



基于数据和知识驱动的超万卡智算集群稳定性保障实践

牛红韦华¹, 黄永宝¹, 丁国强², 黄宝¹, 赵治稳¹, 徐杨¹, 王涛¹, 张锐龄¹, 王旋¹, 张逸翔¹

(1. 中移(苏州)软件技术有限公司, 江苏 苏州 215100;

2. 中国移动通信集团设计院有限公司, 北京 100080)

摘要: 为了解决超万卡智算集群硬件故障多、任务训练故障率居高不下、跨域问题定位困难等稳定性保障问题, 提出了一种基于数据和知识驱动的保障超万卡智算集群稳定性的方案。首先, 通过异构资源一体化采集技术、分布式实时大数据抽取—转换—加载(extract-transform-load, ETL)技术采集集群性能数据; 然后, 基于改进的自注意力机制的双向长短期记忆(self-attention-based bidirectional long short-term memory, SA-BiLSTM)网络深度学习模型实现故障诊断; 最后, 通过知识图谱分析匹配诊断模型输出的结果, 完成故障诊断报告的输出, 提升诊断模型输出的可解释性。在深度学习模型提取时序性特征时引入特征权重系数, 对不同尺度提取的特征加权融合, 提高模型故障诊断精度。在基于1.8万卡智算集群故障诊断仿真实验中, 损失值逐渐收敛并稳定在0.047, 准确率达到了98.4%。实践表明, 该稳定性保障方案能有效保障大模型训练, 提升智算集群的可靠性, 为未来更大规模的智算集群建设与大模型训练提供坚实的基础。

关键词: 智算集群; 故障诊断; SA-BiLSTM; 知识图谱

中图分类号: TP393

文献标志码: A

doi: 10.11959/j.issn.1000-0801.2025151

A data and knowledge-driven practice for ensuring stability in ultra-large intelligent computing clusters

NIU Hongweihua¹, HUANG Yongbao¹, DING Guoqiang¹, HUANG Bao¹, ZHAO Zhiwen¹,

XU Yang¹, WANG Tao¹, ZHANG Ruiling¹, WANG Xuan¹, ZHANG Yixiang¹

1. China Mobile (Suzhou) Software Technology Co., Ltd., Suzhou 215100, China

2. China Mobile Group Design Institute Co., Ltd., Beijing 100080, China

Abstract: A data and knowledge-driven stability assurance scheme for such clusters was proposed to address the issues of frequent hardware failures, persistently high task training failure rates, and difficulties in cross-domain problem localization within ultra-large intelligent computing clusters with over ten thousand computing cards. The cluster performance data was collected by employing heterogeneous resource integrated collection technology and distributed real-time big data ETL techniques. Fault diagnosis was performed using an enhanced SA-BiLSTM deep learning model, improving the explainability of diagnostic model outputs via knowledge graph analysis and matching for the

收稿日期: 2025-03-12; 修回日期: 2025-06-06

通信作者: 黄永宝, huangyongbao@cmss.chinamobile.com



generation of fault diagnosis reports. In the process of extracting time series features with the deep learning model, weighted fusion of features extracted at different scales, thereby improving the accuracy of the fault diagnosis model. In fault diagnosis simulation experiments conducted on an 18 000-card cluster, it was observed that the loss value gradually converged and stabilized at 0.047, achieving an accuracy rate of 98.4%. Practical has shown that the proposed stability assurance scheme can effectively support large-scale model training and enhance the reliability of intelligent computing clusters, providing a solid foundation for the construction of larger-scale intelligent computing clusters and the training of large models in the future.

Key words: intelligent computing cluster, fault diagnosis, SA-BiLSTM, knowledge graph

0 引言

随着人工智能 (artificial intelligence, AI) 技术取得突破性进展, 大模型与智算的紧密结合成为驱动技术进步与革新的关键力量。近年来, 大模型的发展势头尤为迅猛, 如 Llama3.1、GPT-4、DeepSeek V3、Claude-3.5-Sonnet 等, 它们通常包含数十亿甚至数千亿的参数, 需要处理和学习的数量也是前所未有的。业界紧密遵循规模定律^[1], 该定律为大模型的扩展提供了重要的理论依据和实践指导, 但也面临着诸多挑战。随着模型规模的增加, 训练任务涉及千万颗元器件满负荷高速运转, 固有的元器件硬件失效率和海量的器件规模带来硬件故障频发, 涉及的软硬件故障模式繁杂, 故障诊断和治理挑战巨大。超万卡集群急需更有效、更快速且影响更小的故障治理方案来提升模型算力利用率, 实现大模型训练任务长时间不间断地高效运行, 以应对当前故障率居高不下的挑战。

当前, 全球智算领域已形成多个万卡级集群建设范式, 中国移动率先实现单集群 1.8 万个智算加速卡的规模化部署, 成功突破单集群部署规模的技术瓶颈, 依托上述超万卡集群稳定性保障的实践经验, 本文创新性地提出基于数据知识驱动的新型稳定性保障方案。首先, 打造异构资源一体化采集能力, 实现整个智算集群的计算、存储、网络性能监控, 为后续的故障诊断提供实时标准化的多维数据输入; 其次,

提出利用分布式实时大数据进行抽取-转换-加载 (extract-transform-load, ETL) 处理、采用 ClickHouse 作为实时数据库和历史数据库, 采用 Elasticsearch 作为 NoSQL 数据库用于存储实时采集的原始日志数据; 最后, 输出基于自注意力机制的双向长短期记忆网络 (self-attention-based bidirectional long short-term memory, SA-BiLSTM) 模型的智算设备故障诊断以及知识图谱的诊断报告, 确保超万卡智算集群模型训练的高效性和稳定性, 最大限度避免频繁故障导致的训练任务中断、资源浪费、模型收敛速度慢、运维成本增加等问题, 集群稳定性得到加强, 为超万卡智算集群大模型训练故障诊断提供新的解决思路 and 方案。

1 背景

1.1 智算集群与大模型发展趋势

随着 AI 技术的快速发展, 智算集群和大模型训练正经历着深刻变革。从 2018 年 BERT 模型^[2]的 1.1 亿参数, 到 2021 年 GPT-3 模型^[3]的 1 750 亿参数, 再到 2024 年 Llama-3.1-405B 模型^[4]突破 4 050 亿参数, 以及 GPT-4 更是达到了 1.8 万亿参数, 模型规模持续刷新纪录。模型参数规模的急剧扩张以及对计算资源需求的增长, 促使智算集群从千卡级别向超万卡级别跨越。在超万卡智算集群中, 即使单个节点的故障率非常低, 整体故障率仍然呈指数级增加, 会导致训练任务中断或失败, 这极大地影响了大模型任务训练的稳定

性, 延长了训练时间, 增加了维护成本, 给大模型的训练带来了重大的挑战。

1.2 超万卡集群大模型任务训练稳定性挑战

尽管大模型高效训练已取得一定进展, 但在超万卡集群环境下, 由于节点数量庞大、资源类型不统一、数据规模巨大以及网络环境复杂, 实现高效稳定的训练仍面临诸多挑战。尤其是在长时间稳定训练和故障快速诊断方面, 虽然失败和时延是大模型训练的常态, 但其故障成本极高。因此, 如何快速感知故障、精准定位问题、预测潜在风险并缩短故障恢复时间至关重要。在训练过程中, 可能遇到的故障类型主要包括硬件故障、软件故障、网络故障以及其他异常情况, 具体表现如下。

(1) 任务故障概率高。从千卡到超万卡规模, 通信问题成为主要故障源, 包括任务中断、网卡故障、网络超时等; 同时图形处理器 (graphics processing unit, GPU) /神经网络处理器 (neural processing unit, NPU) 异常、作业下发失败、光模块故障等问题导致智算集群故障率呈指数增长, 系统可靠性急剧下降。这些故障会引发训练中断、卡顿或失败, 严重影响训练效率。在大规模分布式训练中, 单点故障可能引发级联效应, 导致故障扩散路径复杂、排查困难、恢复缓慢等问题出现。当前, 超万卡集群亟须具备快速精准的故障诊断和最小化影响的治理能力。

(2) 万卡集群组网复杂度高。万卡级智算集群因其系统复杂度高、规模庞大、层次繁多, 显著增加了运维难度。大模型训练对网络性能要求严苛, 丢包、时延抖动等网络问题会直接影响训练效率。同时, 海量的服务器、网络设备、光模块等组件构成的多层次通信链路, 使得集群组网复杂度呈指数级增长, 导致系统稳定性和可用性显著下降。这种高度复杂的系统环境使得故障预测和根因分析面临巨大挑战。

(3) 故障诊断难度大。为保障断点续训能力, 应建立实时故障感知、定位与预测机制。当前主要面临三大挑战: 故障影响具有隐蔽性, 部分故障类型缺乏明显告警指标, 导致定界困难; 诊断效率提升存在瓶颈, 传统方法依赖海量日志分析, 耗时且低效; 状态感知缺失, 网络流量拓扑不透明, 难以及时获取集群节点真实性能状态; 这些因素共同导致智算集群面临严重的可观测性挑战, 影响了故障的快速诊断, 亟须构建更高效的监控分析体系。

