 DOI: https://doi.org/10.12345/bdai.v7i3.38195

Cloud Computing Resource Load Prediction Method Based on Deep Learning

Zhanqi Chu

Henan University of Science and Technology, Luoyang, Henan, 471023, China

Abstract

With the widespread adoption of cloud computing technology in internet services, industrial internet, and big data platforms, data centers exhibit highly dynamic demands for computing, storage, and network resources. Fluctuations in resource loads not only impact system operational efficiency but also significantly affect service quality and resource utilization rates. Traditional statistical model-based prediction methods demonstrate limitations when handling complex nonlinear load variations. Deep learning technologies, leveraging their advantages in feature extraction and time series modeling, provide innovative solutions for cloud computing resource load forecasting. This study focuses on resource load prediction challenges in cloud computing environments, analyzes the application principles of deep learning in cloud platform load prediction, and constructs a predictive model framework based on deep neural networks. Through systematic exploration of data processing, model training, and performance evaluation, the research aims to enhance cloud platform resource scheduling efficiency and system stability.

Keywords

deep learning; cloud computing; load forecasting; resource scheduling; time series

基于深度学习的云计算资源负载预测方法

楚展奇

河南科技大学，中国 · 河南 洛阳 471023

摘　要

随着云计算技术在互联网服务、工业互联网以及大数据平台中的广泛应用，数据中心的计算、存储和网络资源需求呈现出高度动态化特征。资源负载的波动不仅影响系统运行效率，也会对服务质量和资源利用率产生重要影响。传统基于统计模型的预测方法在处理复杂非线性负载变化时存在一定局限。深度学习技术凭借其在特征提取和时间序列建模方面的优势，为云计算资源负载预测提供了新的解决思路。本文围绕云计算环境下资源负载预测问题，分析深度学习在云平台负载预测中的应用原理，并构建基于深度神经网络的预测模型框架，对数据处理、模型训练与性能评估进行系统探讨，以提升云平台资源调度效率和系统稳定性。

关键词

深度学习；云计算；负载预测；资源调度；时间序列

1 引言

云计算平台通过虚拟化技术整合计算资源，为用户提供按需分配的计算服务。在实际运行过程中，用户请求量具有明显的波动特征，导致计算节点的资源利用率不断变化。资源负载过高会造成系统响应延迟，而资源配置过多又会降低整体利用效率，因此准确预测云计算平台的资源负载具有重要意义。传统预测方法多依赖线性统计模型或简单时间序列模型，在处理复杂业务场景时难以捕捉高维特征与长期依赖关系。随着人工智能技术的发展，深度学习在时间序列分析与模式识别领域表现出较强优势，为资源负载预测提供了新的技术路径。通过构建深度神经网络模型，可以在海量运行数据中挖掘潜在规律，提高预测精度并为云资源调度提供支持。

2 云计算资源负载预测的理论基础

2.1 云计算资源负载特征分析

在云计算平台运行过程中，系统资源负载通常表现为多维度动态变化，包括 CPU 利用率、内存占用率、网络带宽使用情况以及存储访问负载等指标。这些指标共同反映了数据中心的整体运行状态。由于用户访问行为具有明显的不确定性，资源负载数据往往呈现周期性波动与突发性变化并存的特点。在互联网服务平台中，业务访问量通常与时间周

【基金项目】“数字化实验教学平台建设研究”（项目编号：2021BK122）。

【作者简介】楚展奇（1973–），男，中国河南洛阳人，硕士，实验师，从事计算机网络，云计算，人工智能研究。

期密切相关，例如白天访问量较高而夜间相对较低，而在大型促销活动或热点事件发生时，系统负载可能在短时间内迅速上升。资源负载数据在时间序列上通常具有明显的非线性特征，传统线性模型在描述这些变化时容易出现误差积累的问题 [1]。因此，在云计算环境中进行资源负载预测，需要综合考虑数据波动性、周期性以及突发变化等因素，并通过更加灵活的建模方式捕捉系统运行规律。

2.2 资源负载预测的研究意义

准确的资源负载预测能够为云平台资源管理提供重要决策依据。通过对未来负载变化趋势进行预测，云平台可以提前调整虚拟机数量或容器资源配置，从而保证系统在高负载情况下仍然保持稳定运行。在数据中心规模不断扩大的背景下，资源管理效率对企业运营成本具有直接影响。负载预测不仅能够减少资源浪费，还能够提高计算资源利用率，使平台在保证服务质量的同时实现成本优化。对于云服务提供商而言，负载预测模型还可以辅助自动化调度系统进行决策，使资源分配更加合理。通过将预测结果与调度算法结合，可以在负载上升之前完成资源扩展，从而避免系统拥塞或服务中断。

