

请参阅本出版物的讨论、统计数据和作者简介：<https://www.researchgate.net/publication/361820573>

分布式机器学习的边缘计算解决方案

会议论文 · 2022 年 7 月
DOI:10.1109/DASC/PICom/CBDCCom/Cy55231.2022.9927824

引文

3

4 位作者：



法布里齐奥·马罗佐

卡拉布里亚大学

109 篇出版物 1,377 次引用

查看资料



多梅尼科·塔利亚


卡拉布里亚大学

449 篇出版物 7,290 次引用

查看资料

阅读

174



阿莱西奥·奥尔西诺

卡拉布里亚大学

9 篇出版物 26 次引用

查看资料



保罗·特伦菲奥

卡拉布里亚大学

160 篇出版物 3,120 次引用

查看资料

分布式机器边缘计算解决方案学习

Fabrizio Marozzo, Alessio Orsino, Domenico Talia, Paolo Trunfio DIMES,
卡拉布里亚大学, 意大利 {fmarozzo,
aorsino, talia, trunfio}@dimes.unical.it

摘要: 近年来, 机器学习 (ML) 在为语音识别、情感分析、电子邮件垃圾邮件过滤、欺诈预防等许多任务提供解决方案方面取得了巨大的成果。物联网 (IoT) 的快速普及, 连接了数十亿台设备, 产生了大量数据, 需要机器学习去中心化解决方案。然而, 在网络边缘执行复杂的学习任务对数据存储、传输和分析的有效管理提出了巨大的挑战。

由于这些原因, 大量的研究和开发工作致力于适应不同的机器学习算法 (例如, 神经网络、集成算法、SVM、k-means), 以便本地数据的协作训练和推理直接发生在机器学习的边缘。网络 (即靠近数据生成位置)。由于边缘设备的容量有限, 这些设备工作和通信所采用的技术不同以及缺乏可轻松管理它们的通用软件堆栈, 这种情况代表了当今的重大挑战。

在本文中, 我们分析了分布式机器学习算法, 以及如何调整它们在网络边缘运行, 并在需要时与云配合以确保低延迟、节能、隐私保护和可扩展性。我们特别简要讨论了主要的机器学习算法如何适应传统分布式平台 (例如集群、云和 HPC 系统), 以及导致这些算法在资源受限的边缘设备上运行的主要研究工作。然后, 介绍并讨论了一种分层方法, 用于在边缘云架构上调整机器学习算法。这是通过考虑应用程序和设备限制以及多层支持架构的特征来完成的。最后, 我们通过描述一些可以从这种方法中受益的应用场景来总结本文。

索引术语 机器学习、分布式机器学习、物联网、边缘计算、云计算、边缘-云连续体

一、简介

在过去的几十年里, 人工智能, 特别是机器和深度学习, 已经成为解决我们日常生活中多项任务的解决方案, 例如语音识别、电子邮件垃圾邮件过滤器、欺诈预防等。近年来机器学习解决方案发展的推动因素之一是大数据[1]。事实上, 用于训练学习算法的大量数据的可用性以及当今不断增长的计算能力极大地促进了机器学习, 机器学习能够从数据中提取潜在有用的信息以进行决策。传统的机器学习方法依赖于存储和

在传统的分布式系统中处理此类数据, 多年来已被证明是解决复杂学习任务的理想解决方案, 特别是对于那些没有低延迟要求的应用程序。

然而, 物联网 (IoT) 设备 (即无需人工干预即可收集和传输数据的联网物体, 例如智能相机和车辆、可穿戴设备和智能手机) 的普及导致了更多的数据生成在网络边缘。网络边缘通常被定义为设备连接到互联网的地方。将所有这些数据从源传输到集中式服务器, 通过机器学习算法收集、处理和分析这些数据涉及高昂的通信成本, 并可能对延迟产生影响, 这对于低延迟应用程序 (例如健康监控和安全) 可能至关重要。为了解决这个问题, 自然要考虑尽可能靠近数据生成的地方 (即设备所在的网络边缘) 来处理数据。

为了满足这一需求, 边缘计算[2]作为一种将计算任务推向网络边缘的范式应运而生。

尽管如此, 位于网络边缘的物联网设备的计算能力、能源能力、存储容量和带宽都有限, 这使得在此类设备上完全执行繁重的学习任务是不可行的。它们通常被定义为资源受限设备, 这意味着它们无法与额外资源集成[3]。由于这些原因, 过去几年一直致力于调整机器学习算法, 以便直接在边缘设备上对本地数据进行协作训练和推理。然而, 由于边缘设备的容量有限、这些设备工作和通信所用的异构硬件和技术以及缺乏可轻松管理它们的通用软件堆栈, 这项任务在当今是一个重大挑战。除此之外, 当数据必须传输到其他设备或远程服务器以并行训练模型时, 安全和隐私是需要考虑的额外关键问题。

通信是另一个挑战, 因为边缘设备之间的带宽可能比本地计算时间慢得多, 因此有必要为训练过程开发通信高效的方法。

在这项工作中, 我们分析了分布式机器学习算法, 以及如何将它们调整为部署在边缘, 并在需要时与云合作, 以确保低延迟、节能、隐私保护和可扩展性。具体来说, 我们简要讨论主要如何