1.3 超万卡智算集群故障治理现状

目前, 业界通过缩减检查点开销、提升故障预警识别能力、增加任务容错机制等方式实现分钟级故障定位并自动恢复, 不断提升智算集群稳定运行能力, 如Meta在训练Llama 3.1时广泛使用NVIDIA 集合通信库扩展版本 (NVIDIA collective communication library extended version, NCCLX) 进行日志记录与存储, 并通过更改在线配置, 达到无须代码发布或者作业重启而恢复训练的效果^[4]; 文献[5]在对Meta的训练结果进行分析后, 提出通过识别并移除“柠檬节点” (即性能较差或频繁出现故障的节点), 和通过添加适应性路由以绕过故障区域的方式提高训练可靠性; 谷歌Gemini利用CPU内存优化了最优检查点放置策略和检查点流量调度算法, 通过定期发送心跳信号来监测系统状态, 当故障发生时, 会根据故障类型和检查点的可用性来选择合适的恢复机制, 从而实现快速恢复^[6]; 腾讯自研端到端的全栈网络运营系统, 将集群整体故障的排查时间由天级降低至分钟级; 华为盘古大模型在2 048卡训练30天不中断, 断点恢复时长控制在10 min以内; 字节跳动万卡集群采用在GPU上创建守护进程, 定期发送封装了各种性能监控数据的心跳信号的方式, 对集群进行稳定性和性能的监控, 在检测到异常信号后会发出告警, 随后进行人工调查, 恢复训练^[7]; 上海人工智能实验室提出了基于大



语言模型 (large language model, LLM) 对故障作业的日志进行动态编写正则匹配和更新过滤规则, 从而达到辅助自动化诊断的效果^[8]; 文献[9]提出了一个针对分布式大模型规模化训练的实时诊断框架, 采用创新的内核级追踪和全局聚合指标, 可快速定位并解决训练异常; 文献[10]提出了一种基于日志分析的 L4 框架, 通过跨作业过滤 (cross-job filtering)、空间和时间模式分析, 识别故障指示性日志、节点和迭代, 从而提升大规模大语言模型 LLM 训练故障诊断的效率和准确性; 文献[11]提出了一种名为 FALCON 的诊断框架, 该框架能够快速识别 GPU 或通信链路的故障性问题, 并通过创新的多级缓解机制, 提高了对人类标注的故障性检测准确率。

尽管现有方案已取得阶段性成果, 但依然面临以下四大挑战: (1) 故障检测存在较明显的时延; (2) 故障检测需要占用较多的训练硬件资源; (3) 人工介入依赖度过高 (如字节跳动方案中的人工诊断环节); (4) 复杂故障应对能力不足 (如 Meta 与上海人工智能实验室方案)。具体而言, 瞬时硬件故障的精准捕获、诊断引发的资源占用等问题较为突出。尽管现有措施已取得一定成效, 但仍须进一步优化与创新。为此, 本文建议采用更先进的解决方案, 如实时监测、智能诊断和预测性维护等, 以持续提升系统的稳定性和性能表现。

2 智算集群性能采集

故障诊断和治理必须有数据的支撑, 大型云计算服务提供商均致力于提供多元化的智算产品, 公有云平台展现出算力类型丰富、智算集群规模庞大以及基础设施技术架构多样化的显著特征。鉴于大型模型训练任务高度依赖数以千万计的元器件以满载状态高速运行, 这种超大集群的海量器件往往伴随着硬件故障频发以及软硬件故障模式的复杂性增加。因此, 实施全方位、精细粒度且低时延的运行状态实时监测变得至关重

要。基于集群的实时运行数据, 可以被用来精准识别潜在的故障风险。本文提出了一种高度可靠的智算集群数据采集方法, 实现了秒级时延下的数据采集、高效传输、数据清洗与存储等一整套技术方案。该方法旨在解决超大规模且多样化的异构智算集群所面临的稳定性管理技术挑战。

2.1 异构多样智算集群性能采集

针对异构多样的智算集群的性能采集, 本文构建了一套综合性一体化采集体系如图 1 所示, 全面覆盖计算域、存储域以及网络域内的各类资源与设备层面, 实现了对包括英伟达、昇腾、壁仞等国内外主流算力集群高效与灵活地宽 (InfiniBand, IB) 和融合以太网上的远程直接内存访问 (RDMA over converged ethernet, RoCE) 等主流高性能无损网络技术的适配与集成能力, 从而确保了智算集群远程直接内存访问 (remote direct memory access, RDMA) 网络层面的数据覆盖。

(1) 计算域性能采集

本文自主研发了一款高性能的采集器 Agent, 该 Agent 被设计为能够全面且精确地捕获计算域各类关键性能指标、核心软硬件的运行日志以及告警事件信息。通过部署于服务器及交换机上的多种采集工具, 包括但不限于 nvidia-smi、npu-smi、ipmitool、bmc、telemetry 工具以及 shell 脚本等, 自研的采集器能够全面采集计算域各类性能指标。这些指标包括但不限于 AI 卡的使用率、功耗、温度、电压、显存使用率、AI Core 利用率、AI Core 当前频率、AI CPU 利用率、纠错码 (error correction code, ECC) 单比特错误数、ECC 双比特错误数, 服务器的中央处理器 (central processing unit, CPU) 使用率、内存使用率、整机功耗, 存储设备的占用情况, 以及光模块的温度、电压、电流、上下行功率等。

(2) 网络域性能采集

智算领域高性能无损网络主要分为 IB 和 RoCE 两大主流技术架构, 通常设备供应商均会

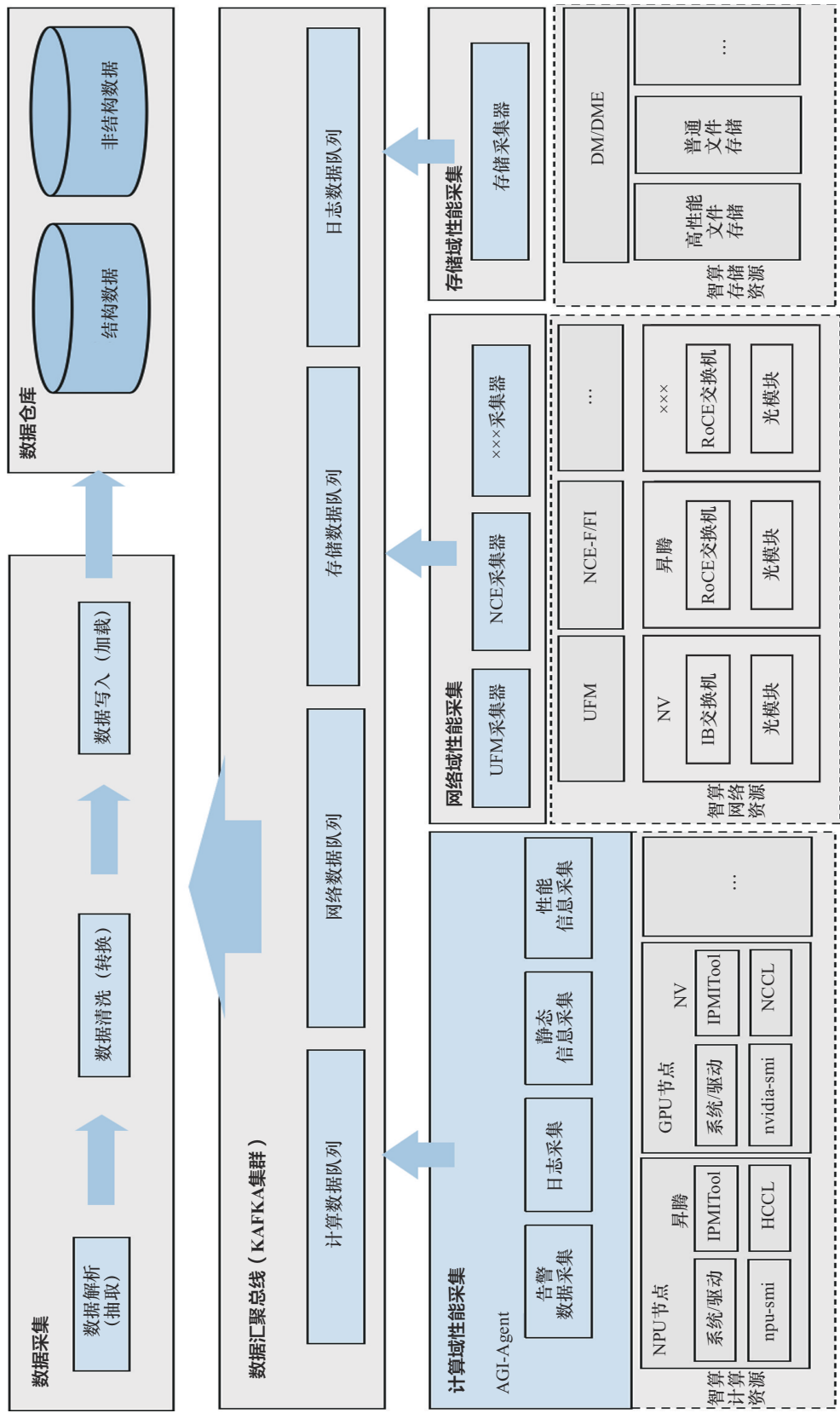


图1 综合性一体化采集体系



提供网络管控系统。通过深度集成设备厂商所提供的原生生态管控能力，如英伟达的UFM、华为的网络云化引擎（network cloud engine, NCE）等，实时获取网络域拓扑结构、各类关键性能指标、核心软硬件的运行日志以及告警事件信息。性能指标主要包括但不限于：交换机的CPU利用率、内存利用率，交换机端口的收发数据速率、收发数据量、收发数据包速率、收发数据包数、收发错误包数、收发丢包数，以及光模块的温度、电压、电流、上下行功率等。

