有限 GPU 显存下的大语言模型训练技术综述

唐宇，乔林波，尹路珈，梁鹏，沈奥，杨智琳，张立志，李东升
国防科技大学计算机学院并行与分布计算全国重点实验室，中国长沙市，410073

摘要：大模型凭借其在多领域应用中的卓越性能，已在计算机视觉、自然语言处理等领域获得广泛关注。然而，此类模型的训练面临图形处理器（GPU）显存容量的显著制约。本文系统梳理了有限GPU显存条件下大模型训练的优化技术体系。首先深入解析训练过程中GPU显存占用的三大核心要素——模型参数、模型状态和模型激活；继而从这三个维度对现有研究成果进行多角度评述；最后展望了该领域未来的发展方向，强调持续创新显存优化技术对推动大语言模型发展的重要性。本综述为研究人员理解大语言模型训练中的显存优化挑战与技术演进提供了系统参考。

关键词：训练技术；显存优化；模型参数；模型状态；模型激活
https://doi.org/10.1631/FITEE.2300710