2.3 深度学习在负载预测中的优势

深度学习模型在处理复杂数据结构方面具有显著优势。相比传统统计方法，深度神经网络能够自动提取数据中的高层特征，并在多层网络结构中逐步形成对复杂模式的表达能力。在云计算环境中，大量运行日志和监控数据构成高维时间序列数据，深度学习模型能够从中提取潜在规律。例如循环神经网络能够在时间序列建模中捕捉长期依赖关系，而卷积神经网络则能够通过局部特征提取识别数据变化趋势 [2]。通过结合多种深度学习结构，可以构建更加灵活的预测模型，从而提升资源负载预测的准确性和稳定性。

3 基于深度学习的负载预测模型构建

3.1 数据采集与预处理方法

在云计算资源负载预测过程中，数据质量对模型性能具有重要影响。数据来源通常包括监控系统采集的 CPU 利用率、内存占用率、网络流量以及磁盘读写速率等指标。由于原始监控数据可能存在噪声和缺失值，需要在模型训练前进行预处理。数据清洗是数据预处理的重要环节，通过异常值检测与缺失值填补，可以提高数据完整性。时间序列数据通常还需要进行归一化处理，使不同量纲的数据在统一尺度下进行计算，从而提高模型训练效率。为了捕捉系统运行规律，还需要对数据进行时间窗口划分，将历史数据转换为适合深度学习模型输入的序列结构。经过预处理的数据能够更加准确地反映系统运行状态，为后续模型训练提供可靠基础。[3]

3.2 深度神经网络预测模型设计

在负载预测模型设计中，可以采用多层神经网络结构对时间序列数据进行建模。循环神经网络能够通过隐藏状态记录历史信息，在时间序列预测任务中表现出良好的性能。为了克服传统循环网络在长序列训练中的梯度消失问题，可以采用长短期记忆网络结构。该模型通过引入门控机制，使网络能够在不同时间尺度上保留重要信息，从而提高预测精度。在模型结构设计过程中，还可以结合卷积网络对输入数据进行特征提取，使模型在时间维度和特征维度上同时学习数据规律。通过构建多层网络结构，模型能够逐步提取复杂特征并形成对资源负载变化趋势的预测能力。

3.3 模型训练与参数优化

深度学习模型在训练过程中需要通过大量历史数据不断调整网络参数，使预测结果逐渐接近真实值。在训练阶段，通常采用均方误差作为损失函数，通过反向传播算法更新模型参数。训练过程中需要合理选择学习率和批量大小，以保证模型能够稳定收敛。为了防止模型在训练数据上出现过拟合，可以采用正则化方法或随机失活机制，使模型在不同数据样本上保持较好的泛化能力。通过不断优化网络结构和训练参数，可以逐步提升模型预测性能，使其能够适应不同云平台环境下的负载变化特征。

4 深度学习负载预测模型的性能分析

4.1 预测精度评估方法

在云计算资源负载预测研究中，模型性能评价是判断预测方法有效性的重要环节。由于云平台负载数据具有波动性强和变化频繁的特点，研究者通常通过多种评价指标对预测模型进行综合分析。平均绝对误差与均方根误差被广泛应用于预测性能评估，这两类指标能够直观反映预测值与真实观测值之间的偏差程度。当误差值保持在较低水平时，说明模型在刻画系统负载变化趋势方面具有较好的稳定性。均方根误差在评价较大偏差时更加敏感，因此能够反映模型在异常波动环境下的预测能力 [4]。除误差指标之外，还可以利用相关系数评价预测序列与真实序列之间的相关程度。当相关系数接近较高水平时，表明模型能够有效反映负载变化的整体规律。通过多维度评价指标进行综合分析，可以更加全面地判断预测模型的性能表现，并为模型结构优化和参数调整提供可靠依据，从而提升负载预测研究的科学性。

4.2 不同模型对比分析

在资源负载预测研究中，模型性能往往需要通过多种算法之间的比较加以验证。传统预测方法在早期研究中得到广泛应用，其中移动平均模型能够通过平滑历史数据反映整体趋势，回归分析模型则通过变量关系建立预测方程，自回归时间序列模型利用时间序列特征进行趋势推断。这些方法在数据结构较为简单的情况下具有一定稳定性，但在云计算环境中，系统负载往往呈现复杂的非线性变化特征，使传统方法在预测精度方面受到一定限制。深度学习模型通过多层网络结构能够提取高维特征信息，并对复杂数据模式进行有

效表达。神经网络在训练过程中可以不断调整权重参数，使模型逐渐适应数据变化规律。在实验对比过程中，深度学习模型通常能够获得更低的预测误差，并在负载波动较大的情况下保持较高稳定性，这种优势使其在云计算资源预测领域逐渐成为重要研究方向。