（3）存储域性能采集

智算领域高性能存储厂商众多，通过深度集成设备厂商所提供的原生生态管控能力，实时获取存储域各类关键性能指标、核心软硬件的运行日志以及告警事件信息。性能指标主要包括但不限于：存储服务器的CPU利用率、内存利用率、读写每秒操作数（operations per second, OPS）、总OPS、读写带宽、总带宽；硬盘的使用率、读写带宽、总带宽、读写每秒输入/输出操作数（input/output operations per second, IOPS）、总IOPS；端口的读写带宽、总带宽、读写IOPS、总IOPS等。

系统将各类异构非标准化的采集数据实时推送到数据总线，通过数据处理层对数据进行解析、清洗、标准化等一系列处理后再持久化到数据仓库，从而构造整个智算集群的数据采集底座，为后续的故障诊断提供实时标准化的多维数据输入。

2.2 智算集群海量性能数据处理

2.2.1 海量数据 ETL

超万卡智算集群规模下，算存网设备的性能指标和软硬件日志数据量庞大。计算设备原始性能数据单日数据量超亿条，容量达百GB级，网络设备性能指标单日数据量约为2亿条，存储设备性能指标单日数据量超千万条，设备日志及软件日志单日上报数据容量达TB级。且所采集的数据类型不统一，性能指标数据以标准JSON方式传输，属于半结构化数据，而软硬件日志数据

属于非结构化的数据。

为解决海量半结构化和非结构化数据难以统一处理的问题，本文提出一种分布式实时大数据ETL方案。针对指标数据、日志数据以及分析工具的产生埋点数据，分别适配不同数据源形成统一数据传输总线，再通过数据抽取模块选择不同特点的数据源完成采集数据的抽取；接下来将抽取出来的原始数据推送到实时处理引擎模块，实时处理引擎加载自定义的数据提取模板，实现数据流的实时解析，同时模板提取控制程序可以实现数据提取模板的可插拔和热加载，解析后的结果加载到联机分析处理（on-line analytical processing, OLAP）数据库。分布式实时大数据ETL方案流程如图2所示。

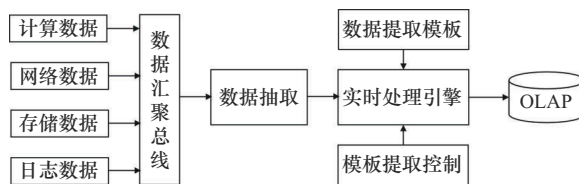


图2 分布式实时大数据ETL方案流程

海量数据实时处理的数据总线和数据抽取模块解决了数据格式不统一以及分布式采集难以灵活抽取的问题。数据处理引擎采用分布式实时处理方案，实现了智算集群软硬件监控数据秒级解析能力。不同数据结构的数据分别存储在不同类型的数据库中，实现了海量数据实时分析和离线分析的功能。

2.2.2 海量数据存储

经过解析清洗后的数据分为半结构化数据和结构化数据，通过以下3种形式存储：实时数据库、历史数据库和NoSQL数据库。系统将实时处理后的结构化数据存于实时数据库和历史数据库，将设备日志、软件日志原始数据存入NoSQL数据库中。

本文采用ClickHouse作为实时数据库和历史数据库，ClickHouse是一个用于联机分析的列式数据库管理系统（database management system，

DBMS)。实时采集的数据经过数据处理引擎处理后的性能数据存储在实时数据库中，当数据存储时间超过3个月或其他自定义周期，该部分数据则转存至历史数据库中。存于ClickHouse的实时数据可以为智算集群故障的诊断提供实时数据源，历史数据可以作为故障诊断模型的训练数据集。

针对半结构化数据，本文选用ElasticSearch作为NoSQL数据库，用于存储实时采集的原始日志数据。ElasticSearch基于Apache Lucene构建，是一个分布式搜索和分析引擎，具备可扩展的数据存储和矢量数据库功能。系统通过ElasticSearch，实现了智算集群原始日志的高效检索和离线数据分析。

3 智算集群故障诊断

本文通过日志异常检测模型、时序异常检测模块以及知识图谱技术模块的有机结合，设计了一套高效的智算集群故障诊断与分析方案。故障元信息主要来源于日志异常检测和设备性能时序数据异常检测。其中，日志异常检测是基于对集群软硬件日志的错误信息，规则化匹配识别潜在

的故障信息，属于常规的故障感知途径。设备性能数据异常检测是基于智算场景下通过特定的算法来智能化的故障感知手段。本文将侧重于后者，详细描述智算集群故障诊断方案，智算集群故障诊断架构如图3所示。

- 日志异常检测模块。该模块具备日志聚类、相关性分析以及异常场景分析这三大核心功能。日志聚类作为整个日志分析模块的基石，主要从Kafka中实时获取日志数据。随后，借助启发式规则对数据进行深度处理，最终生成日志模板库。相关性分析处于日志异常检测的关键环节。此阶段以日志模板库为依托，从ElasticSearch中提取历史数据，为了生成精准的日志分布数据，采用卡方、分箱算法等多种算法相结合的方式进行深入分析得出的日志分布数据。异常场景分析运用最长公共子序列（longest common subsequence, LCS）算法对日志分布进行精准识别，定位典型异常场景。
- 时序数据异常检测模块。系统基于改进

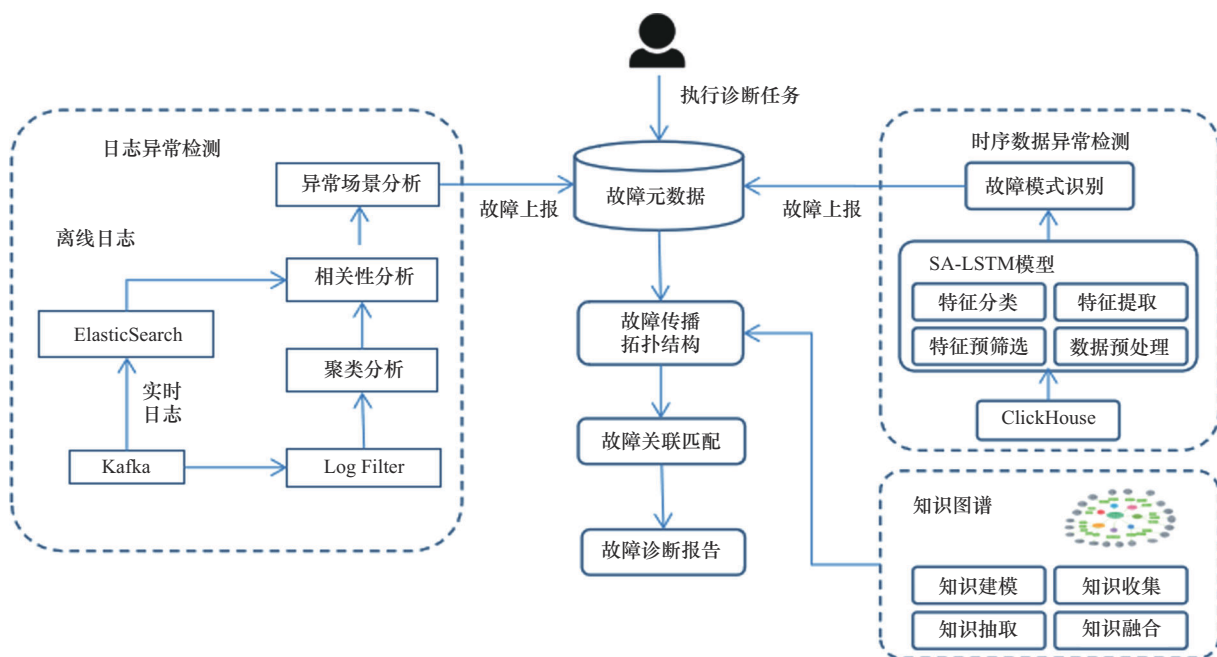


图3 智算集群故障诊断架构



双向长短期记忆 (bidirectional long short-term memory, BiLSTM) 网络^[12-14]深度学习算法构建故障诊断模型。该模块包括数据预处理、特征预筛选、特征提取和特征分类功能, 其将收集的原始数据进行分段处理, 引入自注意力机制^[15-16]对时序数据的特征进行预筛选; 将 BiLSTM 的隐藏层数据再输入批归一化 (batch normalization, BN) 层, 对小批量具有关联关系的记忆单元做 BN 处理, 实现对关键特征的强化; 使用全连接层与丢弃层融合进行特征分类, 最终识别智算设备时序数据故障模式。

- 知识图谱模块。该模块包括知识建模、知识收集、知识抽取和知识融合功能^[17]。知识建模功能对大量典型故障案例进行深入分析和总结; 知识收集功能将日志异常与时序数据异常等历史数据整合进历史资料, 融合专家经验与案例, 经数据预处理, 按故障类型和严重程度进行标注; 知识抽取功能负责提取出其中最具代表性和规律性的特征信息, 并将这些信息进行整合和归纳; 知识融合功能将不同数据源知识进行融合。以上功能构建起了全面、系统的故障知识库和故障传播链。

日志异常检测模块与时序数据异常检测模块实时检测智算设备故障, 利用知识图谱进行故障根因溯源。通过知识图谱所构建的复杂关系网络, 系统能够从故障现象出发, 沿着故障传播链逆向追踪, 逐步排查可能导致故障发生的各种因素。结合故障知识库中的相关知识和历史案例, 系统可以快速定位故障的根本原因。

3.1 基于 SA-BiLSTM 模型的智算设备故障诊断

本文提出一种基于数据和知识驱动的智算设备故障诊断方案, 方案在 BiLSTM 模型的基础上引入自注意力 (self-attention, SA) 机制构建 SA-

