云计算环境下人工智能技术优化路径探析

李姗珊[1]　卢　杰[2]　贾诗音[1]
1.湖北商贸学院 湖北 武汉 430000
2.桂林理工大学 广西 桂林 541000

【摘　要】：在云计算中，人工智能技术大规模部署面临着资源分配、数据传输和体系结构协同的多重约束。当前，异构计算资源调度无序，导致计算效率低下；跨域数据传输机制响应慢，模型训练周期长；云边协作框架尚未建立，智能推理实时性难以保证。针对以上问题，从资源调度、传输信道和体系结构设计三个方面展开研究。构建面向异构资源的统一调度框架，融合 GPU、FPGA 和 TPU 等多种计算资源，提高资源利用率和任务执行效率。利用带宽预留和数据压缩技术拓宽跨域数据传输信道，减少模型迭代时延。构建层次化协作云边缘计算架构，实现推理任务向边缘节点下沉，缩短端到端响应时间。

【关键词】：云计算环境；人工智能技术；优化路径

DOI:10.12417/2705-0998.26.05.086

以云计算与 AI 相结合的实际情景为基础，从语义阐释、问题辨识和优化三个方面进行研究。围绕云计算背景下人工智能技术的优化研究，从挖掘云运算潜力促进建模迭代、降低技术应用门槛以普惠中小场景部署、构建弹性服务体系以支持即时智能反应等三个层次对其进行研究。对现有的多个技术聚合所遇到的实际难题进行分析，由于异质资源的无序调配而引起的计算效率低下；由于跨域数据的传播速度较慢而导致的学习延迟较大，并且由于云端边缘协作的缺乏而引起的推断反应延迟。从优化异质资源调度、拓展数据传输信道和建立云端边缘协作框架三方面入手，最终形成一条完整的解决方案。

1 云计算环境下人工智能技术优化路径的探析意义

1.1 释放云端算力潜能，加速模型训练迭代

云计算环境下人工智能技术优化的首要意义是通过灵活调度算力资源和改进分布式训练机制，充分释放云计算潜力，大幅缩短模型训练周期。传统方式下，计算资源以固定规格分配，训练任务独占物理设备，造成计算资源闲置和任务排队并存的矛盾。云计算的资源池化特性将分散的 GPU 集群集成到统一的、可调度的计算资源池中，从而支持多个任务的并行处理。基于此，分布式学习框架，通过将大规模模型参数分解为多个计算节点协同计算、梯度同步和参数更新并行进行的方法，将原来需要几天时间才能完成的训练任务压缩到几个小时之内。在超参优化环节，引入自寻优机制，取代人工逐轮调试模式，在预定义的参数空间并行抽样评价，快速收敛到最优配置组合。通过以上三个方面的提升，实现云计算能力由静态分配向动态响应转变，持续提升模型训练性能，为人工智能大规模部署奠定计算基础[1]。

1.2 降低 AI 应用门槛，普惠中小场景部署

云计算平台通过对基础设施的具体设计和规范的服务界面进行包装，将人工智能从一个强大的高科技企业独有的功能，转化为面向中小型应用的、易于使用的公众服务。在预训练模型管理中，云计算中已有几十种典型的图像分类、物体检测和自然语言处理等典型实例，在使用过程中，不需要考虑模型的构造和权值，只需要将服务数据进行上载就可以得到相应的推断，而整个过程只需要 500 ms 就可以完成。在自动学习服务中，构建特征工程、算法筛选和评价的自动化程序，通过对已有标签的数据进行处理，完成数据清洗、特征交叉和模型训练等过程，最后生成符合要求的可快速展开的算法，降低人力投入 80%左右。

1.3 构建弹性服务架构，支撑实时智能响应

基于云计算的人工智能服务体系结构，该体系具有毫秒级的可扩展性和快速的推理信道。通过三条主要的研究思路：服务格部署、推理加速引擎的整合和柔性扩展的自主扩展。在服务网格的部署上，通过边缘－车辆代理方式，对各智能推理业务实例进行单独的业务处理，实现对请求路由、负载平衡和保险丝降级的一体化调度，实现 50 ms 内的业务发现时延。在算法的整合上，通过对已建立的算法进行量化和运算的融合，实现浮点参数的 8 比特整数化，并将邻近的运算符整合到单独的运算单位中，从而降低内存和内核间的传输时间，使得单个运算时延由 200 ms 缩短到 30 ms 以下[2]。

作者简介：
第一作者：李姗珊（1998.04-），女，汉族，湖北襄阳人，硕士研究生，高级工程师，研究方向：人工智能、云计算、智能计算、优化算法。
第二作者：卢杰（1998.06-），男，汉族，江西九江人，硕士研究生，工程师，研究方向：云计算、大数据。