机器学习算法已适应在集群、云和高性能计算 (HPC) 系统等传统分布式平台中工作,并概述了在资源受限的边缘设备上运行这些算法的最先进技术。在这方面,文献中的大部分工作都致力于深度学习应用,因此传统的机器学习技术,如支持向量机 (SVM)、k-means、集成学习 (例如随机森林)等,另一方面,在边缘计算环境中经常使用的人工智能应用尚未得到充分研究。

然后,我们讨论一种在边缘云架构上采用机器学习算法的解决方案,即所谓的边缘云连续体。边缘-云连续体利用从网络边缘 (例如物联网设备)到核心 (例如云数据中心)的所有资源。数据在边缘层生成,首先在本地进行处理,而聚合和部分处理则在中间节点完成。仅在必要时,数据才会传输到云端以供进一步分析[4]。

论文结构如下。第二部分分析了边缘-云连续体的分布式机器学习领域的研究工作。第三节回顾了传统分布式高性能平台中的机器学习算法。第四节回顾了建议在网络边缘的资源受限设备上执行的机器学习算法,包括训练和推理。第五节简要讨论了联邦学习范式中的隐私和安全问题。第六节介绍并讨论了在边缘云架构上采用机器学习算法的分层方法。第七节描述了一些可以从这种方法中受益的应用场景,最后第八节总结了本文。

二.相关工作

基于机器学习的智能应用和服务越来越多地应用于边缘计算环境,这主要是由于实时场景中对低延迟的需求。然而,边缘设备的资源受限性质要求它们必须协作来执行分布式训练和推理,并且可能需要云资源的支持。这个场景提供了这项工作的主要动机。之前关于边缘-云连续体的分布式机器学习的工作并未涵盖传统的机器学习算法,而是广泛使用。

例如,周等人。 [5]从应用和软件角度总结了边缘设备上神经网络模型训练和推理的主要解决方案和使能技术。同样,[6] 回顾了加速边缘设备上深度学习模型的新兴技术。

Murshed 等人以与我们类似的方式。 [7]对资源受限的边缘计算环境中的传统机器学习和深度学习进行了分析。特别是,他们讨论了常见的软件和硬件解决方案

用于边缘深度学习,但对于传统机器学习的讨论仅限于边缘计算场景应用的详细描述。 Imteaj 等人提出了另一项有趣的工作。在[3]中,研究了联邦学习范式,并讨论了为资源受限的物联网设备训练分布式机器学习模型的主要问题。在[4]中,罗森多等人。通过回顾边缘、云和边缘-云架构上应用的机器和深度学习以及数据分析领域的工作,重点关注边缘-云连续体上的分布式智能。然而,[4]主要关注跨边缘-云连续体的大数据处理框架和库,以及支持边缘、云和边缘-云连续体实验研究的领先的最先进的模拟和部署系统。

在接下来的章节中,我们将讨论 i) 机器学习算法如何适应传统的分布式高性能平台,如集群、云和 HPC 系统,ii) 在边缘资源受限设备上执行机器学习的主要努力网络和 iii) 联邦学习范式的分布式学习中的隐私和安全问题。表一总结了边缘-云连续体中最先进的分布式机器学习的主要缺点。

三.分布式机器学习

一般来说,机器学习任务可以分为监督学习、无监督学习和强化学习[38]。

简而言之,监督学习中的训练数据是有标签的,与不需要任何标签的无监督学习相反。

不同的是,强化学习涉及从外部交互的反馈中学习。机器学习算法设计为在功能强大的机器上运行,这些机器通常配备 GPU 和 FPGA 等加速硬件。然而,如今由于训练数据和机器学习模型的规模不断增长,由于硬件有限,单机学习无法高效或有效地完成[39]。因此,分布式计算可以帮助缓解这些问题。在分布式机器学习中,多个工作人员相互协作和通信以并行训练模型。特别是它可以通过两种不同的方法来完成:分布数据或分布模型[40]。在第一种方法中,数据在分布式系统的工作节点上进行分区,这些节点都在不同的分区上执行相同的算法。

然后必须聚合通过在各个分区上训练算法而获得的模型。相反,在第二种方法中,工作节点通过执行模型的不同分区来处理相同的数据,因此最终模型是通过所有部分的聚合生成的。

这种方法可以应用于所有参数可以划分的机器学习算法 (例如,神经网络)。另一种方法基于集成学习,其中训练同一模型的多个实例并聚合输出。在所有这些方法中,工作节点可以组织在集中式架构中

类别	机器学习算法	云边缘	边缘云连续体	目标	参考
分布式机器学习 (第三节)	主要机器学习算法 (例如,k-means,DBSCAN、SVM、随机森林等)	✓		使机器学习算法适应传统的分布式高性能系统	[8]–[23]
资源受限边缘设备上的机器学习 (第四节)	主要机器学习算法 (例如,k-means,kNN、树等)		✓	使机器学习算法适应网络边缘资源受限的设备	[24]–[30]
联合机器学习 (第五节)	主要是梯度下降和随机森林		✓	分布式学习中的隐私和安全问题	[31]–[37]

表 1:边缘-云连续体的分布式机器学习的最新技术。

(也称为参数服务器)或分散式服务器。参数服务器架构由一台或多台服务器和多个工作人员组成,通过与中央服务器更新和同步模型参数以迭代方式执行学习过程。相反,在分散设置中,每个工作节点与其邻居进行通信,并且模型在没有中心节点的情况下进行聚合。在所有方法和架构中,分布式学习的主要好处是避免需要在单个机器上收集大量数据进行处理,从而节省时间和能源并提高可靠性[38]。