BiLSTM 诊断模型, 并结合长短期记忆网络基础模型进行优化改进, 结合智算集群的故障知识图谱, 实现 GPU、NPU、高带宽存储器 (high bandwidth memory, HBM)、ECC 错误、GPU 驱动器卡死、光纤收发器故障、网卡过热等故障的高准确性智算集群故障诊断方法。SA-BiLSTM 模型基于自注意力机制并加入 BN 处理和丢弃层的 BiLSTM 优化算法模型, 在仿真实验中表现出良好的故障诊断效果。基于 SA-BiLSTM 网络优化模型结构如图 4 所示。

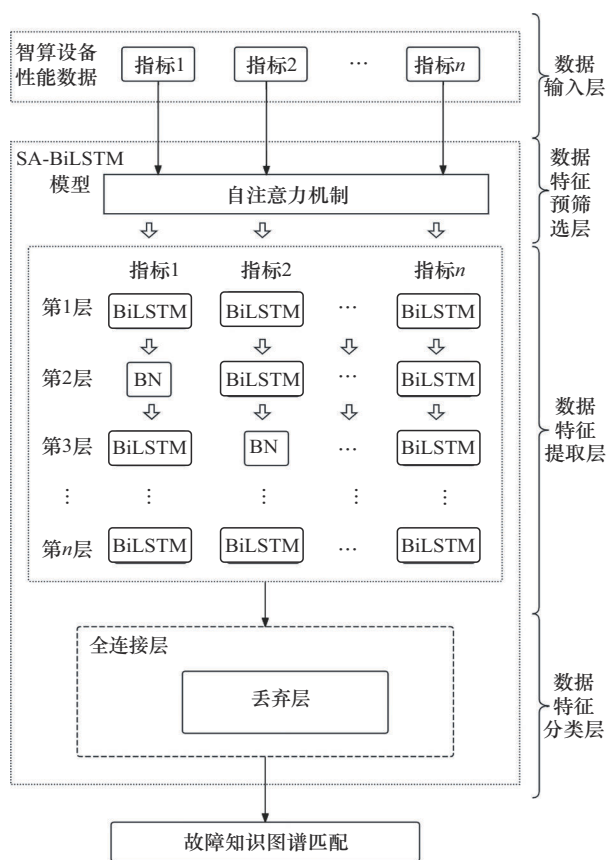


图4 基于SA-BiLSTM网络优化模型结构

模型通过引入自注意力机制对智算设备的性能数据特征进行预筛选, 有效缩小了智算设备性能数据的数据量级, 模型的性能得到了很大提升。通过进行特征提取计算, 在 BiLSTM 模型中引入 BN 处理, 提升了模型训练时的迭代速度, 在全连接层加入丢弃层防止模型出现过拟合问题。本方案从模型

的功能层面大致分为数据输入层、数据特征预筛选层、数据特征提取层、数据特征分类层。

3.1.1 数据输入层

本文所使用的智算设备的性能数据比较繁杂，每种设备的数据采集频率、数据类型、数据单位等都不尽相同，所以需要在数据输入层做模型数据的预处理，拼装不同特征的数据，将收集的原始信号进行分段处理，构造满足时间长度等要求的数据集。部分性能指标见表 1。

表 1 性能指标

性能指标	指标标识
CPU 活跃使用率	cpu_usage_active
内存使用百分比	mem_used_percent
GPU 温度	gpu_temperature
GPU 功耗	gpu_power_draw_watts
NPU 温度过高	npu_info_temperature
光模块接收功率	receive_power
光模块发送功率	transmit_power
光模块 TX1 偏置电流	tx1_bias
光模块 TX2 偏置电流	tx2_bias
端口接收数据速率	port_mb_in_rate
端口发送数据速率	port_mb_out_rate
端口接收包速率	port_pck_in_rate

3.1.2 数据特征预筛选层

由于智算设备性能数据量较为庞大，因此本文设计了数据特征预筛选层，起到对输入数据压缩、过滤、筛选的作用，并进一步引入自注意力机制对性能数据的特征进行预筛选。自注意力机制通过模仿人类视觉选择性注意能力，可以对输入序列进行注意力计算，给不同的元素分配不同的权重信息以获取序列内部的联系。为了从多个特征的多个子空间信息中提取到重点信息，可通过该机制学习映射关系来计算不同特征之间的权重，进而构建全局的关联性信息，实现对输入序列的全局关注，每个序列中的元素与该序列中的其他元素进行注意力权重计算；自注意力机制的权重参数完全是由数据驱动得到的，本文先根据

特征的重要程度初始化可学习参数 W_Q 、 W_K 、 W_V ；经过多轮模型训练，后期再由自注意力机制反向传播 W_Q 、 W_K 、 W_V ；通过可学习参数与输入序列 $X \in \mathbb{R}^{n \times d}$ 运算变换得到查询 Q 、键 K 、值 V 3 个向量， $Q = XW_Q$ ， $K = XW_K$ ， $V = XW_V$ ，注意力权重系数计算式如下。

$$C_{\text{Attention}} = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right)$$

$$\text{Attention}(Q, K, V) = C_{\text{Attention}} V = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (1)$$

其中， $\sqrt{d_k}$ 为缩放因子防止梯度消失。

自注意力机制的核心优势在于动态捕捉不同特征之间的上下文关系，利用注意力权重系数与值 V 向量的加权求和，实现不同尺度特征的加权融合，使模型能够自适应调整不同尺度特征的贡献，显著提升对复杂故障（如智算集群光链路故障、慢节点、慢网络等）的诊断能力，得到最终注意力输出。大量的智算设备性能数据直接通过模型训练学习，往往会出现耗时较长，结果不佳的效果，本文通过自注意力机制进行注意力权重计算，过滤掉无特殊特征的数据，起到数据筛选的作用，从而提取重要特征，为后续的 BiLSTM 减少层级深度，还可以有效避免模型的梯度消失或梯度爆炸问题。

3.1.3 数据特征提取层

为了提高本模型在智算设备的性能数据特征提取能力和训练速度，本文在 BiLSTM 的基础上引入了 BN，将 BiLSTM 的隐藏层数据再输入 BN 层，对小批量具有关联关系的记忆单元做 BN 处理，实现对关键特征的强化，并抑制干扰信号，提高模型的故障诊断能力。BiLSTM 的网络模型结构如图 5 所示。

BiLSTM 模型分为输入层、前向传播层、反向传播层、输出层，前向传播层和反向传播层共同连接输出层，每个单元结构的输出结果由前面

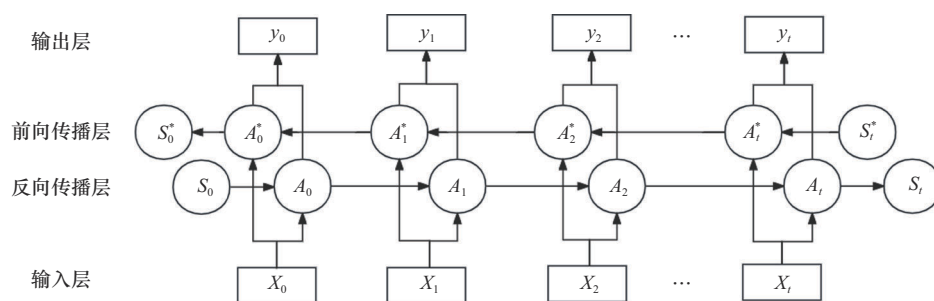


图5 BiLSTM的网络模型结构

的若干输入和后面的若干输入共同决定，使输出结果更加准确。

特征提取层采用若干不同层级的BiLSTM，为了实现模型的快速收敛，在特征提取层中又加入了BN处理，在隐藏层中加入不同数量的BN层，每层BiLSTM中记忆单元的数量各不相同，记忆单元个数越多，提取数据中的高频特征的能力就越强。通过调整每个层级BiLSTM网络的记忆单元数量，可以实现对不同特征提取的时间步长进行个性化调整，从而提取原始数据中的多样化特征。

3.1.4 数据特征分类层

在上一步数据特征提取层处理完成后，只提取出重要的故障特征，无法达到故障诊断的目的，另外需要把不同特征进行分类。本文通过增加一个全连接层达到分类的目的，由于全连接层的每一个节点都与上一层的所有节点相连，相应的参数过

多，在全连接的基础上又引入了丢弃层，丢弃层对每个批次的训练数据进行随机选择，实质是训练了多个子神经网络，在不同的子网络中随机忽略不同位置的权重，最后再将这些小的子网络组合起来，以一种类似投票的机制来作为故障分类的结果，还避免了模型的过拟合问题。

3.2 基于故障知识图谱的故障诊断匹配

利用知识图谱分析匹配故障诊断SA-BiLSTM模型输出分类结果，生成故障诊断报告，进而提高故障诊断模型输出的可解释性。通过对智算设备多源数据进行知识整合、知识抽取、知识融合和表示，并加入故障知识融合反馈机制，将融合发现的问题反馈更新到知识建模的过程中，最终构建完成智算设备故障知识图谱（knowledge graph, KG）。智算设备故障知识图谱构建流程如图6所示。

