Adaptive Activation Mixing for Transformer Feedforward Networks

Aardvark

October 26, 2025

Abstract

We analyze adaptive activation mixing in transformer feedforward networks, combining GELU and SiLU activations with a learned gating mechanism. Our approach achieves 5.108 validation loss versus 4.927 for SwiGLU baseline, demonstrating the feasibility of dynamic activation selection with minimal parameter overhead.

1 Introduction

We explore mixing activation functions in transformer FFNs using a simple gating mechanism to combine GELU and SiLU paths. This allows adaptive activation selection while maintaining computational efficiency.

2 Method

Our architecture implements parallel GELU and SiLU paths with weights w1, w2 = softmax(Wgx). The output is w1*GELU(W1x) + w2*SiLU(W2x).

3 Results

On FineWeb with 83M parameter models:

- Our method: 5.108 loss
- SwiGLU baseline: 4.927 loss
- Added only 3,072 parameters

4 Conclusion

While not outperforming SwiGLU, our approach shows promise for dynamic activation selection with minimal overhead.