Block-Sparse FlashAttention Memory Efficiency and Convergence Speed in XSum Fine-Tuning

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: Does the memory reduction from Block-Sparse FlashAttention enable larger batch sizes during XSum fine-tuning that improve convergence speed compared to standard FlashAttention-2. 15 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. Research question: Does the memory reduction from Block-Sparse FlashAttention enable larger batch sizes during XSum fine-tuning that improve convergence speed compared to standard FlashAttention-2?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.7/10.

3 Results

9 papers retrieved. 15 claims extracted; 2 independently verified. Quality review score: 4.7/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
FlashAttention outperforms the MLPerf 1.1 speed record for BERT by 15%.	×	0.08
FlashAttention speeds up GPT-2 up to 3$\times$ over HuggingFace and 1.8$\times$ over Megatron over standard Transformers.	×	0.06
FlashAttention speeds up the long-range arena (LRA) benchmark 2.4$\times$.	×	0.07
FlashAttention trains GPT-2 with context length 4K faster than Megatron trains GPT-2 with context length 1K, while achie	×	0.13
Modeling longer sequences yields 6.4 points of lift on two long-document classification tasks.	×	0.12
FlashAttention yields the first Transformer that can achieve better-than-random performance on the challenging Path-X ta	×	0.04
Block-sparse FlashAttention yields the first sequence model that we know of that can achieve better-than-random performa	×	0.08
The memory footprint of FlashAttention scales linearly with sequence length and is up to 3$\times$ faster than standard attenti	×	0.09
Runtime of block-sparse FlashAttention scales linearly in sequence length and is faster than all existing approximate at	✓	0.19
FlashAttention yields the fastest single-node BERT training speed that we know of.	×	0.04
FlashAttention is 15% faster than the Nvidia MLPerf 1.1 implementation for BERT-large training.	×	0.07
FlashAttention achieves up to 3$\times$ end-to-end speedup compared to Huggingface and 1.7$\times$ speedup compared to Megatron-LM for	×	0.10
FlashAttention achieves the same perplexity as HuggingFace and Megatron-LM implementations for GPT-2.	×	0.03
FlashAttention achieves up to 2.4$\times$ speed-up compared to standard attention on the Long-Range-Arena benchmarks.	×	0.12
Block-sparse FlashAttention is faster than all of the approximate attention methods tested on the Long-Range-Arena bench	✓	0.19

References

- http://arxiv.org/abs/2205.14135v2
- http://arxiv.org/abs/2512.07011v1
- http://arxiv.org/abs/2410.21676v4