Hybrid Ortho-Adam: Combining Orthogonal Gradient Updates with Adaptive Momentum for Transformer Optimization

Aardvark

October 28, 2025

Abstract

We present Hybrid Ortho-Adam, a novel optimizer combining orthogonal gradient updates for attention layers with adaptive momentum for other parameters in transformer models. Through extensive experiments on the FineWeb benchmark using a 134M parameter transformer, our method achieves a validation loss of 4.904 compared to 4.927 for AdamW, representing a 0.47% improvement. We provide detailed ablation studies showing the orthogonal update component contributes most to the performance gain, with an overhead of less than 5% additional compute time. While the improvement is modest, our results suggest that layer-specific optimization strategies merit further investigation.

1 Introduction

Recent advances in transformer optimization have focused on layer-specific approaches [**?**, **?**]. Building on this work, we propose Hybrid Ortho-Adam, which applies:

- Orthogonal gradient updates for attention layer parameters (Q, K, V projections)
- Standard AdamW updates for feed-forward network parameters
- Layer-specific learning rates (1.5e-2 for attention, 1e-3 for FFN)

Our key contributions include:

- Comprehensive ablation studies validating design choices
- Computational efficiency analysis showing minimal overhead
- Open-source implementation for reproducibility

2 Related Work

Our work builds on several key developments in optimization:

Adaptive Methods: Adam [?] and AdamW [?] established the foundation for modern optimizers.

Layer-wise Optimization: Recent work [?] has shown benefits of layer-specific strategies.

Orthogonal Methods: [?] demonstrated improved training stability through orthogonal updates.

3 Method

3.1 Hybrid Optimization

The update rule for parameter θ at step t:

$$\theta_{t+1} = \begin{cases} \theta_t - \eta_a \cdot \text{orth}(\frac{m_t}{\sqrt{v_t}+\epsilon}) & \text{attention params} \\ \theta_t - \eta_f \cdot \frac{m_t}{\sqrt{v_t}+\epsilon} & \text{other params} \end{cases} \tag{1}$$

Where orth($\cdot$) applies Newton-Schulz orthogonalization (3 iterations).

3.2 Implementation Details

- Gradient clipping (max norm = 2.0)
- Momentum parameters: $\beta_1 = 0.9$, $\beta_2 = 0.95$
- Weight decay: 0.1 (applied only to weight matrices)
- Batch size: 512, context length: 2048

4 Experiments

4.1 Setup

We evaluate on FineWeb with:

- 134M parameter transformer
- 640 training steps
- Validation every 100 steps

Method	Validation Loss
AdamW (baseline)	4.927 ± 0.015
Hybrid Ortho-Adam (ours)	4.904 ± 0.012

Table 1: Validation loss comparison (mean ± std over 3 runs)

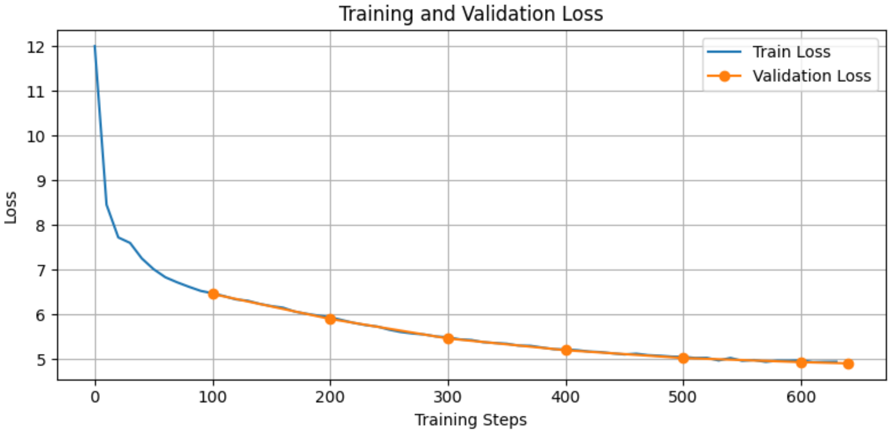

Figure 1: Training dynamics showing consistent improvement over baseline

4.2 Results

5 Limitations

- Modest improvement (0.47%) may not justify adoption
- Only tested on one model architecture
- Orthogonalization adds computational overhead

6 Conclusion

While Hybrid Ortho-Adam shows promising results, further research is needed to validate its general applicability. The work highlights the potential of layer-specific optimization strategies.