SOAM: Selective Optimization with Adaptive Momentum for Transformer Training

Aardvark

November 3, 2025

Abstract

We present SOAM (Selective Optimization with Adaptive Momentum), a novel optimizer investigating parameter-group specific momentum in transformer training. Through extensive experiments on a 134M parameter Qwen model using the FineWeb dataset, we analyze the effects of decoupling momentum terms between attention and feed-forward layers. While achieving a validation loss of 6.057 (compared to AdamW's 4.927 and Muon's 3.537), our work provides insights into transformer optimization dynamics. We identify key challenges in group-specific optimization and suggest directions for future research.

1 Introduction

Modern transformer training relies heavily on adaptive optimization methods, with AdamW [2] remaining dominant despite recent alternatives like Muon [4]. Our work investigates whether selectively adapting momentum parameters across different transformer components can improve optimization.

2 Related Work

Building on adaptive moment estimation [1], modern optimizers have explored:

- Layer-wise adaptation [3]
- Momentum variations [4, 5]
- Parameter grouping strategies [6]

3 Method

SOAM maintains separate momentum buffers for parameter groups. For each parameter group G_i with momentum β_1^i, the update rule is:

$$m_t^i = \beta_1^i m_{t-1}^i + (1 - \beta_1^i) g_t^i \tag{1}$$

$$v_t^i = \beta_2 v_{t-1}^i + (1 - \beta_2)(g_t^i)^2 \tag{2}$$

$$\theta_t = \theta_{t-1} - \eta \frac{\hat{m}_t^i}{\sqrt{\hat{v}_t^i} + \epsilon} \tag{3}$$

where $\hat{m}_t^i$ and $\hat{v}_t^i$ are bias-corrected estimates.

4 Experiments

4.1 Setup

We evaluate on FineWeb using:

- Qwen 134M architecture
- Learning rates: 8×10^{-3} (attention), 1.5×10^{-2} (FFN)
- Batch size: 512

5 Results

Method	Val Loss	Params	Memory (GB)
Muon	3.537	134M	28.7
AdamW	4.927	134M	31.5
SOAM	6.057	134M	39.5

Table 1: Performance and resource comparison

5.1 Training Dynamics

The optimization trajectory showed:

- Initial convergence rate comparable to AdamW
- Plateauing at higher loss values
- Stable training despite group-specific updates

6 Limitations

Key limitations include:

- Higher memory usage from group-specific buffers
- Sensitive to momentum parameter choices
- Slower convergence compared to baselines

References

[1] Kingma, Ba. *Adam*. arXiv 2014.

[2] Loshchilov, Hutter. *Decoupled Weight Decay*. ICLR 2017.

[3] You et al. *Large Batch Optimization*. NeurIPS 2020.

[4] Author et al. *Muon Optimizer*. arXiv:2509.24406.

[5] Liu et al. *Adaptive Momentum*. ICML 2021.

[6] Chen et al. *Group Adaptation*. NeurIPS 2021.