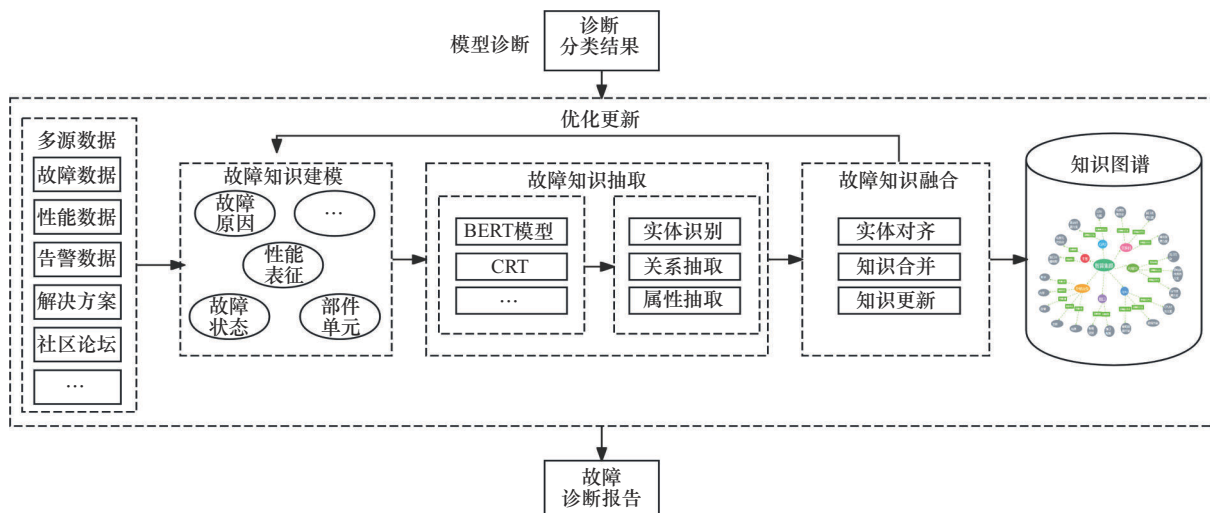


图6 智算设备故障知识图谱构建流程

3.2.1 知识图谱构建

(1) 知识建模。智算设备故障知识图谱属专业领域图谱，构建需要专业知识模型指导：首先，定义实体类型，见表2；其次，确定关系类型，见表3，因果关系是故障诊断的核心，用以关联故障原因与现象，如“光模块脏污致光功率下降或通信中断”；最后，定义实体属性，设备实体属性有运行时间、温度、数量等。

表2 实体类型		
实体类型名称	说明	示例
部件单元	智算设备的各种单元、部件、设备等	NPU、光模块、交换机
性能表征	部件的特征或者性能描述	温度、功率、带宽
故障状态	系统或部件的故障状态描述，多为故障类型	温度过高、脏污、闪断

(2) 知识收集。随着智算设备的发展，集群故障频发，故障数据庞大且多样。鉴于缺乏完整知识库，本文开展知识收集工作，通过多源实时

表3 关系类型

关系类型	客体	关系	主体示例	客体示例
部件单元	故障状态	部件故障	NPU	温度过高
性能表征	故障状态	性能故障	光模块	功率下降
部件单元	部件单元	组成	端口	交换机

采集日志与各类数据，整合历史资料，融合专家经验与案例，经数据预处理，按故障类型和严重程度标注，为构建知识图谱提供优质数据。数据来源示例见表4。

(3) 知识抽取。面对大规模复杂智算设备，单纯依靠人工或专家经验提取知识单元，既耗时又难以获取全面，需要借助基于规则或机器学习的方法。结构化故障数据能直接转化为“实体-关系-实体”三元组；而非结构化/半结构化数据，当前效果出色的是基于BERT-BiLSTM的抽取模型^[18]，它先通过BERT向量化数据，再用BiLSTM分析上下文，最后借助条件随机场（conditional random field, CRF）生成最优标签序列。知识抽取如图7所示。

表4 数据来源示例

数据来源	数据类型	数据内容
故障数据、性能数据告警数据等	结构化	常见故障、故障内容、故障原因、故障解决方法
解决方案、运维手册、专业书等专家知识文档	非结构化	故障类型、故障描述、故障原因、故障解决方法
社区、论坛、百科知识等	半结构化	故障名称、故障现象、故障原因、故障解决方案

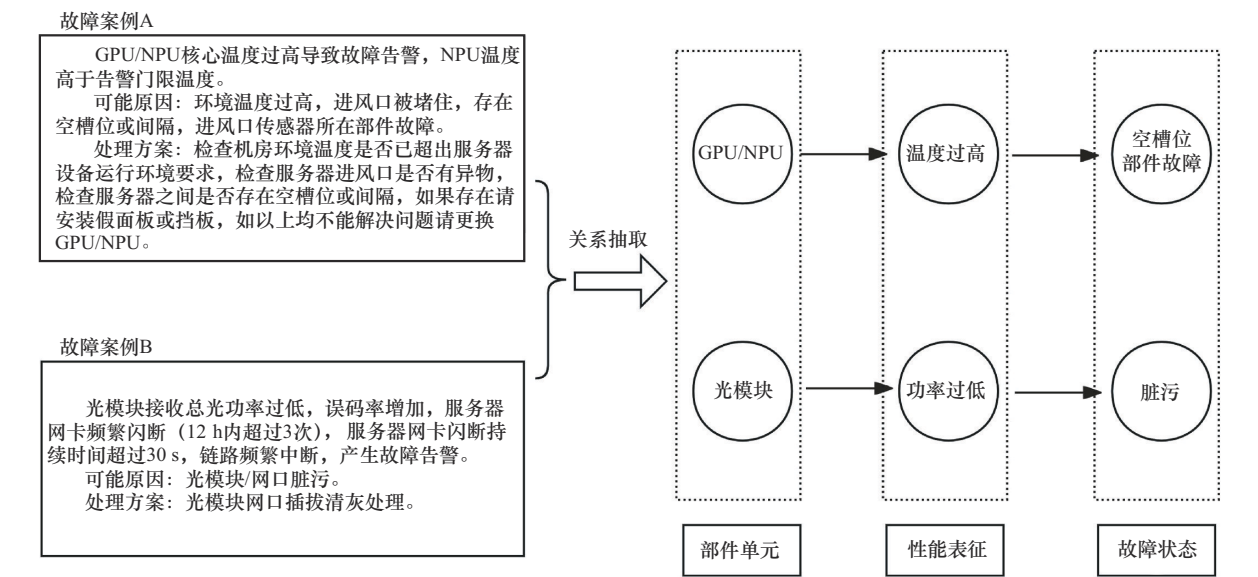


图7 知识抽取



(4) 知识融合。提升智算设备故障知识库的数据质量,需要融合不同数据源知识,其关键在于对不同来源的实体、关系及属性进行对齐,解决实体多表达问题,实现消歧与共指消解。本文采用余弦相似度算法,在抽取实体后,构建特定词典并将其词向量化,计算候选实体相似值,若相似度达预设阈值,便融合实体以完成对齐,最终提升知识抽取模型的准确性与一致性。

(5) 知识存储。对智算设备故障知识进行处理后,可转化为利于计算的数据结构。本文将其分类,以“实体-关系-实体/属性”三元组存储,因知识存储质量影响查询效率,结合智算设备故障知识结构特点,选用Neo4j图数据库,构建知识体系,为故障诊断提供知识匹配依据。

3.2.2 故障诊断结果解释

在故障诊断领域,故障知识图谱可对故障诊断模型分类结果进行深度解释,生成详尽诊断报告,其构建了包含故障模式、原因、现象及处理措施的结构化知识网络。当模型给出分类结果,知识图谱通过关联规则匹配与语义推理,将结果与图谱知识节点关联。如模型判定某故障属特定类型,图谱会依关系网络追溯潜在原因,找出故障现象与历史案例,将抽象结果具象化,展现故障内在逻辑。生成报告时,知识图谱提取匹配节点信息并按逻辑组织。报告含故障类型确认、原因分析、现象描述与处理建议等板块。经图谱解释,报告兼具准确性、可解释性与指导性,能为技术人员提供全面易懂的诊断信息,助力快速定位和解决故障。

4 超万卡集群实践过程与分析

为了验证所提故障诊断方案的准确性和可行性,本文首先采用了仿真实验的方式进行验证,同时对该方案在中国移动大规模智算中心实践的情况展开分析。仿真实验选择光模块故障作为典型故障场景,基于某万卡超大规模智算集群采集的真实数据,构建仿真数据集,进行完整的仿真

实验并开展对比分析。实践基于中国移动目前规模最大的1.8万卡单体集群生产环境,通过深度挖掘真实的生产数据,对故障诊断方案的有效性进行了详尽分析与验证。

4.1 故障诊断仿真实验

4.1.1 仿真数据集

光模块是智算集群设备的核心部件之一,万卡集群由近10万个光模块组成,单模块失效则集群训练中断。据统计,光模块年失效率约为4%,其一旦失效就会导致上下游光链路中断,所以光链路故障已成为影响集群训练的关键因素。光模块输出功率是光模块运行过程中的关键参数之一,因此,为了保证网络的正常运行,光模块的内置寄存器中通常会设置输出功率的告警门限,当光模块检测到输出功率超过预设门限即触发告警,若其发生故障,将对智算训练任务产生严重后果。

基于所提统一采集方案,本文收集了中国移动智算中心某万卡集群光模块近千亿条真实数据,经过对数据进行筛选、清洗、预处理以及标准化,生成标准训练数据集,然后将数据集输入本文提出的SA-BiLSTM模型进行训练,实现对光模块故障的类型识别。

光模块的8种典型光模块故障类别如图8所示,包含正常数据和7种异常数据,其中异常数据分为异常震荡、陡升、陡降、冲击上升、冲击下降、漂移上升、漂移下降等光模块故障现象。数据集包含13 200个样本,样本分布情况见表5。

