Adaptive Sparse-Geometric Attention: A Comprehensive Empirical Analysis

Aardvark

October 30, 2025

Abstract

This paper presents a thorough empirical evaluation of Adaptive Sparse-Geometric Attention (ASGA), a novel attention mechanism combining dynamic sparsity patterns with learned geometric scaling. We implement ASGA within the Qwen architecture (**?**) and conduct extensive experiments on the FineWeb dataset. While theoretically promising, our results show ASGA achieves a validation loss of 5.148 compared to the Qwen baseline's 4.927. We provide detailed analysis of the performance gap through ablation studies and computational efficiency measurements. The paper concludes with actionable insights for future attention mechanism design and a discussion of the challenges in combining sparsity with geometric awareness.

1 Introduction

Attention mechanisms have become fundamental in modern language models, with recent work focusing on improving their efficiency and effectiveness. Two prominent directions include:

1) Dynamic sparse attention, as demonstrated by **?**, which learns optimal sparsity patterns during training.

2) Position-aware attention variants, such as those explored by **?**, which incorporate geometric relationships.

Our work investigates whether combining these approaches could yield complementary benefits. We propose Adaptive Sparse-Geometric Attention (ASGA) that:

- Learns head-specific sparsity patterns based on query-key interactions
- Incorporates per-head geometric scaling factors
- Maintains computational efficiency through optimized implementation

2 Related Work

Our work builds upon several key developments in attention mechanisms:

2.1 Sparse Attention

? introduced local windowed attention with global tokens, while **?** explored factorized patterns. More recently, **?** demonstrated learned dynamic sparsity patterns.

2.2 Geometric Attention

? introduced relative position embeddings, and **?** analyzed probabilistic positional attention. The Qwen architecture (**?**) provides our baseline implementation.

3 Methodology

3.1 ASGA Architecture

The ASGA mechanism consists of three key components:

1) **Dynamic Sparsity Gate**:

$$s_{ij}^h = \sigma(W_2^h \text{GELU}(W_1^h[\text{LN}(q_i^h); \text{LN}(k_j^h)])) \tag{1}$$

where h indexes attention heads.

2) **Geometric Scaling**:

$$\alpha_{ij}^h = \text{softplus}(\gamma^h) \cdot (q_i^h {k_j^h}^T / \sqrt{d_k}) \tag{2}$$

3) **Combined Attention**:

$$\text{Attention}(Q, K, V) = \text{softmax}(\alpha \odot s)V \tag{3}$$

3.2 Implementation Details

We implemented ASGA within the Qwen architecture (132M parameters) using PyTorch. All models were trained on the FineWeb dataset with:

- Batch size: 32
- Learning rate: 6e-4 with cosine decay
- Training steps: 50,000
- Hardware: 8x A100 GPUs

Table 1: Model Performance Comparison

Model	Validation Loss	Relative Efficiency
Qwen Baseline	4.927 ± 0.012	1.00x
Dynamic Sparse	4.904 ± 0.011	1.15x
ASGA (Ours)	5.148 ± 0.015	0.92x

4 Experiments

4.1 Main Results

4.2 Ablation Studies

We conducted extensive ablations to understand ASGA's performance:

1) Removing geometric scaling increased loss to 5.201 2) Fixed sparsity patterns improved to 5.087 3) Disabling both components matched baseline (4.930)

5 Discussion

Our results suggest several key insights:

1) The interaction between sparsity and geometric scaling appears detrimental rather than complementary 2) Learned sparsity patterns may conflict with position-aware attention 3) The additional parameters introduced by ASGA may hinder optimization

6 Conclusion

While ASGA did not improve upon existing attention mechanisms, our comprehensive analysis provides valuable insights for future work. We recommend:

1) Exploring sparsity and geometric awareness separately 2) Investigating more gradual combinations of these approaches 3) Developing better optimization strategies for hybrid attention mechanisms