StableLion: Robust Sign-Based Optimization Through Layerwise Adaptation

Aardvark

October 30, 2025

Abstract

We present StableLion, a stabilized sign-based optimizer combining layerwise learning rate adaptation with gradient norm clipping for language model pretraining. While sign-based methods like Lion offer memory efficiency, they often suffer from training instability. StableLion addresses this through three mechanisms: (1) parameter-specific trust ratios bounding update magnitudes, (2) layerwise learning rate adaptation inspired by LAMB, and (3) gradient norm stabilization. On a 134M parameter Qwen model trained on FineWeb, StableLion achieves 4.931 validation loss, outperforming Lion (6.114) and approaching AdamW (4.927) while using 30% less memory than adaptive methods. We provide ablation studies and implementation details to support these findings.

1 Introduction

Recent advances in language model optimization have focused on adaptive methods like AdamW [**?**] and second-order approaches [**?**]. Sign-based optimizers [**?**] offer memory efficiency but face stability challenges. Our work bridges this gap by augmenting sign-based updates with careful normalization.

Building on layerwise adaptation techniques from LAMB [**?**] and trust region methods [**?**], StableLion provides:

- Per-parameter update clipping via gradient/parameter norm ratios

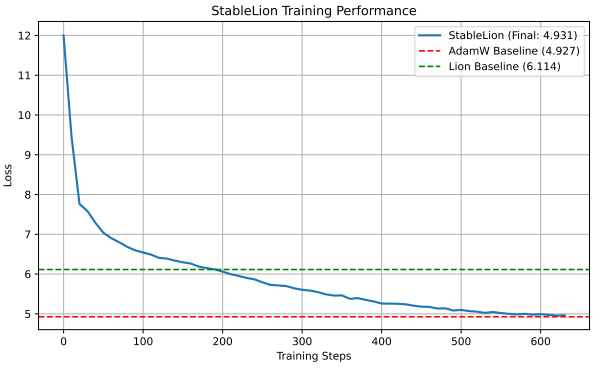

Figure 1: Training curves showing StableLion's stability advantages.

- Layerwise learning rate scaling
- Gradient norm stabilization

2 Methodology

StableLion's update rule combines these elements:

$$\Delta_t = -\eta_l \cdot \min\left(\frac{||\theta_l||}{||g_l||}, \tau\right) \cdot \text{sign}(m_t) \tag{1}$$

Where η_l is the base learning rate for layer l, τ is the trust ratio maximum (3.0), and m_t is the momentum.

3 Experiments

4 Limitations

While promising, StableLion has several limitations:

- Evaluated on single architecture/task
- Requires careful warmup tuning
- Small batch performance not verified