Novelty Assessment Report

Paper: Long-Context Generalization with Sparse Attention
PDF URL: https://openreview.net/pdf?id=PsB6Lynznk
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

Transformer-based architectures traditionally employ softmax to compute attention weights, which produces dense distributions over all tokens in a sequence. While effective in many settings, this density has been shown to be detrimental for tasks that demand precise focus on fixed-size patterns: as sequence length increases, non-informative tokens accumulate attention probability mass, leading to dispersion and representational collapse. We show in this paper that dynamically sparse attention mechanisms using α-entmax can avoid these issues, due to their ability to assign exact zeros to irrelevant tokens. Furthermore, we introduce Adaptive-Scalable Entmax (ASEntmax), which endows α-entmax with a learnable temperature parameter, allowing the attention distribution to interpolate between sparse (pattern-focused) and dense (softmax-like) regimes. Our empirical evaluation on synthetic tasks and language modeling demonstrates that ASEntmax substantially outperforms softmax, scalable softmax, and fixed-temperature α-entmax baselines, achieving up to 1000$\times$ length extrapolation on synthetic benchmarks and superior long-context generalization on language modeling while preserving short-context performance, including better perplexity trends and higher retrieval accuracies at 8$\times$ training length.

Core Task Landscape

This paper addresses: **Length Generalization in Transformer Attention Mechanisms**

A total of **50 papers** were analyzed and organized into a taxonomy with **22 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Positional Encoding and Embedding Schemes**
- **Attention Mechanism Modifications**
- **Hybrid Architectures**
- **Training Strategies and Data Augmentation**
- **Theoretical Foundations**
- **Empirical Studies**
- **Domain-Specific Applications**

Complete Taxonomy Tree

- Length Generalization in Transformer Attention Mechanisms Survey Taxonomy
- Positional Encoding and Embedding Schemes
  - Relative Position Encoding Methods (3 papers)
  - [11] Dissecting transformer length extrapolation via the lens of receptive field analysis (Ta-Chung Chi, 2023) View paper
  - [20] Train short, test long: Attention with linear biases enables input length extrapolation (Press, 2021) View paper
  - [27] Enhancing length generalization for attention based knowledge tracing models with linear biases (Feixiao Lv, 2024) View paper
  - Absolute and Learned Position Embeddings (3 papers)
  - [8] The impact of positional encoding on length generalization in transformers (Kazemnejad, 2023) View paper
  - [21] A formal framework for understanding length generalization in transformers (Huang Xinting, 2024) View paper
  - [42] Toward Length-Extrapolatable Transformers (Chi, 2024) View paper
  - No Position Encoding (1 papers)
  - [3] Length Generalization of Causal Transformers without Position Encoding (Jie Wang, 2024) View paper
  - Extrapolatable Position Embeddings (1 papers)
  - [47] TimelyGPT: Extrapolatable Transformer Pre-training for Long-term Time-Series Forecasting in Healthcare (Ziyang Song, 2024) View paper
- Attention Mechanism Modifications
  - Sparse and Selective Attention ★ (3 papers)
  - [0] Long-Context Generalization with Sparse Attention (Anon et al., 2026) View paper
  - [29] Big bird: Transformers for longer sequences (Zaheer, 2020) View paper
  - [41] Unshackling Context Length: An Efficient Selective Attention Approach through Query-Key Compression (Wang Haoyu, 2025) View paper
  - Dense Attention Variants (2 papers)
  - [12] On Vanishing Variance in Transformer Length Generalization (Li Ruining, 2025) View paper
  - [17] Scalable-Softmax Is Superior for Attention (Nakanishi, 2025) View paper
  - Multi-Query and Grouped-Query Attention (1 papers)
  - [1] Gqa: Training generalized multi-query transformer models from multi-head checkpoints (Ainslie, 2023) View paper
  - Memory-Augmented Attention (3 papers)

