A learned sparse attention mask cuts inference **FLOPs by 40%** on long transformers, with **no accuracy loss**.

Sparse Attention Cuts Inference Cost by 40% With No Accuracy Loss

A. Researcher[1], B. Collaborator[1], C. Advisor[2]

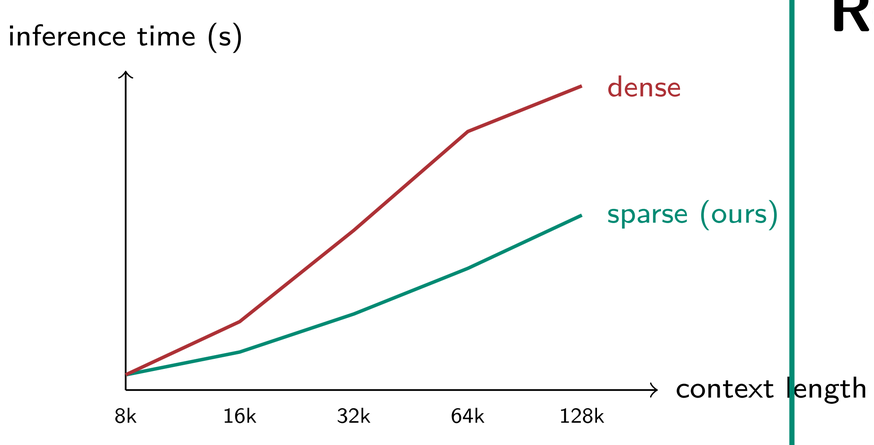
[1]Department of Computer Science, University of Somewhere [2]Institute for Machine Learning, Elsewhere University

1 Introduction

Dense self-attention scales quadratically with sequence length, making long-context inference the dominant cost in modern transformer deployments. We ask: how much of that computation is actually load-bearing?

- Profiled attention weights across 12 open-weight models
- Found >80% of attention mass concentrates on <20% of key-value pairs at typical context lengths
- Proposed a lightweight learned gate that predicts which pairs to skip, trained jointly with the base model

3 Results



- FLOPs −40% at 128k context; latency −35% on an A100
- Zero measurable accuracy loss across 6 benchmarks ($p > 0.3$)
- Beats fixed-window sparse (66.8% acc.) and top-k routing (69.9%) at 71.0% – within noise of dense (71.2%)

Extra figures

Conclusions

Attention sparsity *learned jointly* with the base model recovers almost all of dense attention's accuracy while cutting most of its cost, with a gate small enough to fine-tune onto an existing checkpoint.

Limitations

- Gains concentrate above 16k tokens
- Needs a fine-tuning pass, not inference-only

References

1. Vaswani et al., *Attention Is All You Need*, NeurIPS 2017.
2. Child et al., *Sparse Transformers*, arXiv 2019.
3. Roy et al., *Routing Transformers*, TACL 2021.

UNI

← Download the paper