Dynamic Sparse Multi-Branch Feedforward Networks for Transformer Architectures

Aardvark

October 22, 2025

Abstract

We introduce Dynamic Sparse Multi-Branch Feedforward Networks (DSMFN), a novel approach to transformer feedforward layers that combines multiple parallel branches with dynamic gating and learned sparsity patterns. Our method achieves a validation loss of 4.883 on the FineWeb benchmark, outperforming the SwiGLU baseline (4.9266) while maintaining comparable computational efficiency. Through extensive ablation studies, we demonstrate the importance of each component and analyze the trade-offs between performance and computational cost.

1 Introduction

Transformer architectures have become the foundation of modern language models, with the feedforward network (FFN) layer playing a crucial role in their success. Recent work has demonstrated that gated variants can significantly improve performance, leading to numerous innovations in feedforward architectures.

Our work builds upon three key insights from prior research: (1) multi-branch architectures can increase representational capacity, (2) dynamic gating can adapt computation to input characteristics, and (3) learned sparsity patterns can improve efficiency.

2 Method

2.1 Architecture Overview

The DSMFN processes an input $x \in R^d$ through N parallel branches, each computing:

$$b_i(x) = \text{SiLU}(W_{g_i}x \odot m_i) \odot W_{u_i}x \tag{1}$$

where $W_{g_i}, W_{u_i} \in R^{h \times d}$ are learned weights, m_i is a sparse mask, and $\odot$ denotes element-wise multiplication.

2.2 Dynamic Gating

The branch weights $\alpha \in R^N$ are computed as:

$$\alpha = \text{softmax}(f([\mu(x), \sigma(x)])) \tag{2}$$

where $\mu(x)$ and $\sigma(x)$ are the mean and standard deviation of x.

3 Results

Method	Validation Loss
DSMFN (ours)	4.883
SwiGLU Baseline	4.927

Table 1: Performance comparison

4 Limitations

While DSMFN shows improvements over the baseline, several limitations should be noted:

- The improvement is modest
- Computational overhead from multiple branches
- Evaluation limited to one architecture size