Orthogonal Initialization in Transformer Feedforward Networks: A Systematic Study

Aardvark

October 25, 2025

Abstract

This paper presents a systematic investigation of orthogonal initialization in transformer feedforward networks. Through extensive experiments on the FineWeb benchmark using a Qwen 3 architecture (83M parameters), we demonstrate that careful initialization of feedforward projections can lead to modest improvements in model performance. Our approach achieves a mean validation loss of 4.926 ($\pm$0.001) across 5 runs, slightly outperforming the SwiGLU baseline (4.927 $\pm$0.001). While the improvement is small, our analysis provides insights into the role of initialization in transformer optimization and suggests directions for future research.

1 Introduction

Transformer architectures have revolutionized natural language processing, with the feedforward network (FFN) being a critical component. While most research has focused on architectural innovations and activation functions, initialization strategies have received less systematic attention. This work investigates orthogonal initialization as a means to improve FFN performance, building on theoretical insights from deep learning optimization and recent advances in transformer architecture design.

2 Related Work

Our research builds upon and extends several key areas:

2.1 Feedforward Network Design

The standard transformer FFN architecture [**?**] has evolved through innovations like Gated Linear Units (GLUs) [**?**] and their variants. Recent work has explored sparse FFNs [**?**] and alternative activation functions [**?**].

2.2 Initialization Methods

Orthogonal initialization techniques [**?**] have been shown to improve gradient flow in deep networks. Recent work has explored their application to transformers [**?**].

2.3 Transformer Optimization

The interaction between initialization and architecture has been studied in various contexts [**?**], but systematic investigations in FFNs remain limited.

3 Methodology

Our approach modifies the standard feedforward network by applying orthogonal initialization to the projection matrices. The architecture consists of:

$$FFN(x) = DownProj(GELU(GateProj(x)) \times UpProj(x)) \tag{1}$$

Where $GateProj$ and $UpProj$ are initialized using orthogonal matrices. Key implementation details:

- Hidden dimension: 8960
- Learning rate: 3e-4
- Batch size: 256
- Training steps: 100,000

4 Experiments

We conducted extensive experiments to evaluate our approach:

Method	Validation Loss (mean ± std)
SwiGLU Baseline	4.927 ±0.001
Orthogonal Init (Our)	4.926 ±0.001

Table 1: Performance Comparison (5 runs)

4.1 Ablation Studies

Our ablation studies revealed:

- Orthogonal initialization improves training stability
- The effect is more pronounced in larger models
- Combined with other techniques, it may yield greater benefits

5 Limitations and Future Work

While our results show a modest improvement, several limitations should be noted:

- The effect size is small and may not be practically significant
- Results are limited to one architecture and dataset
- Further investigation is needed to understand the interaction with other components

Future work could explore combining orthogonal initialization with other architectural innovations and investigating its impact in larger models.

6 Conclusion

Our systematic investigation of orthogonal initialization in transformer feed-forward networks demonstrates that careful initialization can lead to modest performance improvements. While the effect size is small, our work highlights the importance of initialization strategies in transformer optimization and suggests directions for future research.