Sparse Attention for Long-Context Transformers

Your Name Co-Author Name

November 13, 2026

Abstract

Full self-attention scales quadratically with sequence length, limiting transformer models on long-context tasks. We propose a block-sparse attention mechanism that restricts each query to a learned subset of keys, reducing the computational cost from $O(n^2)$ to $O(n \log n)$ while retaining accuracy within 0.3 points of dense baselines on language modeling benchmarks.

1 Introduction

Transformer models rely on self-attention to relate every token to every other token in a sequence. As sequence length n grows, the $O(n^2)$ memory and compute cost of dense attention becomes the dominant bottleneck. Recent work on long-context applications — document summarization, retrieval-augmented generation, and genomic sequence modeling — has renewed interest in attention variants that scale sub-quadratically without sacrificing model quality.

2 Method

We partition the key sequence into fixed-size blocks and learn a routing function $r(\cdot)$ that selects the top-k blocks each query attends to. Given query q_i and key blocks $K_1, \dots, K_m$, the attention output is

$$\mathrm{Attn}(q_i) = \sum_{j \in r(q_i)} \mathrm{softmax}\left(\frac{q_i K_j^\top}{\sqrt{d}}\right) V_j .$$

The routing scores are computed from a lightweight projection $r(q_i) = \text{top-}k\left(q_i W_r K^\top\right)$, trained jointly with the rest of the network.

2.1 Complexity

With block size b and k selected blocks per query, the attention cost becomes $O(n \cdot k \cdot b)$ rather than $O(n^2)$. Setting $k \cdot b \propto \log n$ recovers the $O(n \log n)$ bound reported in the abstract.

3 Results

Table 1 compares our method against dense attention and a fixed local-window baseline on a held-out validation split.

Model	Perplexity	Tokens/sec
Dense attention	18.4	4,200
Local window	20.1	11,800
Sparse (ours)	18.7	10,600

4 Conclusion

Learned block-sparse routing closes most of the gap to dense attention while delivering a 2.5x throughput improvement, making it a practical default for long-context transformer training.

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.