4.3 实验结果与应用效果

在实验研究过程中，将深度学习预测模型应用于云平台历史监控数据，可以对系统未来负载变化进行较为准确的预测。通过对大量运行数据进行训练与测试，模型能够识别资源使用过程中存在的周期性变化与突发波动规律。实验结果表明，预测模型在短时间尺度内具有较高精度，能够有效反映系统负载变化趋势。当预测结果显示负载水平即将上升时，云平台可以根据预测信息提前进行资源扩展，例如增加计算节点或启动新的虚拟机实例，以保证系统运行稳定。预测结果在资源调度决策中的应用，使平台能够在业务高峰到来之前完成资源准备，从而减少服务延迟和系统拥塞问题。通过将预测模型嵌入自动化资源管理系统，可以形成以数据分析为基础的智能调度模式，使云计算平台在复杂运行环境中保持较高效率，并进一步提升整体资源利用水平。

5 深度学习负载预测在云平台中的应用

5.1 云资源调度优化

在云计算平台运行过程中，资源调度系统承担着协调计算资源与业务需求之间关系的重要任务。随着云平台规模不断扩大，计算节点数量和业务类型逐渐增多，传统基于静态规则的调度方式已难以满足复杂环境下的资源管理需求。通过引入深度学习负载预测模型，可以利用历史运行数据和实时监测信息对未来资源需求进行分析，从而为调度系统提供决策依据[5]。当系统预测到某一时间段资源需求将出现增长趋势时，调度系统能够提前启动虚拟机或容器实例，并对资源进行合理分配，使计算资源在业务高峰到来之前完成准备。这种基于预测结果的资源调度方式能够减少服务响应延迟，提升系统整体处理效率。同时，将预测模型与自动化调度算法结合，可以在资源分配过程中实现更加精细化的管理，使云平台在不同负载条件下保持稳定运行，从而提升云计算环境的整体服务能力。

5.2 数据中心能耗管理

数据中心作为云计算基础设施的重要组成部分，其运行能耗一直受到广泛关注。服务器设备在高负载状态下运行会消耗大量能源，而在负载较低时仍维持大量设备在线，则容易造成能源浪费。通过应用深度学习负载预测技术，可以对未来一段时间内的资源需求进行分析，从而为数据中心能耗管理提供科学依据。当预测模型判断系统负载将处于较低水平时，可以通过动态调整服务器运行状态，使部分设备进入节能模式或暂时停止运行。随着业务需求增加，再逐步恢复计算资源供给，从而保持系统性能与能源利用之间的平衡。

5.3 智能运维与系统稳定性提升

在大规模云计算平台中，系统运维工作涉及众多服务器节点与网络设备，运行环境复杂且变化频繁。传统运维方式往往依赖实时监测数据进行故障处理，当系统负载突然增加时，可能导致性能下降甚至服务中断。通过在运维管理系统中引入负载预测模型，可以在系统运行状态发生变化之前进行预判，使运维工作从被动响应逐渐转向主动管理。预测模型能够根据历史运行数据识别负载变化规律，当系统预测结果显示某一节点即将出现资源紧张情况时，平台可以提前执行资源迁移或负载均衡操作，使任务在多个节点之间进行合理分配。通过这种方式，可以有效避免局部节点过载问题，从而提升系统运行稳定性。将深度学习预测技术融入运维管理体系，有助于构建更加智能化的云平台管理模式，使系统在复杂业务环境下依然保持稳定可靠的运行状态。

6 结语

随着云计算平台规模不断扩大，资源管理复杂性逐渐提升，准确的负载预测成为提升系统效率的重要手段。深度学习技术在时间序列分析与特征提取方面具有显著优势，为云计算资源负载预测提供了新的技术路径。通过构建基于深度神经网络的预测模型，可以有效挖掘系统运行数据中的潜在规律，提高预测精度并为资源调度提供支持。研究表明，深度学习模型在负载预测任务中表现出良好的性能，能够为云平台智能化管理提供技术基础。

参考文献

[1] 朱义霞.云计算环境下数据中心网络负载均衡策略研究[J].信息与电脑,2026,38(04):170-172.

[2] 何世勋.基于云计算的计算机实验室网络多路径拥塞控制技术研究[J].信息记录材料,2026,27(02):145-147.

[3] 朱睿,蒋雪阳,孙彦赞.面向依赖任务的D2D辅助边缘计算卸载与资源分配优化[J].计量与测试技术,2025,51(12):6-9+15.

[4] 吴明涛,钱鸣静,许梦.改进SSA优化VMD-CNN-BiLSTM的云资源负载预测方法[J].长春师范大学学报,2025,44(12):50-63+73.

[5] 段玉峰,张媛琪.基于云计算架构的高职计算机教学资源共享平台系统规划与设计[J].软件,2025,46(11):93-95.