下面我们将分析分布式机器学习算法在监督和无监督环境下的一些实现以及一些分布式机器学习框架。值得注意的是,所有分析的论文都提出了如何在传统分布式高性能基础设施 (集群、多处理器和多节点环境)中加速传统机器学习算法 (k-means,DBSCAN,SVM 等)和 HPC 平台)。

其中许多都是基于大数据分析范式和框架,例如 Apache Hadoop 和 Spark。因此,它们不能直接适应部署在边缘设备上,除了可扩展性之外,我们还必须考虑其他问题,例如有限的计算和存储容量、节能、数据隐私和有限的通信带宽。特别是,通信开销是边缘计算环境中的主要挑战之一。

A.分布式监督学习在监督学习中,大部分精

力都致力于开发分布式分类算法,特别是支持向量机和基于树的算法,如随机森林。例如,[8]和[9]提出了一种基于MapReduce的分布式SVM算法,该算法对训练数据进行分区,并在云和计算机集群上优化分区子集,从而减少训练时间,同时保持良好的准确性。参考文献 [10] 提出了一种主从设置中的分布式 SVM 算法。分布式 SVM 被视为正则化优化问题,并建模为一系列使用优化技术求解的凸优化子问题。达斯等人。 [11]提出了一种用于 SVM 训练的分布式、可扩展且通信高效的算法,该算法使用核矩阵的紧凑表示来减少训练期间的计算和存储。

培训过程。至于随机森林算法,[12]中提出了基于 Apache Spark 的大数据分类的并行版本。该算法使用结合数据和任务并行性的混合方法进行优化。

B.分布式无监督学习

在无监督学习中,聚类是最常用的任务之一,k-means 和 DBSCAN (基于密度的噪声应用空间聚类)是两种最流行的聚类算法。许多研究和开发工作都致力于通过并行化和/或分布式来提高其性能。例如,在[13]中

张等人。提出了一种基于数据并行分布方法和动态负载平衡参数服务器架构的k-means算法并行策略。

同样,在[14]Zhao等人中。提出了一种基于MapReduce编程范式的并行k-means聚类算法。最近,Balcan 等人在 [15] 中分析了分布式集群问题,其中数据在节点之间进行分区,这些节点的通信仅限于图结构的边缘。描述了一种具有低通信成本的分布式 k 均值算法,该算法基于充当整个数据集代理的一小组点的构造。在[16]秦等人。开发了一种用于无线传感器网络的分布式 k 均值算法,其中每个节点都配备了传感器。所提出的分布式实现能够将数据划分为具有小的组内距离和大的组外距离的组。

在 [17] Patwary 等人中。提出了一种使用图算法概念的并行 DBSCAN 算法,适用于共享和分布式内存系统。具体来说,他们利用不相交集数据结构来打破 DBSCAN 的内在顺序性,并使用基于树的方法来构建集群,确保工作负载平衡。同样,戈茨等人。 [18]提出了一种 DBSCAN 并行方法,该方法采用三种技术来打破算法的内在顺序性并增强分布式处理环境中的工作负载平衡。这些技术是: i)用于域分解的计算分割启发式; ii) 数据索引预处理步骤,以及 iii) 基于规则的集群合并方案。在[19]陈等人。提出了分布式环境中 DBSCAN 算法的并行版本,该算法通过对数据进行分区,然后每个节点独立构建集群,并将子结果聚合为一个

最后结果。[20] 中提出了一个非常相似的解决方案（即数据分区、并行集群构建和最终合并），其中 Luo 等人。利用分布式 Spark 框架。

对于大多数机器学习算法（即使是较少使用的算法），至少有一种专用于分布式环境的实现。例如，在[21]中，Pizzuti和Talia提出了AutoClass的分布式内存多计算机并行实现，AutoClass是一种基于贝叶斯分类的聚类算法。

C. 分布式机器学习框架

除了特定的算法之外，不同的并行/分布式软件框架还包括分布式机器学习库，例如，Apache Mahout [22] 是一个用于开发可扩展机器学习算法的开源库。它构建在 Hadoop 之上，包括推荐挖掘、聚类、分类和频繁项集挖掘算法。另一方面，MLlib [23] 是 Apache Spark 的机器学习库，它提供高级数据分析和并行机器学习算法，例如构建在 Spark [41] 之上的分类、回归、聚类和协作过滤。

四. 资源受限的机器学习

边缘设备

在边缘部署机器学习应用程序对于不同的现实世界应用程序场景 iOS 来说是一个关键机会，它可以受益于在数据源附近执行训练和推理所带来的低延迟（参见第七节）。然而，物联网边缘设备的性质（即有限的计算和能源、硬件和技术的异构性、安全和通信问题）对在此类设备上执行繁重的学习任务提出了巨大的挑战。这是一个相对较新的研究领域，已经提出了一些系统。这里我们认为边缘学习（即边缘学习）包括训练（即边缘训练）和推理（即边缘推理）过程。

A. 边缘训练 虽然推

理通常在边缘设备上执行，但边缘训练却不太常见[24]。在边缘设备中训练机器学习模型的技术的主要工作主要涉及深度学习。目标是获得轻量级深度学习模型，可以在边缘设备上协作学习。文献主要关注基于梯度下降技术进行训练的算法。一般来说，分布式梯度下降学习过程包括局部更新步骤，其中每个边缘设备执行梯度下降以改进局部模型参数，从而最小化其自己的局部数据集上的损失函数。然后，需要进行全局聚合步骤，将不同边缘设备获得的模型参数发送到聚合器，聚合器是通常在远程云上运行的组件。聚合后，更新后的参数被发送回边缘设备进行下一轮迭代。

