



THE UNIVERSITY OF
CHICAGO

DEPARTMENT OF STATISTICS

Master's Thesis Presentation

Yinjie Wang

Department of Statistics
The University of Chicago

“Simple and Effective Query-Adaptive Sparse Attention”

July 23, 2026, at 3:00 PM
Jones 111, 5747 S. Ellis Avenue

Abstract

Sparse attention enables language models to efficiently process very long contexts by restricting computation and focus to the most relevant tokens.

In this paper, we study block sparse attention, which explicitly computes attention over only a sparse subset of contiguous blocks. A common approach is to assign each block a lightweight relevance score and select the highest-scoring blocks for each query. Our key insight is that both the scoring function and the number of selected blocks should adapt dynamically to the query itself.

We begin with the empirical observation that, within each layer and head, the residual variation in a query's attention scores is approximately constant along the key sequence.

Based on this observation, we propose a simple procedure that assigns each block a score based on contiguous sub averaging of the constituent scores followed by a query-dependent standardization.

These standardized scores make adaptive block selection straightforward: blocks can be selected by direct thresholding or by more refined rules guided by statistical insights.

We show empirically that this yields a highly effective method for query-adaptive sparse block selection, strongly outperforming methods that use a fixed budget of keys (top-k), a fixed budget of attention probability mass (top-p), and related dynamic sparse attention baselines.