直接从**亚马逊云科技 Marketplace** 获取，让用户能够无缝地在 Amazon SageMaker 或其他亚马逊云科技服务（如 Amazon Elastic Kubernetes Service（Amazon EKS）或 Amazon Elastic Container Service（Amazon ECS））中提取和运行 GPU 优化的容器和模型。

将机器学习应用于您的业务

通过在云中运行机器学习工作负载，企业可以按需使用最强大的 GPU 实例和机器学习工具，这些实例和工具可以在几分钟内启动，从一个实例扩展到数千个实例，同时还能控制基础设施成本。

英伟达支持的亚马逊云科技服务和基础设施适用于所有经验水平的企业，无论是擅长构建机器学习工作负载且希望自行管理其基础设施的企业，还是更喜欢完全托管式方案的企业。亚马逊云科技在机器学习开发生命周期的每个步骤中，使用计算、联网、存储和机器学习工具为您的企业提供支持，包括收集和准备数据，选择正确的算法，调整模型以获得最高准确性，以及长期部署并监控模型的性能和质量。

了解亚马逊云科技和英伟达的潜力 ›

开始使用 Amazon SageMaker ›

亚马逊云科技 | NVIDIA