

FastCheck: 一种基于并行传输与定制化压缩的神经网络训练检查点快速保存与恢复方法

滕云¹, 孙大为¹, 胡世鹏², 李之悦², 张广艳², 田海东³, 常锐³

¹中国地质大学(北京)人工智能学院, 中国北京市, 100083

²清华大学计算机科学与技术系, 中国北京市, 100084

³中兴通讯股份有限公司移动网络和移动多媒体技术国家重点实验室, 中国深圳市, 518057

摘要: 大规模深度神经网络训练常面临软硬件故障问题, 出现关键故障时往往需要整个机器重启, 极大延长了训练时间。现有检查点保存与恢复方案有的无法应对此类关键故障, 有的受限于输入/输出带宽导致检查点保存与恢复速度缓慢。提出 FastCheck 框架, 通过并行传输与定制化压缩技术加速检查点保存与恢复过程。首先, FastCheck 将检查点数据分片, 利用多节点并行执行检查点保存与恢复过程; 其次, 通过权重增量压缩与动量索引压缩进一步减小检查点规模与开销; 最后, 采用轻量级共识协议精准追踪节点健康状态, 避免向故障节点传输检查点数据。在 PyTorch 中实现了 FastCheck, 并在多种深度神经网络模型上与两种基线方案进行了对比评估。实验结果表明, FastCheck 将检查点保存时间最高降低了 78.42%, 恢复时间最高减少了 77.41%, 且在不同训练阶段均能持续提升系统效率。

关键词: 深度神经网络模型; 关键故障; 并行传输; 数据压缩; 检查点保存与恢复

<https://doi.org/10.1631/ENG.ITEE.2025.0034>