2 云计算环境下人工智能技术优化中存在的问题

2.1 资源调度秩序紊乱，算力利用效率偏低

在云计算背景下，AI 的培训往往被应用在由图处理器、张量处理器和 CPU 组成的异质机群上。由于不同厂家的处理器具有明显的指令集结构、内存带宽和核心数目等方面的差别，且已有的多个并行处理算法均基于通用的资源提取，没有考虑到底层的硬件性能。以 GPU 为例子，在一家公共云计算平台上进行了实时监控，发现采用这种方法时，单个显卡的运算效率可以高达 75%，但是平均性能却只保持在 45%—55%左右，起伏在 30%以上。究其根本原因，是由于该算法没有充分利用处理器的计算强度和处理器浮点处理的特性，而忽略了算法的计算强度和处理器浮点处理的适应性。而在多个用户的共用簇中，由于存在着资源的争夺和锁定竞争，使得调度的无序更加严重[3]。

2.2 数据跨域传输迟缓，模型训练时长增加

在云计算中，由于存储节点和计算节点的隔离，使得其在学习过程中面临着学习的“瓶颈”问题。在目标内存中，当运算开始时，需要将其从内存中取出到内存节点的缓存中。在高度并行的访问模式下，内存输出带宽因多任务争用而产生激烈竞争，致使单一任务的实际读取吞吐率由理想状态下的 2.5 GB/s 降低至 1.2 GB/s。于是，读时延由毫秒提升到秒量级，CPU 在等待的时候会进入空闲，CPU 的运算能力闲置时间可以达到整个训练时间的 30%—40%。在不同区域的培训环境下，数据的传递是非常重要的问题。由于分布式学习过程中，多个节点间存在着频繁的梯度信息交互，且各可用区域内的带宽一般不到其带宽的 1/10，且随着节点数目的增大，延时成指数增长[4]。

2.3 云边协同机制缺失，智能推理响应滞后

云边协作的不足主要体现在云计算和边界推断的连接上存在缺陷，云计算中的模型在云计算中已经被训练完毕，却缺少将其传输到边缘终端的过程。目前，在实际应用中，需要通过算法工程师对已有的数据进行人工导出和压缩，然后将其上传到边界设备的镜像库中，并进行周期性的拉取。该过程在装置数目不多的情况下可以保持，但是一旦边界结点的数目超过 100 个，人工作业的时间限制就暴露出来。由于模式从钟点扩展到天区，且边界设施仍然使用多天前的老模式，导致推断的准确性与云计算的实际情况有较大差异。这种偏向因训练样本和推断样本之间的不一致性而更加严重。云计算中的训练数据主要来自以往的学习和仿真，而实际的学习环境受到光照、传感器噪声和网络扰动等多种因素的干扰，使得其与训练样本之间的统计特性发生偏移。由于缺少自动反馈和重培训的方法，边推断精度会随着时间的推移而逐渐降低，实际测试结果表明，经过 30 天的持续运行，一幅影像的推断精度降低 12 个百分点。

3 云计算环境下人工智能技术优化的有效策略

3.1 优化异构资源调度，提升算力利用效率

在云平台上，AI 的培训任务对于计算能力的要求也是不同的，CPU 更适于进行逻辑控制和数据预处理，而图和张量则更善于并行，而可编程门阵列则更适合于低时延分析。提高计算效率的核心是构建针对异质数据的精细调度机制。基于“资源池化抽象”“任务特征描述”和“动态调度”三个步骤。对于资源池，将 CPU 核心、显卡和张量处理器等核心集成到资源管理体系中，并在各个资源节点上分别装载代理，并定时报告可用资源和目前的负荷状况。其中，CPU 的颗粒级最少为 1000 个核，而显卡的计算则是兆字节的内存，而 CPU 的计算则是以百万亿的浮点计算来计算。对于任务特性的描述，在提出学习任务时，需要说明所要采用的框架类型，模型的参数数量和样本数量[5]。

基于数据的多维数据驱动的分布式计算方法，并将其分为计算、存储和通讯三种类型。针对运算密度较大的应用，如卷积神经网络，将高频率的图处理器和张量处理器作为主要的运算资源；针对存储需求较大的应用，如大型的嵌入式系统，将其作为主要的存储单元，并启动存储切换机制。对于动态的计划，使用了二段式的排程方法。对资源的预定，由系统对所需的资源进行预测，并按照优先到优先的原则，在任务队列中为每一个任务保留一个位置，保留 30 秒，超过 30 秒后，将其释放并再次进入队列。抢占调度，通过对高优先权的工作进行评价，如果这个工作已经存储工作点并且超出了预定的临界值，那么就暂停这个工作，然后将它的工作分配给更高的工作。该程序以 5 分钟为单位，对网络中的长时间空闲或超满的资源进行分析，并对其进行动态的任务分配。

