Rotation-Based Feedforward Networks: A Geometric Approach to Transformer Layers

Aardvark

October 28, 2025

Abstract

We present Rotation-Based Feedforward Networks (RBFN), a novel architecture that replaces traditional feedforward layers with learned 4D rotational transformations. Drawing inspiration from geometric deep learning, RBFN parameterizes hidden space transformations as compositions of rotations rather than pointwise nonlinearities. On the FineWeb benchmark with an 83M parameter model, RBFN achieves a validation loss of 4.916, representing a 0.011 improvement over the SwiGLU baseline while maintaining comparable computational requirements. Detailed analysis reveals that the rotational formulation provides particular benefits in later training stages, suggesting advantages for modeling hierarchical linguistic structures. We provide both theoretical analysis of the rotation mechanism's properties and empirical validation of its effectiveness compared to existing feedforward variants.

1 Introduction

Transformer architectures have become foundational in modern machine learning, yet their feedforward components remain relatively understudied compared to attention mechanisms. While accounting for approximately two-thirds of parameters and computations in large language models, feedforward layers typically employ simple pointwise transformations that may not fully exploit the geometric structure of learned representations.

Recent work has introduced various gating mechanisms and architectural variants, but these modifications largely preserve the fundamental feedforward computation paradigm. We propose a fundamentally different approach inspired by rotational transformations from geometric deep learning. Our Rotation-Based Feedforward Network (RBFN) replaces traditional feedforward operations with learned 4D rotations in the hidden space, offering several theoretical advantages:

- Norm preservation: Rotation matrices maintain vector norms
- Hierarchical composition: Rotations naturally compose

- Rich interactions: 4D rotations enable sophisticated mixing

2 Method

The RBFN transforms an input x through learned rotational transformations. The complete computation is:

$$RBFN(x) = W_{down}(R(x) \cdot V(x)) + b \tag{1}$$

Where:

- $V(x) = W_{val}x$ projects the input
- $R(x)$ computes rotation matrices
- $\cdot$ denotes rotation application
- W_{down} projects back

The rotation matrices are computed as:

$$R(x) = \prod_{i=1}^{k} R_i(x) \tag{2}$$

Where each $R_i(x)$ is constructed from learned parameters. The rotation application is defined as:

$$(R(x) \cdot V(x))_j = \sum_{k=0}^{3} R(x)_{j,k} V(x)_{j+k \bmod d} \tag{3}$$

We initialize parameters randomly and apply layer normalization.

3 Results

RBFN achieves a final validation loss of 4.916, compared to 4.927 for SwiGLU. Training dynamics reveal:

- Slower initial convergence
- Faster improvement after step 200
- More stable optimization

Memory usage is higher (47.4GB vs 42.3GB). The results compare favorably to recent variants.

4 Conclusions

We presented Rotation-Based Feedforward Networks, a novel approach applying geometric transformations to Transformer feedforward layers. While showing modest improvements, the method opens new research directions in geometric language modeling.