    - [16] Leave no context behind: Efficient infinite context transformers with infini-attention (Munkhdalai, 2024) View paper
    - [28] Beyond attention: Breaking the limits of transformer context length with recurrent memory (A. Bulatov, 2024) View paper
    - [40] Landmark attention: Random-access infinite context length for transformers (Mohtashami, 2023) View paper
    - Attention Bias and Calibration (2 papers)
    - [23] A Training-Free Length Extrapolation Approach for LLMs: Greedy Attention Logit Interpolation (Yan Li, 2025) View paper
    - [38] From interpolation to extrapolation: Complete length generalization for arithmetic transformers (Shaoxiong Duan, 2023) View paper
    - Attention Tensorization and Structural Modifications (2 papers)
    - [35] Long Sequence Modeling with Attention Tensorization: From Sequence to Tensor Learning (Aosong Feng, 2024) View paper
    - [46] BiLSTM-MLAM: A Multi-Scale Time Series Prediction Model for Sensor Data Based on Bi-LSTM and Local Attention Mechanisms (Yongxin Fan, 2024) View paper
- Hybrid Architectures
    - State Space Model Hybrids (3 papers)
    - [25] Samba: Simple Hybrid State Space Models for Efficient Unlimited Context Language Modeling (Ren Liliang, 2024) View paper
    - [31] Block-state transformers (Fathi, 2023) View paper
    - [33] Mega: Moving average equipped gated attention (Ma, 2022) View paper
    - Linear Attention Mechanisms (2 papers)
    - [10] Parallelizing linear transformers with the delta rule over sequence length (Yoon Kim, 2024) View paper
    - [30] Linear attention sequence parallelism (Qin Zhen, 2024) View paper
    - State Space Model Analysis (2 papers)
    - [6] The hidden attention of mamba models (Itamar Zimerman, 2025) View paper
    - [19] Generalization Error Analysis for Selective State-Space Models Through the Lens of Attention (Arya Honarpisheh, 2025) View paper
- Training Strategies and Data Augmentation
    - Task-Specific Training Techniques (2 papers)
    - [4] Improving length-generalization in transformers via task hinting (Awasthi, 2023) View paper
    - [39] Length generalization in arithmetic transformers (Jelassi, 2023) View paper
    - Training-Free Adaptation Methods (3 papers)
    - [14] Lm-infinite: Simple on-the-fly length generalization for large language models (Han, 2023) View paper
    - [36] LM-Infinite: Zero-Shot Extreme Length Generalization for Large Language Models (Chi Han, 2023) View paper
    - [37] Mambaextend: A training-free approach to improve long context extension of mamba (S Azizi, 2025) View paper
- Theoretical Foundations (3 papers)
    - [18] What algorithms can transformers learn? a study in length generalization (Zhou, 2023) View paper
    - [22] Quantitative Bounds for Length Generalization in Transformers (Izzo, 2025) View paper
    - [24] A theory for length generalization in learning to reason (Xiao, 2024) View paper
- Empirical Studies
    - Algorithmic and Reasoning Tasks (3 papers)
    - [7] Exploring length generalization in large language models (Anil, 2022) View paper
    - [32] Transformers can achieve length generalization but not robustly (Zhou Yongchao, 2024) View paper
    - [45] The neural data router: Adaptive control flow in transformers improves systematic generalization (CsordÃ¡s, 2021) View paper
    - Language Modeling Studies (1 papers)
    - [13] Attention Mechanisms in Transformers: A General Survey (Hosseinzadeh, 2025) View paper
- Domain-Specific Applications
    - Speech and Audio Processing (2 papers)
    - [2] Exploring Length Generalization for Transformer-Based Speech Enhancement (Zhang Qiquan, 2025) View paper
    - [5] An Exploration of Length Generalization in Transformer-Based Speech Enhancement (Qiquan Zhang, 2024) View paper
    - Text-to-Speech and Generation (1 papers)
    - [9] Robust and Unbounded Length Generalization in Autoregressive Transformer-Based Text-to-Speech (Eric Battenberg, 2024) View paper
    - Video Generation (1 papers)
    - [15] Phenaki: Variable Length Video Generation From Open Domain Textual Description (Villegas, 2022) View paper
    - Specialized Prediction Tasks (7 papers)
    - [26] Neural attention shaping with contextual embedding recalibration in language models (Dorian Osatov, 2024) View paper
    - [34] GET-Zero: Graph Embodiment Transformer for Zero-Shot Embodiment Generalization (Austin Patel, 2024) View paper
    - [43] Temporal Logical Attention Network for Log-Based Anomaly Detection in Distributed Systems (Yang Liu, 2024) View paper
    - [44] Quora Insincere Questions Classification Using Attention Based Model (Snigdha Chakraborty, 2022) View paper
    - [48] I-BERT: Inductive Generalization of Transformer to Arbitrary Context Lengths (Nam, 2020) View paper
    - [49] Attention-Based Encoder for Online Data Anomaly Classification in Multivariate Time Series (Qun Yang, 2025) View paper
    - [50] AttABseq: an attention-based deep learning prediction method for antigenâ□□antibody binding affinity changes based on protein sequences (Ruofan Jin, 2024) View paper