3.2 拓宽数据传输通道，降低模型训练时延

在分布式人工智能系统地学习过程中，网络的梯度信息和网络模型的参数传递是限制其性能的重要原因。为拓展数据传输信道，减少网络延迟，需要在网络拓扑结构优化、数据压缩传输以及通讯和运算叠加三个层次上进行系统性的研究。对于网络的拓扑结构，基于肥树形或者非堵塞的 Ridge 结构来代替三层网络。将交换机配置到服务器柜的顶端，利用多条上行链路将各个交换机接入汇聚层，链路的聚集带宽不小于 400 吉比特每秒；在路由器的缓存中设置明确的阻塞通告机制，如果路由器的缓存队伍超出一定的阈值，就会给发送者阻塞标志，由发送者相应地减少传输率；研究一种新型的远距离存储接入机制，使得该算法可以在不通过 OS 核心的情况下，实现对其他节点的存储信息的读取，使得单个通信时延从毫秒量级下降到亚毫秒量级[6]。

为减少通讯开销，在分级同步过程中采用压缩方法。运算过程是：在每个运算结点进行局部小规模学习之后，对所求的

梯度张量进行量化，将 32 比特的浮点数据进行压缩，得到 8 比特或 16 比特的浮点。利用稀疏处理方法，只留下 10%的最大值作为传递信号，将剩余的部分设置为 0。接收机接收到原始信号后，对信号进行去噪处理和信号的稀疏性处理，并根据信号的历史梯度来估算出信号中的信息。压缩速率随实际情况而变化，一般情况下设定 50%，拥堵状态下提高到 75%。在信息和运算的交叠过程中，采用分层传递和下一次正演算法进行迭代。其方法为：当各计算节点通过逆传递得到完备的梯度值后，不再等其他节点的梯度累加，直接将当前的梯度值进行转发，并进行下一批的正向运算。采用双重缓存技术，实现数据的传递和运算的分离，并在此基础上实现数据转发和数据处理的交互。

3.3 构建云边协同架构，加快推理响应速度

在云训练与边缘推理相结合的框架下，云端到边缘的数据传递时延和边缘节点本身的计算能力受限，成为推理过程中的“瓶颈”。为此，本项目拟从“轻量化压缩”“边缘存储部署”和“云边缘协作”三个方面展开研究。对大规模的云数据进行建模，并对其进行精简处理。以原始大规模数据为训练对象，对未标注数据进行统计推断，获得软标记的概率分布。从教师网络的训练样本中抽取五分之一作为学生网络的训练样本。学生网络采用可分离卷积层替代传统卷积神经网络中的标准卷积层，其单次前向传播的计算时间约为原始网络的六分之一。利用蒸馏损失函数对学生网络加以训练，使其输出逼近教师网络产生的软标记分布。经上述实验流程后，所得学生模型在推理速度提升三至五倍的同时，分类准确率仍保持在百分之九十五以上。

在边界缓存配置中，将模式缓冲层配置于边界节点和邻近使用者的集中节点。根据建模频率和实时性需求，设置缓存机制，频率较高的模块驻留在内存中，而在运行过程中将其保存在磁盘中，在运行过程中将其加热到记忆体中。高速缓存利用最新最小占用算法，在高速缓存内存不够的情况下，剔除长期没有被调用的模式。基于增量式的替代方法，当模型中的变量超出临界值时，只会将有差别的部分推出来，并将所需的数据量缩减到原来的 5%—10%。针对云计算环境下的协作计算问题，基于信任门槛的任务转移策略。边界节点根据使用者的要求，进行一种轻型模式的初始推断，并给出相应的预测值和可信度分数。如果可信度分数大于预定的门限值，则将该结果直接送回；当其小于临界值时，将其特征矢量或者中层特性进行打包，并将其上载到云计算平台，再通过云计算对其进行第二次精细化判定。当网络的负荷超过 70%的时候，可以将网络的流量上限提高 10%，从而缓解网络的负荷。

4 结语

综上所述，面向云计算的人工智能算法，其关键科学问题是：解决算法性能的瓶颈和数据的传递限制，实现对海量数据的建模和快速的智能化处理。通过充分挖掘云计算的计算潜力，大大降低建模迭代周期，大幅提高技术创新的效率；通过减少对中小规模场景的使用要求，使其有更多的中小规模场景实现智能升级；通过建立灵活的业务体系结构，有效支持了系统的即时反应能力。通过对云原生和 AI 架构的深入适配和云端边缘的协作机理的不断改进，云计算中的 AI 算法将实现高效、低延迟、强可伸缩等方面的发展，为各行各业的智慧变革奠定更扎实的基础。

参考文献：

[1] 王宇慧,祝川惠,康丽馨.云计算环境下的审计模式变革研究[J].商业经济,2025,(10):170-172+183.

[2] 曾思敏.云计算环境下高校会计信息化建设策略[J].中国乡镇企业会计,2025,(16):10-12.

[3] 方卫洪.云计算环境下计算机信息安全技术的深度应用与策略分析[J].中国战略性新兴产业,2025,(18):48-50.

[4] 李海燕.基于云计算的大数据技术分析与应用[J].集成电路应用,2025,42(02):182-183.

[5] 唐小波.云计算环境下的数据安全管理策略优化与技术研究[J].电工技术,2024,(S2):124-125+128.

[6] 张峰.基于人工智能的云计算环境智能运维技术应用[J].电脑编程技巧与维护,2024,(12):151-154.