计算机世界/2020 年/9 月/28 日/第 007 版
产品潮流汇

如何选择云机器学习平台

InfoWorld 网站特约编辑兼评论员　Martin　Heller　编译　陈琳华

为了能够支持完整的机器学习生命周期,每个云机器学习平台均应具备 12 种功能。

用户需要大量的数据,对数据进行清洗,并在合理的时间内训练数据模型,这样才能创建高效的机器学习和深度学习模型。有了高效的机器学习和深度学习模型之后,用户需要部署和监视这些模型。如果发生了变化,用户还得根据需要重新对它们进行训练。

对于那些已在计算资源和 GPU 等加速器上投入了巨资的用户,他们可以在本地执行所有这些操作,不过这些用户可能会发现,在资源足够的情况下,许多资源其实在很多时间都处于闲置状态。与此同时,用户可能还会发现,在云端运行整个管道可能成本效益更为出色,因为云服务可以根据实际需求调用大量的计算资源和加速器,在不需要的时候再把这些资源释放出来。

目前,为了支持从项目规划到维护生产模型一整套完整的机器学习生命周期,多家主要的云提供商和众多小型云提供商都在构建自己的机器学习平台,并投入了大量精力。那么用户如何确定哪些云服务能够满足自己的需求呢?以下 12 种功能是所有云机器学习平台都应具备的功能。

控制成本

用户需要控制模型的成本。通常情况下,在深度学习成本当中,生产预测模型的部署成本占了 90%,训练仅占 10%。用户的负载和模型的复杂性决定了预测成本的控制。

如果负载很高,那么用户可以使用加速器来避免增加虚拟机实例。如果负载是波动的,那么用户可以根据负载的变化动态调整实例和容器的数量或规模。如果负载较低或者偶尔才有负载,那么用户则可以选择带有局部加速器的微型实例来处理预测工作。

支持在线建模环境

以往的做法是用户将数据导到桌面上进行建模。如今,构建机器学习和深度学习模型需要大量的数据,这颠覆了用户以往的经验。对于探索性的数据分析和模型构建,用户只需将少量量数据样本下载到桌面上即可,但是想要构建生产模型,用户仍需访问完整的数据。目前,适于构建模型的 Web 开发环境主要有 Jupyter Notebooks、Jupyter Lab 和和 Apache Zeppelin。如果数据与开发环境在相同的云服务上,那么用户可最大程度地减少数据移动,从而节约时间。

支持 ETL 或 ELT 管道

数据库中最常见的两种数据管道配置是 ETL(导出、转换和加载)和 ELT(导出、加载和转换)。机器学习和深度学习放大了对这些环节的需求,尤其是转换环节。在用户的转型需要进行调整时,ELT 可提供更高的灵活性,因为对于大数据而言加载环节是最耗时的。

因此对于机器学习来说,用户还必须将变量控制在标准化范围之间,防止范围波动过大。至于用户到底将标准范围设置为多少,要取决于模型采用的算法。因为原始数据通常都掺杂着大量的无用数据,所以需要进行过滤。另外,原始数据的变化范围非常大,例如一个变量的最大值可能高达数百万,而另一个变量的范围可能是-0.1 至-0.001 之间。

更靠近数据

因为数据传输速度不可能大于光速,距离过长就意味着等待时间过长。即使在具有无限带宽的完美网络上,情况也是如此。如果用户拥有的大量数据足以建立起精确模型,那么理想的状态是在存储数据的地方就近建立模型,这样可以避免传输大量数据。许多数据库也支持这种做法。

其次是将数据与模型构建软件放在同一个高速网络上,这通常意味着数据和模型构建软件在同一数据中心内。即便是在同一个云可用区域内将数据从一个数据中心迁移到另一个数据中心,

如果数据量过大也会出现严重的延迟。那么用户可能通过增量更新的方法来缓解延迟。如果用户不得不在带宽受限且存在高延迟的网络上长距离移动大数据,那么这将是最糟糕的情况。

支持 AutoML 和自动提取特征

通常情况下,AutoML 系统会尝试使用大量的模型,以查看哪些模型有最佳的目标函数值。优秀的 AutoML 系统还可以自动提取特征,并有效地利用资源寻找含有优秀特征集的最佳模型。因为并非所有的用户都擅长选择机器学习模型和模型所使用的变量,以及从原始观察中提取新的特征。即便用户擅长,他们也需要花费大量的时间,因此这些工作有必要实现自动化。

支持纵向和横向训练

除训练模型外,Notebooks 需要的计算和内存资源都不高。如果 Notebooks 能够执行在多个大型虚拟机或容器上运行的训练任务,并且如果训练可以使用 GPU、TPU 和 FPGA 等加速器,那将会带来许多好处。其中,最大的好处就是训练时间可以由数天时间缩短为数小时。

支持最佳的机器学习和深度学习框架

在机器学习和深度学习方面,大多数数据科学家都有自己偏爱的框架和编程语言。对于喜欢 Python 的人来说,他们在机器学习方面更偏爱 Scikit-learn,而 TensorFlow、PyTorch、Keras 和 MXNet 通常是深度学习的首选。云机器学习和深度学习平台通常都有自己的算法集合,并且它们通常使用至少一种语言支持外部框架。部分云平台还针对一些主要的深度学习框架进行了修改。在某些情况下,用户还可以将自己的算法和统计方法与平台的 AutoML 设备集成在一起。

提供预训练的模型并支持迁移学习

以 ImageNet 为例,其数据集不仅非常庞大,而且训练能够使用这些数据集的深度神经网络可能需要花上数周的时间。因此针对 ImageNet 数据集的预训练模型就变得意义重大。

另外,并非所有人都愿意花费大量的时间和计算资源来训练自己的模型。如果可以使用预训练模型,用户就不必如此了。但预训练模型的不足之处在于其可能无法一直标识出用户关注的对象。在这种情况下,迁移学习可以帮助用户针对特定数据集定制神经网络的最后几层,不需要用户再花时间和资金训练整个网络。

支持模型部署预测

在找到了适合自己的最佳模型后,用户还需要能够快捷地部署这些模型。如果用户出于相同目的部署了多个模型,那么用户则还需要再从中进行挑选。

监控用于预测的数据

整个世界是不断变化的,数据也随着世界的变化而变化。用户不能部署完模型就甩手不管了。相反,用户需要不断监控那些出于预测目的而提交的数据。如果数据的变化远远超过了训练数据集的最初设定范围,那么用户则需要重新训练自己的模型。

提供经过优化的 AI 服务

云平台不仅提供了图像识别功能,还为许多应用程序提供了强大的且经过优化了的 AI 服务,例如语言翻译、语音转文本、文本转语音、预测和推荐。为了确保良好的响应时间,目前这些经过优化的人工智能服务已经部署在了计算资源充足的服务端点上。这些服务已使用了大量数据进行了训练和测试,数据在数量上远远大于企业在正常情况下可用的数量。

对试验进行管理

对所有的模式都尝试一遍是为数据集找到最佳模型的唯一方法,无论是手动的还是 AutoML 都要尝试一下。这时紧随而来的另外一个问题就是如何管理这些试验。优秀的云机器学习平台可帮助用户查看并比对训练集和测试数据试验的所有目标函数值,以及模型和混淆矩阵的大小。

本文作者 Martin Heller 此前曾担任 Web 和 Windows 编程顾问。从 1986 年至 2010 年,Heller 一直从事数据库、软件和网站的开发工作。近期,Heller 还出任了 Alpha Software 的技术兼训练副总裁和 Tubifi 的董事长兼首席执行官。