Narrative

Core task: length generalization in transformer attention mechanisms. The field addresses how transformers can maintain or improve performance when processing sequences longer than those seen during training. The taxonomy organizes research into seven main branches: Positional Encoding and Embedding Schemes explore how position information affects extrapolation (e.g., ALiBi[20]); Attention Mechanism Modifications redesign core attention operations through sparse patterns, compression, or alternative formulations; Hybrid Architectures blend transformers with recurrent or state-space models (e.g., Samba[25], Mega[33]); Training Strategies and Data Augmentation develop curriculum or augmentation methods; Theoretical Foundations provide formal analyses of generalization bounds and expressiveness; Empirical Studies systematically evaluate length extrapolation across tasks; and Domain-Specific Applications tailor solutions to speech, vision, or reasoning domains. These branches reflect complementary perspectives—some focus on architectural innovation, others on training regimes or theoretical guarantees—yet all converge on enabling transformers to handle longer contexts reliably.

Within Attention Mechanism Modifications, sparse and selective attention methods form a particularly active line of work, balancing computational efficiency with representational capacity. Big Bird[29] introduced structured sparsity patterns combining local, global, and

random attention, demonstrating that carefully designed sparse schemes can preserve model quality while reducing quadratic complexity. Query-Key Compression[41] takes a different approach by compressing attention matrices to manage memory and computation. Sparse Attention Generalization[0] sits within this cluster, emphasizing how sparsity patterns themselves can be designed or learned to improve length extrapolation rather than merely reduce cost. Compared to Big Bird[29], which fixes sparsity structure a priori, and Query-Key Compression[41], which focuses on compression mechanics, Sparse Attention Generalization[0] appears to investigate adaptive or principled sparse designs that explicitly target generalization to longer sequences, bridging efficiency concerns with the core challenge of length robustness.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Big bird: Transformers for longer sequences

Authors: Zaheer, Manzil, Manzil Zaheer, Guruganesh, Guru, et al. (34 authors total) | **Year/Venue**: 2020 | **URL**: View paper

Abstract

Transformers-based models, such as BERT, have been one of the most successful deep learning models for NLP. Unfortunately, one of their core limitations is the quadratic dependency (mainly in terms of memory) on the sequence length due to their full attention mechanism. To remedy this, we propose, BigBird, a sparse attention mechanism that reduces this quadratic dependency to linear. We show that BigBird is a universal approximator of sequence functions and is Turing complete, thereby preserving...

Relationship Analysis

Both papers belong to the Sparse and Selective Attention category, addressing length generalization through sparsity mechanisms in transformer attention. They overlap in their core approach of reducing quadratic attention complexity to linear by introducing structured sparsity patterns that selectively attend to relevant tokens. The key difference is that the original paper uses α-entmax transformations with learnable temperature scaling (ASEntmax) to achieve dynamic sparsity through exact zeros in attention distributions, while the candidate paper (Big Bird) employs a fixed sparse attention pattern combining random attention, local windows, and global tokens without adaptive sparsity control.

---

2. Unshackling Context Length: An Efficient Selective Attention Approach through Query-Key Compression

Authors: Wang Haoyu, Teng Tong, Haoyu Wang, Guo Tianyu, Tong Teng, et al. (17 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Handling long-context sequences efficiently remains a significant challenge in large language models (LLMs). Existing methods for token selection in sequence extrapolation either employ a permanent eviction strategy or select tokens by chunk, which may lead to the loss of critical information. We propose Efficient Selective Attention (ESA), a novel approach that extends context length by efficiently selecting the most critical tokens at the token level to compute attention. ESA reduces the compu...