按照这种方法，Wang 和合著者 [25] 提出了一种无需与云服务器合作即可在边缘训练机器学习模型的技术。该技术仅使用边缘设备来最小化学习模型的损失函数。局部梯度下降是在多个边缘设备上对本地数据执行的。本地模型被发送到另一个边缘设备（即聚合器），该边缘设备计算加权平均值并将其发送回所有边缘设备以进行下一个迭代步骤。作者仅使用三个 Raspberry Pi 设备和一台笔记本电脑作为实验设置，展示了该技术的有效性，在不同数据集上实现了接近最佳的性能。基于梯度下降的分布式学习在[26]中也从理论角度得到了广泛的研究。

B. 边缘推理

推理过程通常发生在边缘，以确保本地数据的低延迟和隐私。在本节中，我们讨论最近的工作，这些工作提出了用于在资源受限的边缘设备上推理并降低预测成本的框架和算法。例如，[27]中提出了一种名为 Bonsai 的基于树的算法，用于对物联网设备进行有效预测。Bonsai 能够保持准确性，同时最小化模型大小和预测成本。这是通过开发一个学习浅层稀疏树的树模型来完成的。然后将数据投影到学习树的低维空间中。Bonsai 在 Arduino Uno 板上进行了部署和评估。Gupta 和合著者在[28]中进行了类似的实验方法。作者提出了 ProtoNN，一种基于 k 最近邻 (kNN) 的算法，用于对资源受限设备进行准确预测。ProtoNN 基于三个关键方面：i) 学习少量原型来表示整个训练集，ii) 数据的稀疏低维投影，以及 iii) 投影和原型的联合判别学习。在 [29] Yazici 等人中，使用 10 个不同的数据集在 Raspberry Pi 上测试了三种不同的算法（随机森林、SVM 和多层感知器）。他们特别评估了推理过程时间、准确性和功耗方面的性能。

结果表明，随机森林算法的准确率最高，而SVM算法的推理速度更快，功耗更高效。为了减少集成模型的大小，在[30]中提出了一种在资源受限设备上的随机森林修剪方法，以优化成本和准确性。特别地，剪枝问题被提出为整数程序并用大规模原对偶算法来解决。

对于深度学习，在边缘设备中部署模型的常用方法是在强大的机器（即云或集群）上训练大型且准确的模型，然后使用压缩技术（即低秩近似、知识蒸馏、剪枝、参数量化）以减小尺寸。在这个方向上，Tiny机器学习范式是一个快速发展的领域，机器学习算法能够在设备上执行数据分析

具有极低的功耗。然而,压缩模型通常会导致精度较低[7],因此必须进一步研究精度和成本之间的权衡。

讨论的研究工作证明了减少数据传输和机器学习模型大小的技术的有效性,从而提高了资源受限的边缘设备的推理性能。然而,它们都不是为了部署在具有许多也可以协作执行推理过程的边缘设备的现实边缘计算环境中。

五、联邦机器学习

大多数分布式机器学习技术都以集中方式管理数据,而不考虑训练或推理过程中的隐私和安全问题。特别是,在训练过程中更新全局模型取决于边缘设备发送的信息,而边缘设备通常具有有限的防御能力,并且可能受到潜在攻击的影响。为了满足这一需求,联邦学习是一种机器学习范式,其目的是在数据保持分布在大量客户端的同时训练集中式模型[42]。虽然联邦学习广泛用于基于梯度下降的算法,但传统机器学习算法中的隐私问题尚未得到充分研究。

当联邦范式应用于集群时,主要目标是将存储在每个节点上的全局相似的本地数据分组。Kumar 和合著者 [31] 提议将联合平均技术应用于基于 [43] 提出的分布式版本的 k 均值算法。边缘设备产生的数据永远不会发送到集中节点,从而确保隐私保护和减少延迟。通过迭代模型平均得到最终模型。Dennis、Li 和 Smith [32] 开发了一种一次性联合集群方案,称为 k-FED,它只需要与中央服务器进行一轮通信。每个设备解决本地 k 均值问题,然后通过消息传递来传达其本地集群均值。

相反,[33] 中提出了一种使用联邦学习 SVM 来检测 Android 恶意软件的隐私保护联邦学习系统。它允许移动设备协作训练分类器,而不会暴露敏感信息。

我们致力于在集成技术中应用联邦学习范式,特别是随机森林算法,它是广泛工业场景中最常用的机器学习算法之一。

例如,在[34] Wu等人中。提出了 Pivot,一种保护隐私的垂直决策树训练和预测的解决方案,可确保不会暴露中间信息。所提出的解决方案还可以扩展到树集成模型,例如随机森林和梯度增强决策树。还是在垂直联邦学习领域,在[35] Yao等人中。提出了一种专为高效训练和推理而设计的联邦随机森林算法,以及一种利用随机森林并行性的分布式系统,实现高分区容错性。该系统涉及