表5 样本分布情况

故障现象	数量/个
正常	4 800
异常震荡	1 200
陡升	1 200
陡降	1 200
漂移下降	1 200
冲击上升	1 200
冲击下降	1 200
漂移上升	1 200

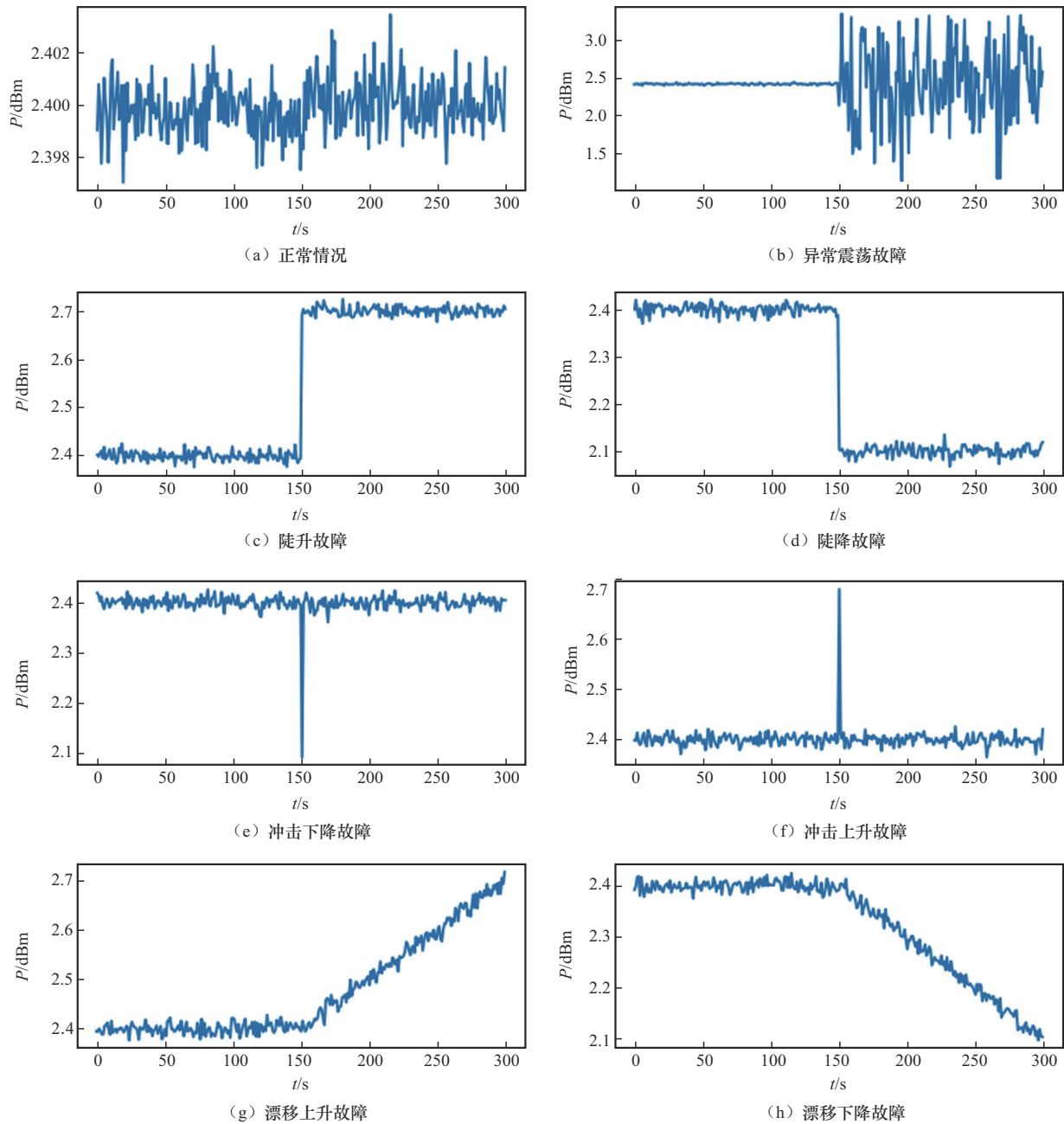


图 8 光模块的 8 种典型光模块故障类别

4.1.2 SA-BiLSTM 仿真实验

实验将上述包含 13 200 个样本的数据集按照 8:2 划分为训练集和测试集。数据划分采取随机采样的方式切割各个集合, 并采用分层抽样处理数据集, 确保训练集和测试集中各类别样本的比例与原始数据集保持一致, 而且每种分类训练样本量相同。实验使用 SA-BiLSTM 模型, 采用

Adam 优化器, 学习率设置为 0.01; 模型参数全连接层神经元个数调整为 32, 每轮的批处理样本数为 20, 共计训练 50 轮。

模型评价标准采用指标如下。

(1) 准确率: 在诊断结果中, 正确诊断的数量/样本总数。

(2) 损失: 损失函数在最后一轮时的大小。



(3) 收敛速度：训练过程中准确率提高的速度。

SA-BiLSTM 故障诊断模型不同权重下损失值变化曲线、准确率变化曲线分别如图9、图10所示，按照输出光功率、接收光功率、温度、误码率、链路时延等特征的不同重要程度，设置了3组初始权重系数进行对比实验。对比实验中训练集和测试集的结果曲线都具有趋同的特征，有效表明了自注意力机制的作用及其可学习的特点。总体来看，测试集故障识别准确率均与训练集相近，无过拟合与欠拟合现象，曲线波动较小并无较大的差异，模型训练效果较为理想，其中损失值逐渐收敛并稳定在0.047左右，同时准确率可达到98.4%。

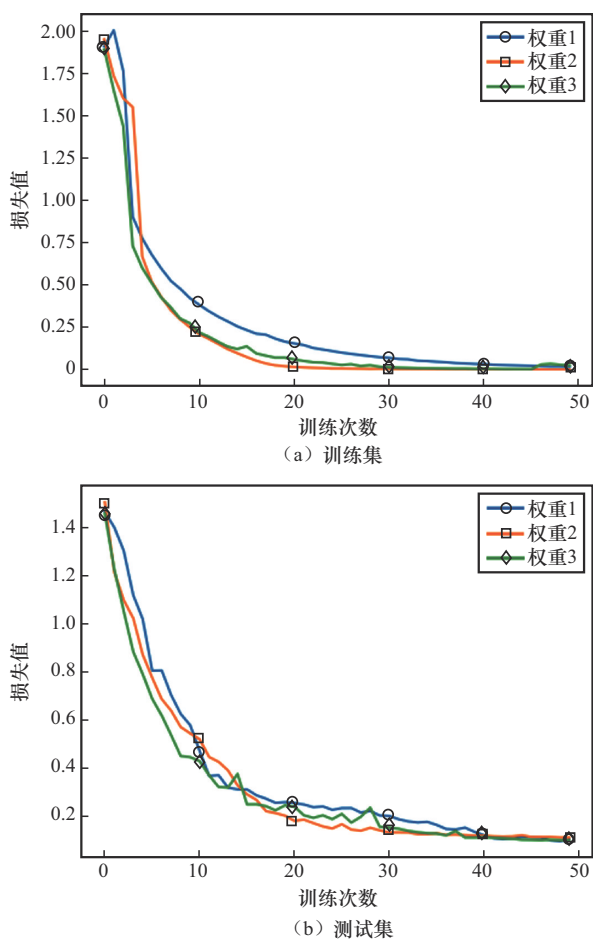


图9 不同权重下损失值变化曲线

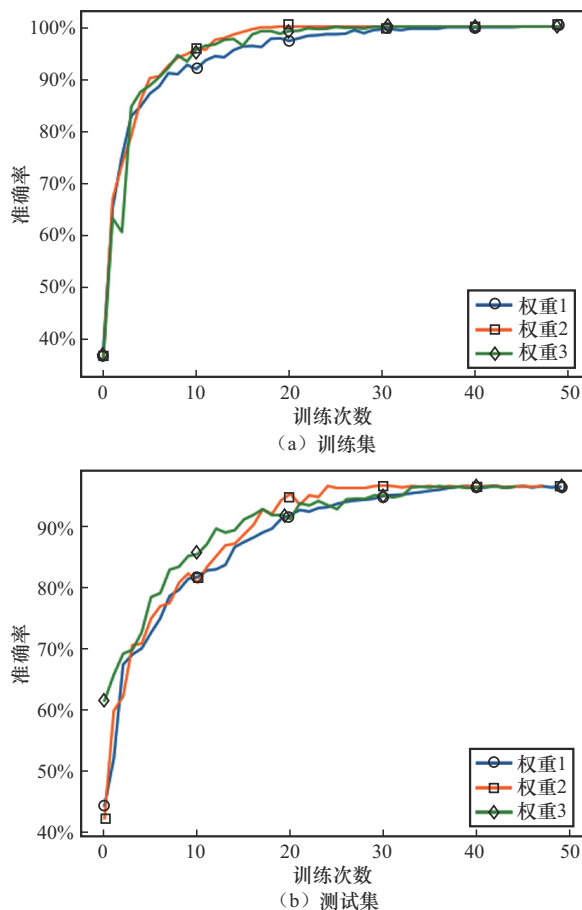


图10 不同权重下准确率变化曲线

4.1.3 算法对比实验

本文为了验证 SA-BiLSTM 模型诊断效果进行了对比实验，分别将原始数据输入 SA-BiLSTM 模型、CNN-1D 模型、BiLSTM 模型和支持向量机 (support vector machine, SVM) 模型进行对比。其中，BiLSTM 模型需要对原始信号进行归一化处理，为控制单一变量，其训练超参数和全连接层神经元个数与 SA-BiLSTM 模型保持一致。同时，为了验证训练集和测试集不同划分比例情况下，不同模型故障的诊断准确率，设置了9组不同比例的训练集和测试集，输入模型中进行实验。不同训练集下准确率比较结果如图11所示。