Relationship Analysis

Both papers belong to the Sparse and Selective Attention category, addressing length generalization through sparsity mechanisms in attention distributions. While the original paper introduces α-entmax with adaptive temperature scaling (ASEntmax) to achieve dynamic sparsity through exact zero assignments to irrelevant tokens, the candidate paper (ESA) proposes efficient token-level selection via query-key compression and proximity influence. The key difference lies in their approach: the original paper modifies the attention transformation function itself (replacing softmax with α-entmax), whereas the candidate paper maintains standard attention but selectively computes it over compressed, adaptively chosen tokens.

Contributions Analysis

Overall novelty summary. The paper introduces ASEntmax, a learnable-temperature variant of α-entmax attention, and provides theoretical analysis of sparse attention for long-context modeling. It sits within the Sparse and Selective Attention leaf of the taxonomy, which contains only three papers total. This is a relatively sparse research direction compared to denser areas like Positional Encoding methods or Hybrid Architectures. The two sibling papers focus on structured sparsity patterns (Big Bird) and compression-based approaches, while this work emphasizes adaptive sparsity through learnable temperature parameters that interpolate between sparse and dense regimes.

The taxonomy reveals that Sparse and Selective Attention is one of six subtopics under Attention Mechanism Modifications, alongside Dense Attention Variants, Memory-Augmented Attention, and others. Neighboring leaves include Dense Attention Variants (which maintain full distributions) and Memory-Augmented Attention (which extends context through compression or recurrence). The scope note for this leaf emphasizes 'sparsity or selectivity to handle longer sequences efficiently,' distinguishing it from dense variants that modify softmax while preserving full distributions. This work bridges efficiency and generalization concerns, connecting to both the computational motivations of sparse attention and the length extrapolation goals central to the broader taxonomy.

Among 29 candidates examined across three contributions, none were found to clearly refute the proposed work. The theoretical analysis of α-entmax examined 10 candidates with no refutations; ASEntmax examined 9 candidates with no refutations; and the empirical demonstration of extreme length extrapolation examined 10 candidates with no refutations. This suggests that within the limited search scope, the combination of learnable-temperature sparse attention and its application to length extrapolation appears relatively unexplored. However, the search examined only top-K semantic matches and citations, not an exhaustive survey of sparse attention or entmax literature.

Based on the limited literature search, the work appears to occupy a distinct position within sparse attention research by combining adaptive temperature learning with length generalization objectives. The taxonomy context shows this is a less crowded area compared to positional encoding or hybrid architecture research. The absence of refuting candidates among 29 examined suggests novelty within the search scope, though broader entmax or sparse attention communities may contain relevant prior work not captured by semantic search.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Theoretical analysis of α-entmax for long-context modeling

Description: The authors provide theoretical guarantees demonstrating that α-entmax attention avoids attention dispersion, prevents representational collapse, and alleviates over-squashing in long-context transformers. They prove that α-entmax maintains bounded normalized entropy and reduces gradient paths from O(nL) to O(sL).

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Bridging the divide: Reconsidering softmax and linear attention

URL: View paper

Brief Assessment

Softmax Linear Divide[66] focuses on the injective property and local modeling capability of attention mechanisms in vision transformers, not on sparse attention mechanisms like α-entmax for preventing attention dispersion and representational collapse in long-context language modeling.

---

2. Selective Attention: Enhancing Transformer through Principled Context Control

URL: View paper

Brief Assessment

Selective Attention[57] focuses on temperature scaling for softmax attention to control sparsity, not on α-entmax's theoretical properties for preventing attention dispersion and representational collapse in long contexts.

---

3. Resonant pattern shaping through iterative latency induction in contextual token expansion of transformer-based language models

URL: View paper

Brief Assessment

Resonant Pattern Shaping[69] focuses on iterative latency induction and contextual token expansion in transformers, not on α-entmax attention mechanisms or their theoretical properties for preventing attention dispersion and representational collapse.