一种高效的同态密码系统,可以提供数据隐私保护。Han等人还在[36]中提出了一种垂直联邦学习模型,名为联邦梯度提升森林 (Federated Gradient Boosting Forest),它通过并行构建决策树作为基础学习器,同时集成了Boosting和Bagging。在[37]中,Liu 和合著者重点研究了随机森林的隐私保护学习系统,该系统达到了与非隐私保护方法相同的准确性。更详细地说,开发了一个学习系统,可以使用相同的用户样本但不同的属性在不同的客户端上协作训练模型,而无需交换原始数据。还提出了一个预测过程来减少客户端之间的通信开销。

尽管这些提议表明联邦学习具有多种优势,尤其是可扩展性和数据隐私,但它们没有考虑边缘设备的硬件特性,而边缘设备通常在计算和存储资源方面受到限制。他们仅关注边缘分布式学习的隐私和安全方面,假设每个边缘设备都有可用的计算资源。

六.分布式机器学习
边缘-云连续体

在分布式环境中执行机器学习算法需要 (a)分离组成算法的任务,(b)根据不同计算节点之间存在的依赖关系来协调它们在不同计算节点上的执行,以及 (c)管理计算节点和不可靠的通信链路上可能发生的故障。因此,选择合适的分布式算法来解决给定问题不仅取决于问题的特征,还取决于运行该算法的系统的功能和配置、进程之间的通信和同步类型可以执行的。当我们考虑以有限的计算能力、能源消耗和延迟问题、不同的技术和软件堆栈为特征的物联网系统时,传统分布式系统的所有这些问题都会被放大。

我们在这里讨论的方法旨在使机器学习算法的分布式版本适应物联网环境。

如图 1 所示,在边缘-云连续体中,我们可以识别四个不同的层 [44]:

- 设备层。这是物联网设备生成或收集数据 (数据收集)的层。该数据通常可以以持久或临时方式 (数据存储)存储在设备的存储系统中。在存储或使用它在设备上执行分析之前,可以根据应用程序要求过滤数据 (数据过滤)。然后,该本地数据可用于训练学习模型 (本地学习),该模型在更高级别上可用于联邦学习任务。
- 边缘层。这一层是路由器、基站或微型数据中心等网关的作用,使计算更接近物联网设备。这种接近的目标

是收集物联网设备的感知数据（数据聚合），对其进行预处理并可能进行缓存（数据过滤和缓存）并将其发送到云（或雾）以进行存储或执行无法执行的复杂学习任务在这个级别或那个设备中。在这里可以聚合从许多设备学习到的本地模型（模型聚合）。

雾层。这是一个中间层，可以受益于更接近云端且比边缘层提供的计算更强大的计算。数据可以被存储和利用以进行集体学习。特别是，在设备级别对私有数据进行本地学习之后，可以在该级别进行集体训练阶段，其中通过基于共识的算法将标签分配给共享和未标记的数据。云层。该层充当网络的主干，并提供持久性数据存储以及其他层不可用的强大且非常大的处理资源[45]。如果需要，可以利用它来聚合全局模型（全局学习）。

并非上述所有级别都是必需的（例如，雾是可选的），但如果减少到设备-云两个极端级别，则架构将变成传统的基于云的解决方案。在这种四层架构上实现去中心化算法必须考虑以下几个方面：

1) 数据位置 - 计算尽可能靠近数据，以最大限度地减少节点之间的数据移动。由于各个级别由具有不同计算能力的异构软件组件组成，如果任务在给定级别的某个节点上无法执行（或效率低下），则需要向上级请求支持（等等）在）。例如，无法在设备上执行的任务被卸载到边缘节点，依此类推，将节点雾化到云端。

2) 地理分布 有必要考虑执行分布式算法的硬件组件的物理分布。

硬件组件可能位于彼此远离的不同位置。因此，设备的本地化可能会影响通信开销和算法性能。

3) 算法特征和配置 每种算法的适应方式都可以与其他算法不同，并且在这些环境中可能没有唯一的算法（代码）迁移方式。可能存在允许用户在物联网环境中定义和执行机器学习算法类的模式，但有许多变量需要考虑。此外，在混合云/边缘架构中部署分布式算法是一个极其复杂的问题，因为存在不同且丰富的有效配置参数（NP-hard问题[46]）。

4)任务调度和数据持久化 需要一个数据感知调度器来有效地执行以下任务：

编写分布式学习算法。调度程序必须确保所有计算节点之间的负载平衡（某些节点可以具有不同的计算能力），更加关注高度关键的任务，并在必要时执行任务的复制以提高执行时间和容错能力[47]。通常必须确定集中式或分布式主节点来协调所有这些调度活动。此外，请考虑临时数据可能存储在非持久组件上，因此可能会丢失。

5) 应用程序约束 - 必须尊重将使用机器学习算法的应用程序的功能和非功能约束（服务质量或 QoS）。通常，满足关键的应用程序和系统要求（例如低延迟、节能、减少网络流量、隐私保护和高可扩展性）至关重要。由于这些原因，机器学习算法也必须是可配置的才能满足要求。事实上，相同的算法可以在具有不同自由度的架构上以不同的方式执行（例如，边缘节点上的模型不太精确，云上的模型更精确）。

6) 编码和测试 - 考虑每一层可以由不同的硬件实现，并且可以由不同的软件工具/库编程。这使得编写和彻底测试算法变得困难。此外，使用大量硬件节点的实际测试可能非常昂贵且不灵活，并且基准测试和设置实际实验可能非常具有挑战性。仿真工具功能强大且灵活，可用于再现和测试物联网系统和网络[48]–[50]。