对比实验表明 SA-BiLSTM 模型在故障诊断准确率以及稳定性方面相比于其他模型均有较明显

显的优势。在超万卡集群的 7 类故障场景下，SA-BiLSTM 模型在测试集上的故障识别率最高能够达到 98.4%，平均为 96.2%，相较于 CNN-1D、BiLSTM 和 SVM 模型有明显提升。同时，SA-BiLSTM 模型在不同数据集下均表现出较高的准确率，表明该模型具有较好的稳定性。其主要原因在于 SA-BiLSTM 使用多层双向长短期记忆网络强化特征学习，使得多分类下效果同样明显。数据在特征预筛选层通过自注意力机制进行注意力权重计算，过滤掉无特殊特征的数据，起到数据筛选的作用，从而提取重要特征，为后续的 BiLSTM 减少层级深度，有效避免模型的梯度消失或梯度爆炸问题；在特征提取层中加入了 BN 处理，同时在隐藏层中加入不同数量的 BN 层，提升获取数据中的高频特征能力，以达到提高准确率与加快训练速度的效果；第三层在全连接的基础上又引入了丢弃层，丢弃层对每个批次的训练数据进行随机选择；两层 BiLSTM 层后的正则化与丢弃层有效减轻过拟合问题，提高模型的泛化能力，保证了模型的稳定性。

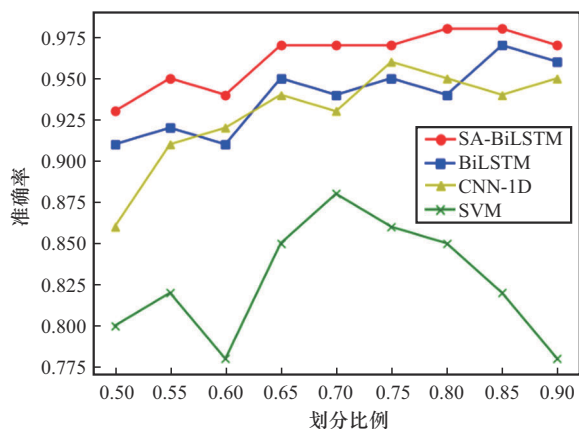


图 11 不同训练集下准确率比较结果

4.1.4 故障知识图谱应用

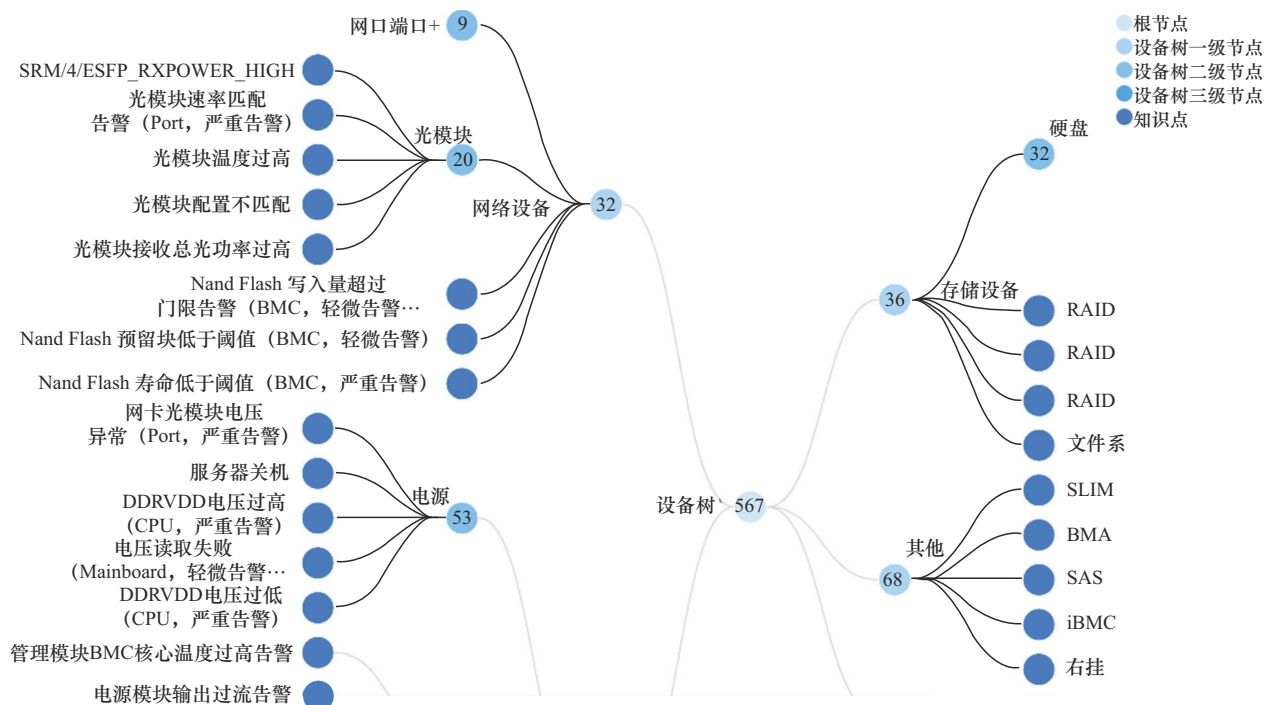
本文利用知识图谱分析匹配上述故障诊断 SA-BiLSTM 模型输出分类结果，构建了常见的几类智算设备故障知识图谱，覆盖计算设备、网络设备、存储设备。根据大模型训练场景下智算

设备特点，将设备实体划分为 NPU、光模块、RDMA 网卡、交换机、硬盘等几大类，并构建完成各个实体间的部件单元、性能表征、故障状态关系。通过组织结构和可视化，可以呈现故障现象、故障根因、解决方法、潜在关联关系等信息。以设备根节点网络设备为例详细说明故障诊断知识图谱，故障知识图谱局部如图 12 所示。图 12 (a) 展示了网络设备、光模块等实体关系，图 12 (b) 展示了“光模块接收总光功率过低”故障的关键词、问题现象、问题原因、处理步骤等内容。在实际的智算设备维护过程中，不仅需要智能化的故障诊断手段，还需要维护人员能够根据专业故障知识库和专家经验进行故障分析。然而，智算设备故障相关知识来源较多且类型多样，不便于维护人员查找。本文利用知识图谱技术可以将这些专业领域知识进行结构化展示和查询，并最终生成故障诊断报告，有效提高了故障诊断结论的可解释性。

4.2 超万卡集群故障诊断实践

中国移动 2024 年大规模上线了内蒙古、黑龙江、京津冀、长三角、粤港澳大湾区、成渝等地区 13 个智算中心节点，全网智能算力规模超 20EFLOPS，其中单集群最大算力达 6.9EFLOPS (1.8 万卡)。目前本文提出的故障诊断方案支持的主要算力类型包括英伟达、昇腾、天数和壁仞等，均在各个智算中心成熟应用。本文围绕中国移动目前规模最大的 1.8 万卡单体集群进行故障诊断实践分析。

在 1.8 万卡智算集群中，计算域、网络域和存储域的总机器数量超过 3 700 台，性能数据采集 Agent 能支持秒级的采集频率，采集指标数量超过 500 个，覆盖算存网的各种资源和设备，性能数据的实时吞吐量达到 80 Mbit/s，数据仓库中共计存储的性能数据量超过 90 TB，非结构化数据共计存储超过 100 TB (日志数据保留 30 天)。1.8 万卡数据采集规模见表 6。



(a) 实体关系

光模块接收总光功率过低

关键词:

1044、故障

文章内容:

问题现象描述

光模块接收总光功率过低。

问题影响

光模块故障

可能的原因

光模块接收总光功率过低。

处理步骤

1. 用户视图下display interface查看本端口当前接收光功率值 (Rx power total) ; 如果在接收范围内, 联系技术支持工程师处理, 转6;
2. 如果对端端口不是down且没有告警, 转4;
3. 修复对端端口down以及告警, 如果告警未恢复, 转6;
4. 确认光纤是否有问题, 可通过更换光纤尝试;
5. 查看本端接口信息, 是否有告警, 可通过更换光模块尝试修复;
6. 如果更换光模块后仍有告警, 请收集告警信息、日志信息和配置信息, 联系工程师进行处理。

(b) “光模块接收总光功率过低” 故障

图12 故障知识图谱局部

表6 1.8万卡数据采集规模

数据类别	包含范围	指标数量	采集频率	数据规模
性能数据	计算	200+	30 s	超3亿条/天
	网络	150+	30 s	超5亿条/天
	存储	100+	30 s	超1亿条/天
资产数据	计算、存储、网络	100+	1 h	超35万条
日志数据	软硬件、应用日志	—	实时	TB级/天