---

4. How Sparse Attention Approximates Exact Attention? Your Attention is Naturally -Sparse

URL: View paper

Brief Assessment

Naturally Sparse[68] focuses on proving attention is naturally nc-sparse through analyzing softmax sparsity patterns, not on α-entmax mechanisms. The candidate establishes sparsity bounds for standard attention computation, while the original paper provides theoretical guarantees specifically for α-entmax preventing dispersion and collapse.

---

5. Sp2t: Sparse proxy attention for dual-stream point transformer

URL: View paper

Brief Assessment

Sp2t[61] focuses on sparse attention mechanisms for 3D point cloud processing using proxy-based methods, not on theoretical guarantees for α-entmax in long-context transformers or language modeling.

---

6. Multimodal Fusion And Sparse Attention-based Alignment Model for Long Sequential Recommendation

URL: View paper

Brief Assessment

Multimodal Sequential Recommendation[63] focuses on multimodal fusion and sparse attention for sequential recommendation in e-commerce/content platforms, not on theoretical properties of α-entmax attention mechanisms for long-context transformers or language modeling.

---

7. Beyond black-box ai: A theory of interpretable transformers for asset pricing

URL: View paper

Brief Assessment

Interpretable Asset Pricing[65] focuses on financial asset pricing with sparse attention for interpretability in finance applications, not on long-context transformers or the theoretical properties of α-entmax for preventing attention dispersion and representational collapse in NLP tasks.

---

8. Mixture of Contexts for Long Video Generation

URL: View paper

Brief Assessment

Mixture of Contexts[67] addresses long video generation through sparse attention routing for memory retrieval, not theoretical analysis of α-entmax attention mechanisms or their properties in preventing attention dispersion and representational collapse.

---

9. Sparse moe as the new dropout: Scaling dense and self-slimmable transformers

URL: View paper

Brief Assessment

Sparse MoE Dropout[64] focuses on mixture-of-experts training strategies for transformers, not on sparse attention mechanisms or α-entmax. The paper addresses representation collapse through random expert routing during training, which is fundamentally different from the original paper's theoretical analysis of attention dispersion and gradient flow in α-entmax attention.

---

10. On the role of attention masks and layernorm in transformers

URL: View paper

Brief Assessment

Attention Masks LayerNorm[62] focuses on rank collapse and representational preservation in transformers through attention masks and LayerNorm, not on sparse attention mechanisms like α-entmax. The paper does not address attention dispersion prevention or gradient path reduction claims specific to α-entmax.

Contribution 2: Adaptive-Scalable Entmax (ASEntmax)

Description: The authors propose ASEntmax, a novel attention mechanism that extends α-entmax with learnable, head-specific and query-specific temperature parameters. This allows the model to adaptively adjust sparsity based on sequence length and content, balancing between sparse and dense attention regimes.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Is Temperature Sample Efficient for Softmax Gaussian Mixture of Experts?

URL: View paper

Brief Assessment

Temperature Sample Efficiency[55] focuses on temperature parameters in Gaussian mixture of experts for sample efficiency in maximum likelihood estimation, not on adaptive temperature for controlling sparsity in attention mechanisms for transformers.

---

2. Sparse-sensor reconstruction of oblique detonation-wave temperature fields using a diffusion-guided residual coordinate-attention U-shaped network

URL: View paper

Brief Assessment

Detonation Temperature Reconstruction[60] focuses on sparse-sensor reconstruction of temperature fields in detonation waves using diffusion models and coordinate-attention networks, not on attention mechanisms with learnable temperature parameters for controlling sparsity.

---

3. Selective Attention: Enhancing Transformer through Principled Context Control

URL: View paper

Brief Assessment

Selective Attention[57] proposes temperature scaling for softmax attention with query-dependent and position-dependent parameters, but does not extend α-entmax with learnable temperature parameters as ASEntmax does.

---

4. pFedKA: Personalized Federated Learning via Knowledge Distillation with Dual Attention Mechanism

URL: View paper

Brief Assessment

pFedKA[59] focuses on federated learning with knowledge distillation and uses adaptive temperature in a cross-attention module for aligning feature spaces, not for controlling sparsity in attention mechanisms for sequence modeling.