七.应用场景

必须使用分布式学习的许多不同的现实应用场景可能会受益于边缘-云连续体中边缘和云之间的合作。他们中的许多人需要实时计算、低能耗、高可扩展性和良好的隐私级别，可以从使用可能与云解决方案集成的边缘解决方案中受益。在这里，我们讨论利用所提出的方法的三个重要场景。

A. 智慧城市

边缘计算范式使车辆和人类能够连接和集成各种服务，以增强智能城市的个性化体验[51]。这可以使用能够从网络边缘的数据源提取知识的机器学习技术来完成。边缘云连续体可以为所有这些先进的移动服务提供广泛而高效的支持，例如 i) 将出租车移动到可能找到新客户的城市区域； ii) 根据位置和偏好向汽车驾驶员提供广告； iii) 根据之前访问过的地点向行人建议参观地点。智能可以嵌入到可以协作训练的边缘服务器和设备中

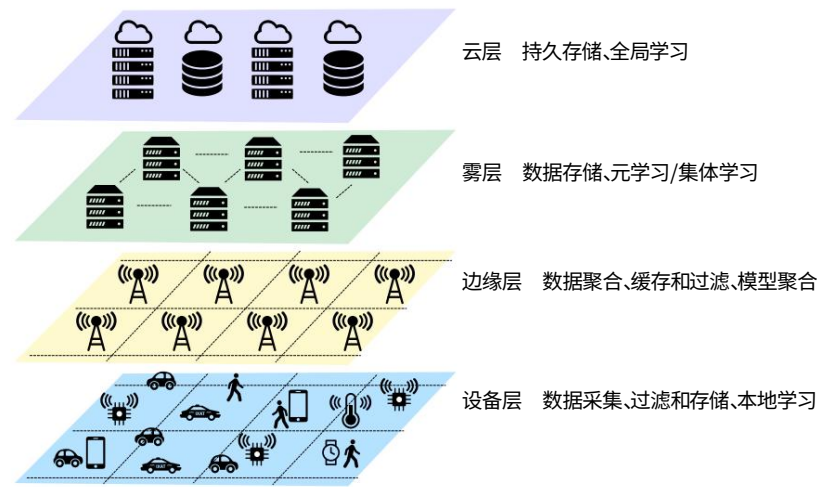


图 1:边缘-云连续体架构。

机器学习模型对其产生的数据进行建模,确保学习过程的可扩展性和本地数据的隐私保护。然后可以通过聚合在边缘学习的模型来在云层或雾层获得全局模型。

B. 工业物联网

工业物联网使机器能够提高工业流程的性能和生产力,同时减少浪费[52]。结合来自站点网络的数据可以更有效控制物料流,并及早发现、识别和消除生产或供应瓶颈,从而优化工业运营。

我们提出的分层架构可以实现先进的 IIoT 应用,例如: i) 在边缘执行机器学习任务,以减少响应时间和带宽; ii) 工业设备的预测性维护分析,以在潜在故障影响生产之前识别它们; iii) 先进的物流管理。

C. 智慧医疗

由于需要通过标准机制访问服务并扩展资源以执行繁重的智能任务,许多医疗保健组织已采用云计算解决方案。医疗设备被视为通过临床工作流程直接连接到网络的物联网应用程序的一部分。然而,各种类型医疗设备的可用性和复杂性不断增加,以及此类设备生成的大量数据,可能会限制当前基于云的物联网应用。在这种情况下,智能医疗解决方案[53]可以利用Edge-Cloud Continuum解决方案来持续监控和协助家庭患者,从医院设备中获取见解,以改善患者的治疗效果并优化医疗供应。

八. 结论

传统的分布式机器学习方法依赖于大颗粒分布式系统或云中的大数据存储和处理。他们要求传输所有数据

从源到集中服务器。收集、处理和分析它涉及高昂的通信成本,这对于低延迟应用程序可能至关重要。边缘计算范式的出现就是为了满足这一需求,通过在尽可能靠近数据生成的地方处理数据。尽管如此,网络边缘的物联网设备的计算能力和能量能力有限,这使得在此类设备上完全执行繁重的学习任务是不可行的。出于这些原因,必须像我们讨论的那样致力于调整机器学习算法,以对边缘设备上可用的本地数据执行协作训练和推理。

参考

[1] L. Belcastro, F. Marozzo and D. Talia, “大数据分析的编程模型和系统”,《国际并行、紧急和分布式系统杂志》,卷. 34,第 632-652 页,2019 年。

[2] W. Shi and S. Dustdar, “边缘计算的前景”,计算机,卷. 49,没有. 5,第 78-81 页,2016 年。

[3] A. Imteaj, U. Thakker, S. Wang, J. Li and M. Amini, “资源受限物联网设备联邦学习调查”,IEEE 物联网杂志,卷. 9,不. 1,第 1-24 页,2021 年。

[4] D. Rosendo, A. Costan, P. Valduriez and G. Antoniu, “边缘到云连续体的分布式智能:系统文献综述”,《并行与分布式计算杂志》,2022 年。

[5] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo and J. Zhang, “边缘智能:利用边缘计算铺平人工智能的最后一英里”,IEEE 论文集,卷. 107,没有. 8,第 1738-1762 页,2019 年。

[6] J. Chen and X. Ran, “边缘计算深度学习:综述”,IEEE 会议录,卷. 107,没有. 8,第 1655-1674 页,2019 年。

