MicroAdam 论文调研报告

1120232535 汪隽宁 数据科学与大数据技术

摘要

本文调研了由Ionut-Vlad Modoranu等人提出的**MicroAdam**优化器 (Modoranu et al., 2024)，该工作发表于2024年。MicroAdam通过**梯度稀疏化**和**可压缩误差反馈**机制，将Adam类优化器的内存开销从$O(2d)$降至$O(0.5d)$，同时保持理论收敛性。实验证明其在BERT、LLaMA等模型微调中，内存减少50%以上且精度无损。本报告解析其核心算法、理论贡献及实验结果。

1 引言

传统Adam优化器因存储一阶矩m_t和二阶矩v_t需$O(2d)$内存，成为大模型训练的瓶颈。现有解决方案如8-bit Adam (Dettmers et al., 2022)和GaLore (Zhao et al., 2024) 或缺乏理论保证，或依赖强假设。MicroAdam的创新在于:

- 首次证明**误差反馈可量化**而不破坏收敛性
- 动态滑动窗口实现**稀疏动量计算**
- 在十亿级参数模型上验证有效性

2 核心算法

2.1 关键步骤

MicroAdam的每步迭代流程如下:

Algorithm 1 MicroAdam with Quantized EF

Input: $\beta_1, \beta_2, \epsilon, \mathcal{G}, T, d, k$
Initialize $m_0, v_0 \leftarrow 0_d$, $e_1 \leftarrow 0_d^{4b}$
for $t = 1$ to T **do**
  $g_t \leftarrow \nabla_\theta f(\theta_t)$ // 1. 梯度计算
  $a_t \leftarrow g_t + Q^{-1}(e_t)$ // 累积误差
  $\mathcal{I}_t, \mathcal{V}_t \leftarrow \text{TopK}(|a_t|)$ // 2. Top-K稀疏化
  $a_t[\mathcal{I}_t^c] \leftarrow 0$ // 丢弃非Top-K元素
  $e_{t+1} \leftarrow Q(a_t)$ // 3. 误差4-bit量化
  $\mathcal{G} \leftarrow (\mathcal{I}_t, \mathcal{V}_t)$ // 更新滑动窗口
  $\hat{m}_t \leftarrow \text{AdamStats}(\beta_1, \mathcal{G})$ // 4. 动态动量计算
  $\hat{v}_t \leftarrow \text{AdamStats}(\beta_2, \mathcal{G}^2)$
  $\theta_{t+1} \leftarrow \theta_t - \eta_t \frac{\hat{m}_t}{\epsilon + \sqrt{\hat{v}_t}}$ // 5. 参数更新
end for

2.2 内存优化分析

Table 1: 内存开销对比（LLaMA-7B实例）

优化器	内存表达式	实际占用
Adam (bfloat16)	$O(2d)$	25.10 GB
8-bit Adam	$O(2d/4)$	12.55 GB
GaLore (r=256)	$O(6d_r + 2\epsilon_1)$	1.36 GB
MicroAdam	$O(0.5d + 4mk)$	**5.65 GB**

3 理论贡献

MicroAdam在以下条件下保持收敛：

1. **非凸光滑函数**：收敛速率$O(1/\sqrt{T})$，匹配AMSGrad

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\theta_t)\|^2] \leq 2C_0\left(\frac{f(\theta_1)-f^*}{\sqrt{T}} + \frac{L\sigma^2}{\epsilon\sqrt{T}}\right) + \mathcal{O}\left(\frac{G^3(G+d)}{T}\right)$$

2. **PL条件**：线性收敛速率$O(\log T/T)$

关键条件是压缩算子满足$(1+\omega)q < 1$（q为梯度压缩率，ω为误差量化界）。

4 实验结果

主要结论：

- 在相同超参下，MicroAdam达到与Adam相当的准确率（34.72% vs 34.50%）
- 相比8-bit Adam，内存减少55%（5.65GB vs 12.55GB）
- 在ResNet-50上表现出隐式正则化效果，验证精度提升1%以上

5 局限性与展望

- **局限性**：
  - 预训练任务中稀疏更新可能不足（需稠密更新注意力层）
  - 小模型上因PyTorch内存分配策略，实测开销可能高于理论值
- **未来方向**：
  - 扩展至低秩梯度压缩
  - 研究动态稀疏度调整策略

参考文献

Ionut-Vlad Modoranu, Mher Safaryan, Grigory Malinovsky, Eldar Kurtic, Thomas Robert, Peter Richtarik, and Dan Alistarh. 2024. MicroAdam: Accurate Adaptive Optimization with Low Space Overhead and Provable Convergence. *arXiv preprint* arXiv:2405.15593. https://arxiv.org/abs/2405.15593

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit Optimizers via Block-wise Quantization. *arXiv preprint* arXiv:2110.02861. https://arxiv.org/abs/2110.02861

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. 2024. GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection. *arXiv preprint* arXiv:2403.03507. https://arxiv.org/abs/2403.03507