---

5. Measurable shifts in emergent representational forking through probabilistic context folding in large language models

URL: View paper

Brief Assessment

Representational Forking[54] focuses on probabilistic context folding and representational clustering in LLMs, not on adaptive temperature parameters for attention sparsity control as proposed in ASEntmax.

---

6. Scatterbrain: Unifying sparse and low-rank attention

URL: View paper

Brief Assessment

Scatterbrain[52] focuses on combining sparse and low-rank attention approximations using LSH and kernel methods, not on adaptive temperature parameters for controlling sparsity in attention mechanisms as proposed in ASEntmax.

---

7. Semantic flux anchoring in large language models: A framework for stability-oriented representation reinforcement

URL: View paper

Brief Assessment

Semantic Flux Anchoring[51] focuses on semantic stability in LLM representations through anchor mechanisms, not on adaptive temperature parameters for attention sparsity control as in ASEntmax.

---

8. Enhanced Multimodal Recommendation System for Personalized Lifestyle Recommendations

URL: View paper

Brief Assessment

Multimodal Lifestyle Recommendations[53] focuses on adaptive temperature scaling for contrastive loss in multimodal recommendation systems, not on attention mechanisms for transformers. The adaptive temperature mechanism described operates on similarity distributions in embedding spaces for recommendation tasks, which is fundamentally different from ASEntmax's learnable temperature parameters for controlling sparsity in transformer attention distributions.

---

9. Probabilistic contextual resonance in large language model decoding through selfmodulated semantic interference

URL: View paper

Brief Assessment

Contextual Resonance[56] focuses on probabilistic semantic interference in LLM decoding, not on adaptive temperature parameters for attention sparsity control as proposed in ASEntmax.

Contribution 3: Empirical demonstration of extreme length extrapolation

Description: The authors demonstrate through extensive experiments that ASEntmax achieves superior long-context generalization, including 1000× length extrapolation on synthetic tasks and improved perplexity trends and retrieval accuracies at 8× training length on language modeling tasks.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Griffin: Mixing Gated Linear Recurrences with Local Attention for Efficient Language Models

URL: View paper

Brief Assessment

Griffin[78] focuses on architectural efficiency and demonstrates extrapolation capabilities, but does not specifically address attention mechanism sparsity or the 1000× length extrapolation on synthetic tasks that the original paper claims with ASEntmax.

---

2. A comprehensive survey on long context language modeling

URL: View paper

Brief Assessment

Long Context Survey[75] is a comprehensive survey paper that reviews existing work on long context language modeling, including various length extrapolation methods. However, it does not present original empirical experiments demonstrating 1000× length extrapolation on specific synthetic tasks or 8× training length improvements on language modeling tasks with ASEntmax, which are the specific claims of the original paper.

---

3. Linrec: Linear attention mechanism for long-term sequential recommender systems

URL: View paper

Brief Assessment

LinRec[74] focuses on efficient linear attention mechanisms for sequential recommender systems, not on length extrapolation capabilities in language models or synthetic tasks as demonstrated in the original paper.

---

4. mPLUG-Owl3: Towards Long Image-Sequence Understanding in Multi-Modal Large Language Models

URL: View paper

Brief Assessment

mPLUG-Owl3[73] focuses on multi-modal (vision-language) long sequence understanding in MLLMs, not on attention mechanism length extrapolation in language models. The candidate addresses architectural efficiency for processing long image sequences, while the original contribution concerns α-entmax attention's mathematical properties for length generalization in text-based transformers.

---

5. Lightning attention-2: A free lunch for handling unlimited sequence lengths in large language models

URL: View paper

Brief Assessment

Lightning Attention[70] focuses on efficient implementation of linear attention for unlimited sequence lengths during training, not on length extrapolation capabilities of attention mechanisms. The candidate addresses computational efficiency rather than generalization to longer sequences than seen during training.

---

6. Hyena Hierarchy: Towards Larger Convolutional Language Models

URL: View paper

Brief Assessment

Hyena Hierarchy[72] focuses on subquadratic attention alternatives using long convolutions and gating for language modeling, not on α-entmax attention mechanisms or their length extrapolation properties in the context of sparse attention distributions.