[7] M. S. Murshed, C. Murphy, D. Hou, N. Khan, G. Ananthanarayanan and F. Hussain, “网络边缘的机器学习:一项调查”,ACM 计算调查 (CSUR),卷. 54,没有. 8,第 1-37 页,2021 年。

[8] N. K. Alham, M. Li, Y. Liu and M. Qi, “用于可扩展图像分类和注释的基于 mapreduce 的分布式 svm 集成”,计算机与数学与应用,卷. 66,没有. 10,第 1920-1934 页,2013 年。

[9] F. Ozgur, C. Atak and M. Erdal Balaban, “基于 Mapreduce 的二元分类分布式 SVM 算法”,arXiv e-prints,第 10-11 页. arXiv-1312,2013。

[10] Q. Chen and F. Cao, “主从模式下的分布式支持向量机”,神经网络,第 1 卷. 101,第 94-100 页,2018 年。

[11] J. D. V. Sarin and R. N. Mahapatra, “用于分布式支持向量机训练的快速且通信高效的算法”,IEEE Transactions on Parallel and Distributed Systems,卷. 30,没有. 5,第 1065-1076 页,2018 年。

[12] J. Chen,K. Li,Z. Tang,K. Bilal,S. Yu,C. Weng 和 K. Li,“Spark 云计算环境中大数据的并行随机森林算法”,IEEE并行和分布式系统交易,卷. 28,没有. 4,第 919–933 页,2016 年。

[13] 张勇,熊正,毛建,欧丽,“并行k均值算法研究”,2006年第六届世界智能控制与自动化大会,2006年第1期. 2. IEEE,2006 年,第 5868–5871 页。

[14] W. Zhu,H. Ma 和 Q. He,“基于 MapReduce 的并行 k 均值聚类”,IEEE 国际云计算会议。施普林格,2009 年,第 674–679 页。

[15] M.-FF Balcan,S. Ehrlich 和 Y. Liang,“一般拓扑上的分布式 k 均值和 k 中值聚类”,神经信息处理系统进展,卷. 2013 年 26 日。

[16] 秦建,付伟,高浩,郑文新,“基于多智能体共识理论的传感器网络分布式k-means算法和模糊c-means算法”,IEEE控制论学报,2016年第1期. 47,没有. 3,第 772–783 页,2016 年。

[17] MMA Patwary,D. Palsetia,A. Agrawal,W.-k. Liao,F. Manne 和 A. Choudhary,“使用不相交集数据结构的新的可扩展并行 dbscan 算法”,载于 SC 12:高性能计算、网络、存储和分析国际会议论文集. IEEE,2012 年,第 1–11 页。

[18] M. Gotz,C. Bodenstein 和 M. Riedel,“Hpdbscan:高度并行 dbscan”,《高性能计算环境中机器学习研讨会论文集》,2015 年,第 1–10 页。

[19] M. Chen,X. Gau 和 H. Li,“具有优先级 r 树的并行 dbscan”,2010 年第二届 IEEE 国际信息管理与工程会议. IEEE,2010 年,第 508–511 页。

[20] G. Luo,X. Luo,TF Gooch,L. Tian 和 K.qin,“基于 Spark 的并行 dbscan 算法”,2016 年 IEEE 大数据与云计算国际会议 (BDCloud). IEEE,2016 年,第 548–553 页。

[21] C. Pizzuti 和 D. Talia,“P-autoclass:用于挖掘大数据集的可扩展并行集群”,IEEE 知识与数据工程期刊,卷. 15,没有. 3,第 629–641 页,2003 年。

[22] “Apache Mahout”,<https://mahout.apache.org/>,2022 年 6 月访问。

[23] “Apache Spark 的 MLlib”,<https://spark.apache.org/mlib/>,6 月访问 2022 年。

[24] N. ea Kukreja,“边缘训练:原因和方式”,2019 年 IEEE 国际并行和分布式处理研讨会. IEEE,2019 年,第 899–903 页。

[25] S. Wang,T. Tuor,T. Salonidis,KK Leung,C. Makaya,T. He 和 K. Chan,“当边缘遇到学习时:资源受限分布式机器学习的自适应控制”,IEEE INFOCOM 2018 - IEEE 计算机通信会议,2018 年,第 63–71 页。

[26] Y.Zhang,MJ Wainwright 和 JC Duchi,“统计优化的通信高效算法”,神经信息处理系统进展,卷. 2012 年 25 日。

[27] A. Kumar,S. Goyal 和 M. Varma,“物联网的 2 kb 内存中的资源高效机器学习”,国际机器学习会议. PMLR,2017 年,第 1935–1944 页。

[28] C. Gupta,AS Suggala,A. Goyal,HV Simhadri,B. Paranjape,A. Kumar,S. Goyal,R. Udupa,M. Varma 和 P. Jain,“Protonn:压缩且准确”knn 用于资源稀缺设备”,国际机器学习会议. PMLR,2017 年,第 1331–1340 页。

[29] MT Yazici,S. Basurra 和 MM Gaber,“边缘机器学习:实现智能物联网应用”,大数据和认知计算,卷. 2,没有. 3,第 3 页. 2018 年 26 日。

[30] F. Nan,J. Wang 和 V. Saligrama,“修剪随机森林以进行预算预测”,神经信息处理系统进展,卷. 2016 年 29 日。

[31] HH Kumar,V. Karthik 和 MK Nair,“联合 k 均值聚类:一种新颖的基于边缘人工智能的隐私保护方法”,2020 年 IEEE 新兴市场云计算国际会议 (CECM). IEEE,2020 年,第 52–56 页。

