Efficient Sparse Attention Routing for Long-Context Inference

Your Name

November 13, 2026

Abstract

We present a routing-based sparse attention mechanism for transformer inference that reduces memory bandwidth without retraining the base model. Our method achieves a 3.1x speedup on 32k-token contexts with less than 1% accuracy degradation.

Keywords: sparse attention, long-context inference, transformers, efficient deep learning

1 Introduction

Serving transformer models over long contexts is dominated by attention cost. Prior sparsification approaches typically require retraining, which is impractical for large pretrained models. We instead propose a post-hoc routing scheme that plugs into existing checkpoints.

2 Related Work

Sparse Transformers [1] introduced fixed sparsity patterns, while Longformer [2] combined local and global attention. Reformer [3] used locality-sensitive hashing to approximate nearest-neighbor attention. Our approach differs by learning routing weights from activations at inference time, requiring no gradient updates.

3 Method

For each query q_i we estimate block relevance via a cheap similarity score

$$s_{ij} = \frac{\bar{q}_i \cdot \bar{k}_j}{\|\bar{q}_i\| \|\bar{k}_j\|},$$

where $\bar{q}_i$ and $\bar{k}_j$ are low-rank projections of the query and block-mean key. We retain the top-k blocks by s_{ij} and mask the rest before the softmax.

4 Experiments

We evaluate on three long-context benchmarks; results are averaged over five runs.

Context	Dense (ms)	Sparse (ms)	Accuracy Δ
8k	42	19	-0.2%
16k	118	44	-0.4%
32k	390	126	-0.7%

5 Conclusion

Our training-free routing method recovers most of the throughput benefits of sparse attention while preserving accuracy, and can be applied directly to existing checkpoints.

References

[1] Child, R. et al. Generating Long Sequences with Sparse Transformers. 2019.

[2] Beltagy, I. et al. Longformer: The Long-Document Transformer. 2020.

[3] Kitaev, N. et al. Reformer: The Efficient Transformer. 2020.