LongNet and FlashAttention-2 Memory Scaling for Ultra-Long Sequence Training

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does the training memory consumption of LongNet scale relative to FlashAttention-2 when processing sequence lengths between 64k and 256k tokens on large-scale corpus pretraining. 13 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: LongNet: Scaling Transformers to 1,000,000,000 Tokens. Research question: How does the training memory consumption of LongNet scale relative to FlashAttention-2 when processing sequence lengths between 64k and 256k tokens on large-scale corpus pretraining?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.3/10.

3 Results

15 papers retrieved. 13 claims extracted; 2 independently verified. Quality review score: 6.3/10.

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
LONGNET can scale the sequence length to 1 billion tokens.	✓	0.24
The computation complexity of dilated attention is O(Nd) where N is the sequence length and d is the hidden dimension.	✓	0.17
LONGNET can be transformed into a dense Transformer, supporting off-the-shelf optimization for Transformers.	×	0.06
LONGNET can parallelize the training across nodes, breaking the constraint of both computation and memory with a distrib	×	0.05
LONGNET achieves nearly constant runtime when scaling to 1B tokens while vanilla Transformer suffers from quadratic comp	×	0.07
LONGNET is implemented on language modeling using the MAGNETO architecture with XPOS relative position encoding and base	×	0.03
The Stack dataset is used for pre-training the model.	×	0.04
The backbone architecture uses dilated attention to replace standard attention.	×	0.10
The models are trained with a batch size of 0.5M tokens for 300K steps.	×	0.03
All experiments are conducted based on the torchscale codebase.	×	0.03
LONGNET uses segment lengths of w = {2048, 4096, 8192, 16384, 32768}, and the dilated ratios are r = {1, 2, 4, 6, 12}.	×	0.03
We adjust their sparse ratios to match the computation flops with LONGNET so that the comparison is fair.	×	0.03
The attention layers in vanilla Transformers are dense and fully connected, so the computation cost is much higher.	×	0.03

References

- http://arxiv.org/abs/2512.18834v2
- http://arxiv.org/abs/1804.00857v1

- http://arxiv.org/abs/2307.02486v2