---

7. Beyond the limits: A survey of techniques to extend the context length in large language models

URL: View paper

Brief Assessment

Context Extension Survey[79] is a survey paper that reviews existing techniques for extending context length in LLMs. It does not present original empirical experiments demonstrating 1000× length extrapolation or specific results on synthetic/language modeling tasks, and therefore cannot refute the novelty of the original paper's experimental contributions.

---

8. Exposing attention glitches with flip-flop language modeling

URL: View paper

Brief Assessment

Flip-Flop Glitches[76] focuses on attention mechanism failures in flip-flop language modeling tasks, not on length extrapolation capabilities of attention mechanisms in general language modeling contexts. The candidate examines sporadic reasoning errors rather than systematic length generalization performance.

---

9. A length-extrapolatable transformer

URL: View paper

Brief Assessment

Length-Extrapolatable Transformer[77] focuses on training transformers on short texts (1024 tokens) and evaluating on longer sequences up to 8192 tokens, demonstrating length extrapolation capabilities. However, this work does not refute the original paper's novelty claim of achieving 1000× length extrapolation (64 to 65k tokens) using ASEntmax with sparse attention mechanisms, as the candidate achieves more modest extrapolation ratios (8×) and uses different technical approaches (xpos position encoding and blockwise causal attention rather than α-entmax sparse attention).

---

10. LongLoRA: Efficient Fine-tuning of Long-Context Large Language Models

URL: View paper

Brief Assessment

LongLoRA[71] focuses on efficient fine-tuning methods (shifted sparse attention and improved LoRA) for extending context windows of pre-trained LLMs, not on analyzing attention mechanisms' inherent extrapolation capabilities like ASEntmax does with α-entmax.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Long-Context Generalization with Sparse Attention View paper
- [1] Gqa: Training generalized multi-query transformer models from multi-head checkpoints View paper
- [2] Exploring Length Generalization for Transformer-Based Speech Enhancement View paper
- [3] Length Generalization of Causal Transformers without Position Encoding View paper
- [4] Improving length-generalization in transformers via task hinting View paper
- [5] An Exploration of Length Generalization in Transformer-Based Speech Enhancement View paper
- [6] The hidden attention of mamba models View paper
- [7] Exploring length generalization in large language models View paper
- [8] The impact of positional encoding on length generalization in transformers View paper
- [9] Robust and Unbounded Length Generalization in Autoregressive Transformer-Based Text-to-Speech View paper
- [10] Parallelizing linear transformers with the delta rule over sequence length View paper
- [11] Dissecting transformer length extrapolation via the lens of receptive field analysis View paper
- [12] On Vanishing Variance in Transformer Length Generalization View paper
- [13] Attention Mechanisms in Transformers: A General Survey View paper
- [14] Lm-infinite: Simple on-the-fly length generalization for large language models View paper
- [15] Phenaki: Variable Length Video Generation From Open Domain Textual Description View paper