[32] DK Dennis,T. Li 和 V. Smith,“异质性取胜:一次性联合集群”,国际机器学习会议。PMLR,2021 年,第 2611–2620 页。

[33] R.-H. 许,Y.-C. 王,C.-L. 范,孙,T. Ban,T. Takahashi,T.-W. 吴,和S.-W. Kao,“基于边缘计算的 Android 恶意软件检测的隐私保护联邦学习系统”,2020 年第 15 届亚洲信息安全联合会议 (AsiaJCIS). IEEE,2020 年,第 128–136 页。

[34] Y. Wu,S. Cai,X. Xiao,G. Chen 和 BC Ooi,“基于树的模型的隐私保护垂直联合学习”,arXiv 预印本 arXiv:2008.06170,2020。

[35] H. Yao,J. Wang,P. Dai,L. Bo 和 Y. Chen,“一种高效且稳健的垂直联合随机森林系统”,arXiv 预印本 arXiv:2201.10761,2022 年。

[36] Y. Han,P. Du 和 K. Yang,“Fedgbf:通过梯度提升和装袋的高效垂直联邦学习框架”,arXiv 预印本 arXiv:2204.00976,2022 年。

[37] Y.Liu,Y.Liu,Z.Liu,Y.Liang,C.Meng,J.Zhang 和 Y.Zheng,“联邦森林”,IEEE 大数据汇刊,2020 年。

[38] J. Qiu,Q. Wu,G. Ding,Y. Xu 和 S. Feng,“大数据处理机器学习综述”,EURASIP 信号处理进展杂志,第 1 卷. 2016 年,没有. 1,第 1–16 页,2016 年。

[39] C. Savaglio,P. Gerace,G. Di Fatta 和 G. Fortino,“物联网边缘的数据挖掘”,2019 年第 28 届计算机通信与网络国际会议 (ICCCN). IEEE,2019 年,第 1–6 页。

[40] T. Kraska,A. Talwalkar,JC Duchi,R. Griffith,MJ Franklin 和 MI Jordan,“MLbase:分布式机器学习系统”,在 Cidr,卷. 1,2013 年,第 2–1 页。

[41] L. Belcastro,R. Cantini,F. Marozzo,A. Orsino,D. Talia 和 P. Trunfio,“大数据分析编程:原理和解决方案”,大数据杂志,卷. 9,没有. 2022 年 4 月。

[42] J. Konecny,HB McMahan,FX Yu,P. Richtarik,AT Suresh 和 D. Bacon,“联邦学习:提高通信效率的策略”,arXiv 预印本 arXiv:1610.05492,2016。

[43] G. Jagannathan 和 RN Wright,“任意分区数据上的隐私保护分布式 k 均值聚类”,第十一届 ACM SIGKDD 国际数据挖掘知识发现会议记录,2005 年,第 593–599 页。

[44] MS Aslanpour,SS Gill 和 AN Toosi,“云、雾和边缘计算的性能评估指标:未来研究的回顾、分类、基准和标准”,物联网,卷. 12,p. 100273, 2020。

[45] D. Talia,P. Trunfio 和 F. Marozzo,云中的数据分析:模型、技术和应用.爱思唯尔,2015 年 10 月,iSBN 978-0-12-802881-0。

[46] 陶锋,赵德,胡勇,周子,“制造网络系统中基于粒子群优化的资源服务组合及其优化选择”,IEEE工业信息学汇刊,2014年第1期. 4,没有. 4,第 315–327 页,2008 年。

[47] S. GiampA.L. Belcastro,F. Marozzo,D. Talia 和 P. Trunfio,“用于执行大规模分布式工作流程的数据感知调度策略”,IEEE Access,卷. 9,第 47 354–47 364 页,2021 年,iSSN:2169-3536。

[48] A. Barbieri,F. Marozzo 和 C. Savaglio,“通过参数扫描进行物联网平台和服务配置:基于仿真的方法”,2021 年 IEEE 国际系统、人与控制论会议 (SMC),17 2021 年 10 月–20 日,第 1803–1808 页。

[49] M. Sinqadu 和 ZS Shibeshi,“使用 ifogsim 对交通监控应用程序进行性能评估”,无线智能和分布式通信环境国际会议。施普林格,2020 年,第 51–64 页。

[50] L. Belcastro,A. Falcone,A. Garro 和 F. Marozzo,“边缘计算环境中大规模 roi 挖掘应用的评估”,第 25 届国际分布式仿真和实时应用研讨会 (DS-RT),2021 年,第 1–8 页。

[51] LU Khan,I. Yaqoob,NH Tran,SA Kazmi,TN Dang 和 CS Hong,“支持边缘计算的智慧城市:全面调查”,IEEE 物联网杂志,卷. 7,没有. 10,第 10 200–10 232 页,2020 年。

[52] Q. Wang,X. Zhu,Y. Ni,L. Gu, and H. Zhu,“物联网和工业物联网的区块链:综述”,物联网,第 1 卷. 10,p. 100081, 2020。

[53] S. Tian,W. Yang,JM Le Grange,P. Wang,W. Huang,Z. Ye,“智慧医疗:让医疗更智能”,《全球健康杂志》,第 1 卷. 3,没有. 3,第 62–65 页,2019 年。