最小化内存冗余的自动并行策略生成方法

时彦琦，梁鹏，郑浩，乔林波，李东升
国防科技大学并行与分布处理国家重点实验室，中国长沙市，410000

摘要：受内存和计算资源限制，大规模深度学习模型通常以分布式方式训练。现有策略生成方法很少以最小化内存占用作为目标。为此，提出一种新算法，能够生成以最小化内存冗余为目标的自动并行策略。提出一种冗余内存代价模型来计算给定并行策略中每个算子的内存开销。为确保生成最优的并行策略，将并行策略搜索问题形式化为整数线性规划问题，使用高效求解器寻找具有最小内存占用的算子内并行策略。所提方法在多维并行训练框架中实现；实验结果表明，与最新Megatron-LM方法相比，可节省高达67%的内存开销，而吞吐量相差不大。

关键词：深度学习；自动并行；最小化内存冗余
https://doi.org/10.1631/FITEE.2300684