中国移动1.8万卡集群目前已稳定运行半年以上，已积累超大规模的存量数据，覆盖各种软硬件故障场景。其中，光模块、NPU/GPU故障为智算集群发生频率最高的故障场景，且在生产环境中这两大部件故障对大模型训练任务影响较大。本文中提到的故障诊断方案在超万卡集群生产环境中进行了深度实践，在光模块设备分类下有光模块弱光、网口闪断、脏污、接收光功率异常等故障；在NPU设备分类下有HBM ECC异常、比特错误、多比特错误等故障。方案在这7个故障场景下使用本文所提诊断方法进行了长达3个月的大规模应用，通过故障诊断的结果进行正确率统计分析，这7类故障场景诊断准确率均高于85%，最高的能达到93.7%。通过对1.8万卡的智算集群稳定性保障的深度实践，某大模型在15天的长稳测试中集群的有效训练时长超过了96%，部分故障在5 min内完成诊断和定界定位，极大地提升了集群的可靠性。超万卡集群故障诊断实践结果见表7。

在中国移动1.8万卡集群实践中，对光模块故障诊断统计分析测算出本文所提方案在集群稳定性维护上的效率优势，说明本文的方案可以较明显地提升集群稳定性。超万卡集群光模块故障诊断效率对比结果见表8。通过表8可知，与技

术人员进入机房近端检测的传统故障处理方案相比，本文提供的方案只需要基于自动化的设备性能、日志采集数据即可主动进行算法诊断分析。传统方案中一般需要2名计算专业和2名网络专业的人员接入设备执行复杂的指令查询相关数据，并基于专业知识进行综合分析、故障定位判断，单个光链路故障平均需要1 h才能完成定位，且强依赖技术人员的专业能力。本文方案通过算法和知识图谱实现了诊断过程的标准化、自动化，整个诊断过程不需要额外投入专业技术人员，且定位时长缩短到5 min以内。通过本文提供的方案，系统实现了主动感知和诊断，部分故障场景（如光模块脏污、光模块弱光等）甚至可以在性能劣化阶段就会被发现，可在5 min内完成故障定位分析，并结合故障隔离技术（对故障节点进行驱离）将任务中断控制在5 min内。

表7 超万卡集群故障诊断实践结果

序号	故障场景	设备分类	诊断次数/次	准确率
1	光模块弱光	光模块	56	85.4%
2	网口闪断	光模块	329	93.7%
3	光模块脏污	光模块	256	91.3%
4	光模块接收光功率异常	光模块	37	90.6%
5	HBM ECC异常	NPU	238	88.3%
6	单比特错误	NPU	130	86.2%
7	多比特错误	NPU	204	92.2%

5 结束语

本文提出基于数据和知识驱动的保障超万卡智算集群稳定性的方案，有效解决了超万卡智算集群规模大、硬件故障多、任务训练故障率居高

表8 超万卡集群光模块故障诊断效率对比结果

方案	诊断方式	人力投入	定位时长	任务中断时长
传统方案	技术人员携带专业工具进入机房，近端检测	熟练专业技术人员4名 (计算专业2名+网络专业2名)	1小时/单个光链路故障	>1 h
本文方案	基于采集数据，通过本文诊断算法实时自动化分析	0人	<5 min	<5 min



不下、跨域问题定位困难等问题。该方案通过异构资源一体化采集技术和分布式实时大数据ETL技术实现了集群性能数据的全面采集；利用改进BiLSTM深度学习模型进行故障诊断，使用多层BiLSTM强化特征学习，使得多分类下效果同样明显；通过引入自注意力机制，模型在提取时序性特征的同时能够引入特征权重系数，从而对不同尺度提取的特征加权融合，提高模型故障诊断精度；同时引入丢弃层有效减轻过拟合问题，提高了模型的泛化能力和稳定性；最后借助知识图谱完成故障诊断报告的输出，使得诊断结果更加直观和易于理解。该方案在兼顾诊断正确率与故障的可解释性方面效果良好，并在中国移动1.8万卡大规模智算集群实践中，7类故障场景诊断准确率平均高达90.6%。通过与传统故障诊断方式进行统计对比分析，本文提出的故障诊断方案可明显缩短故障定位时长，减少训练任务中断时间，有效地提升了集群稳定性，优化了运维流程，提升了运维效率。

在未来的研究中，模型可通过广泛引入更为丰富的领域专业知识以及海量多元的数据样本，聚焦超万卡集群，结合多种诊断手段，扩大故障诊断覆盖范围，解决万卡集群训练任务故障定位定界困难的问题，保障集群稳定高效训练。本文的工作为智算集群在日益复杂多变场景下的故障诊断提供了实践经验，为智能运维领域开拓和发展提供了助力。

参考文献:

- [1] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models[J]. arXiv preprint, 2020, arXiv: 2001.08361.
- [2] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint, 2018, arXiv: 1810.04805.
- [3] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[J]. arXiv preprint, 2020, arXiv: 2005.14165.
- [4] GRATTAFIORI A, DUBEY A, JAUKHRI A, et al. The Llama 3 herd of models[J]. arXiv preprint, 2024, arXiv: 2407.21783.
- [5] KOKOLIS A, KUCHNIK M, HOFFMAN J, et al. Revisiting reliability in large-scale machine learning research clusters[C]// Proceedings of the 2025 IEEE International Symposium on High Performance Computer Architecture (HPCA). Piscataway: IEEE Press, 2025: 1259-1274.
- [6] WANG Z, JIA Z, ZHENG S, et al. GEMINI: fast failure recovery in distributed training with in-memory checkpoints[C]// Proceedings of the 29th Symposium on Operating Systems Principles. New York: ACM Press, 2023: 364-381.
- [7] JIANG Z, LIN H, ZHONG Y, et al. MegaScale: scaling large language model training to more than 10 000 GPUs[J]. arXiv preprint, 2024, arXiv: 2402.15627.
- [8] HU Q, YE Z, WANG Z, et al. Characterization of large language model development in the datacenter[J]. arXiv preprint, 2024, arXiv: 2403.07648.
- [9] CUI W H, ZHANG J, ZHAO H, et al. XPUTimer: anomaly diagnostics for divergent LLM training in GPU clusters of thousand-plus scale[J]. arXiv preprint, 2025, arXiv: 2502.05413.
- [10] JIANG Z H, HUANG J J, CHEN Z B, et al. L4: diagnosing large-scale LLM training failures via automated log analysis[J]. arXiv preprint, 2025, arXiv: 2503.20263.
- [11] WU T Y, WANG W, YU Y H, et al. FALCON: pinpointing and mitigating stragglers for large-scale hybrid-parallel training[J]. arXiv preprint, 2024, arXiv: 2410.12588.
- [12] SCHUSTER M, PALIWAL K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673-2681.
- [13] PAN H, HE X, TANG S M F, et al. An improved bearing fault diagnosis method using one-dimensional CNN and LSTM[J]. Journal of Mechanical Engineering, 2018, 64(7-8): 443-452.
- [14] 邓志光, 吴茜, 朱加良, 等. 基于改进LSTM的核电厂传感器故障诊断研究[J]. 自动化仪表, 2023, 44(6): 115-120.
DENG Z G, WU Q, ZHU J L, et al. Research on sensor fault diagnosis in nuclear power plants based on improved LSTM[J]. Process Automation Instrumentation, 2023, 44(6): 115-120.
- [15] 王太勇, 王廷虎, 王鹏, 等. 基于注意力机制BiLSTM的设备智能故障诊断方法[J]. 天津大学学报(自然科学与工程技术版), 2020, 53(6): 601-608.
WANG T Y, WANG T H, WANG P, et al. An intelligent fault diagnosis method based on attention-based bidirectional LSTM network[J]. Journal of Tianjin University (Science and Technology), 2020, 53(6): 601-608.
- [16] 茅健, 郭玉荣, 赵嫚. 基于注意力机制的滚动轴承故障诊断方法[J]. 计算机集成制造系统, 2023, 29(7): 2233-2244.
MAO J, GUO Y R, ZHAO M. Fault diagnosis method of rolling bearing based on attention mechanism[J]. Computer Integrated Manufacturing Systems, 2023, 29(7): 2233-2244.

- [17] WANG X, HE X N, CAO Y X, et al. KGAT: knowledge graph attention network for recommendation[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2019: 950-958.
- [18] 潘庆亚, 张衡, 刘文杰, 等. 基于数据和知识的5G网络故障诊断方法[J]. 电信科学, 2023, 39(12): 53-64.
- PAN Q Y, ZHANG H, LIU W J, et al. Fault diagnosis method for 5G networks based on data and knowledge[J]. Telecommunications Science, 2023, 39(12): 53-64.

[作者简介]



牛红韦华 (1989-), 女, 中移(苏州)软件技术有限公司计划部副总经理, 主要研究方向为云计算、人工智能等。



黄永宝 (1990-), 男, 现就职于中移(苏州)软件技术有限公司, 主要研究方向为云计算、人工智能等。



丁国强 (1986-), 男, 现就职于中国移动通信集团设计院有限公司, 主要研究方向为云计算和人工智能技术等。

黄宝 (1992-), 男, 现就职于中移(苏州)软件技术有限公司, 主要研究方向为云计算、人工智能等。

赵治稳 (1990-), 男, 现就职于中移(苏州)软件技术有限公司, 主要研究方向为云计算、人工智能等。

徐杨 (1991-), 男, 现就职于中移(苏州)软件技术有限公司, 主要研究方向为人工智能、故障诊断等。

王涛 (1991-), 男, 现就职于中移(苏州)软件技术有限公司, 主要研究方向为无线安全、智算中心等。

张锐龄 (1988-), 男, 现就职于中移(苏州)软件技术有限公司, 主要研究方向为云计算、人工智能等。

王旋 (1995-), 男, 现就职于中移(苏州)软件技术有限公司, 主要研究方向为云计算、人工智能等。

张逸翔 (1996-), 男, 现就职于中移(苏州)软件技术有限公司, 主要研究方向为网络安全、算力网络等。