- [16] Leave no context behind: Efficient infinite context transformers with infini-attention View paper
- [17] Scalable-Softmax Is Superior for Attention View paper
- [18] What algorithms can transformers learn? a study in length generalization View paper
- [19] Generalization Error Analysis for Selective State-Space Models Through the Lens of Attention View paper
- [20] Train short, test long: Attention with linear biases enables input length extrapolation View paper
- [21] A formal framework for understanding length generalization in transformers View paper
- [22] Quantitative Bounds for Length Generalization in Transformers View paper
- [23] A Training-Free Length Extrapolation Approach for LLMs: Greedy Attention Logit Interpolation View paper
- [24] A theory for length generalization in learning to reason View paper
- [25] Samba: Simple Hybrid State Space Models for Efficient Unlimited Context Language Modeling View paper
- [26] Neural attention shaping with contextual embedding recalibration in language models View paper
- [27] Enhancing length generalization for attention based knowledge tracing models with linear biases View paper
- [28] Beyond attention: Breaking the limits of transformer context length with recurrent memory View paper
- [29] Big bird: Transformers for longer sequences View paper
- [30] Linear attention sequence parallelism View paper
- [31] Block-state transformers View paper
- [32] Transformers can achieve length generalization but not robustly View paper
- [33] Mega: Moving average equipped gated attention View paper
- [34] GET-Zero: Graph Embodiment Transformer for Zero-Shot Embodiment Generalization View paper
- [35] Long Sequence Modeling with Attention Tensorization: From Sequence to Tensor Learning View paper
- [36] LM-Infinite: Zero-Shot Extreme Length Generalization for Large Language Models View paper
- [37] Mambaextend: A training-free approach to improve long context extension of mamba View paper
- [38] From interpolation to extrapolation: Complete length generalization for arithmetic transformers View paper
- [39] Length generalization in arithmetic transformers View paper
- [40] Landmark attention: Random-access infinite context length for transformers View paper
- [41] Unshackling Context Length: An Efficient Selective Attention Approach through Query-Key Compression View paper
- [42] Toward Length-Extrapolatable Transformers View paper
- [43] Temporal Logical Attention Network for Log-Based Anomaly Detection in Distributed Systems View paper
- [44] Quora Insincere Questions Classification Using Attention Based Model View paper
- [45] The neural data router: Adaptive control flow in transformers improves systematic generalization View paper
- [46] BiLSTM-MLAM: A Multi-Scale Time Series Prediction Model for Sensor Data Based on Bi-LSTM and Local Attention Mechanisms View paper
- [47] TimelyGPT: Extrapolatable Transformer Pre-training for Long-term Time-Series Forecasting in Healthcare View paper
- [48] I-BERT: Inductive Generalization of Transformer to Arbitrary Context Lengths View paper
- [49] Attention-Based Encoder for Online Data Anomaly Classification in Multivariate Time Series View paper
- [50] AttABseq: an attention-based deep learning prediction method for antigenâ□□antibody binding affinity changes based on protein sequences View paper
- [51] Semantic flux anchoring in large language models: A framework for stability-oriented representation reinforcement View paper
- [52] Scatterbrain: Unifying sparse and low-rank attention View paper
- [53] Enhanced Multimodal Recommendation System for Personalized Lifestyle Recommendations View paper
- [54] Measurable shifts in emergent representational forking through probabilistic context folding in large language models View paper
- [55] Is Temperature Sample Efficient for Softmax Gaussian Mixture of Experts? View paper
- [56] Probabilistic contextual resonance in large language model decoding through selfmodulated semantic interference View paper
- [57] Selective Attention: Enhancing Transformer through Principled Context Control View paper
- [58] Sparse structure search for delta tuning View paper
- [59] pFedKA: Personalized Federated Learning via Knowledge Distillation with Dual Attention Mechanism View paper
- [60] Sparse-sensor reconstruction of oblique detonation-wave temperature fields using a diffusion-guided residual coordinate-attention U-shaped network View paper
- [61] Sp2t: Sparse proxy attention for dual-stream point transformer View paper
- [62] On the role of attention masks and layernorm in transformers View paper
- [63] Multimodal Fusion And Sparse Attention-based Alignment Model for Long Sequential Recommendation View paper
- [64] Sparse moe as the new dropout: Scaling dense and self-slimmable transformers View paper
- [65] Beyond black-box ai: A theory of interpretable transformers for asset pricing View paper
- [66] Bridging the divide: Reconsidering softmax and linear attention View paper
- [67] Mixture of Contexts for Long Video Generation View paper
- [68] How Sparse Attention Approximates Exact Attention? Your Attention is Naturally -Sparse View paper
- [69] Resonant pattern shaping through iterative latency induction in contextual token expansion of transformer-based language models View paper
- [70] Lightning attention-2: A free lunch for handling unlimited sequence lengths in large language models View paper
- [71] LongLoRA: Efficient Fine-tuning of Long-Context Large Language Models View paper
- [72] Hyena Hierarchy: Towards Larger Convolutional Language Models View paper
- [73] mPLUG-Owl3: Towards Long Image-Sequence Understanding in Multi-Modal Large Language Models View paper
- [74] Linrec: Linear attention mechanism for long-term sequential recommender systems View paper
- [75] A comprehensive survey on long context language modeling View paper
- [76] Exposing attention glitches with flip-flop language modeling View paper
- [77] A length-extrapolatable transformer View paper
- [78] Griffin: Mixing Gated Linear Recurrences with Local Attention for Efficient Language Models View paper
- [79] Beyond the limits: A survey of techniques to extend the context length in large language models View paper