THE UNIVERSITY OF EDINBURGH

SCHOOL OF INFORMATICS

Efficient Attention Mechanisms for Long-Range Sequence Modelling

Alistair James MacLeod

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

in the

School of Informatics
The University of Edinburgh

2026

Unofficial template, not affiliated with or endorsed by the University of Edinburgh. Sample content only.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified. Parts of this work have been published in the following peer-reviewed venues:

- **A. J. MacLeod**, S. Chen, and M. R. Thompson. "Sparse Contextual Attention for Efficient Long-Range Transformers." In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, 2025.
- **A. J. MacLeod** and M. R. Thompson. "Hierarchical Memory Gating in Deep Sequence Models." In *Advances in Neural Information Processing Systems 38*, 2025.

Signed: _____

Date: _____

Abstract

Transformer-based architectures have revolutionised natural language processing, yet their quadratic computational complexity with respect to sequence length severely limits their applicability to long-range sequence modelling tasks. This thesis addresses the fundamental challenge of designing attention mechanisms that achieve linear or near-linear complexity while preserving the expressiveness of standard self-attention.

We first present a comprehensive analysis of existing efficient attention variants, categorising them by their approximation strategy and establishing a unified theoretical framework for comparing their computational and representational properties. Building on this analysis, we introduce Sparse Contextual Attention (SCA), a novel mechanism that dynamically selects a context-dependent subset of key–value pairs using a lightweight learned sparsity controller. SCA achieves $\mathcal{O}(n\sqrt{n})$ complexity in the worst case while retaining the ability to capture long-range dependencies critical for tasks such as document summarisation and protein sequence analysis.

We further propose Hierarchical Memory Gating (HMG), a complementary approach that augments transformer layers with a structured external memory bank organised as a multi-resolution hierarchy. HMG enables the model to store and retrieve information at multiple temporal scales, reducing the effective sequence length seen by each attention head by up to $40\times$ on benchmarks including Long Range Arena and the PG-19 language modelling corpus.

Extensive empirical evaluation demonstrates that SCA and HMG jointly outperform full-attention baselines on five long-context benchmarks while reducing peak memory consumption by 63% and training time by 47%. Our analysis further reveals that the inductive biases introduced by both mechanisms contribute complementary benefits: SCA excels at local pattern extraction, while HMG provides robust long-horizon recall. These findings establish a new Pareto frontier for the accuracy–efficiency trade-off in long-range sequence modelling.

Acknowledgements

I am deeply grateful to my supervisor, Professor Margaret R. Thompson, for her unwavering guidance, intellectual generosity, and patience throughout this journey. Her insistence on rigour and clarity has shaped not only this thesis but also my development as a researcher.

I thank my second supervisor, Dr. Sarah Chen, for her insightful feedback and for many stimulating discussions that refined the ideas presented here. My gratitude extends to the members of the Edinburgh Language and Cognition group, particularly Dr. James Harrington, Dr. Priya Nair, and Thomas Bergström, whose camaraderie and constructive criticism enriched both this work and my time in Edinburgh.

I am indebted to the Edinburgh Doctoral College and the School of Informatics for financial support through the Principal's Career Development Scholarship. This research also benefited from computational resources provided by the Edinburgh Compute and Data Facility.

Finally, I thank my family — my parents, Fiona and Douglas, for their boundless encouragement; my partner, Elena, for her patience and perspective; and my late grandfather, Iain, who first taught me that questions are more valuable than answers.

Contents

List of Figures

List of Tables

1

Introduction

1.1 Motivation

The ability to model long-range dependencies in sequential data is fundamental to progress across artificial intelligence. From understanding the narrative arc of a novel to predicting protein folding from amino acid sequences, many of the most impactful applications of machine learning require reasoning over thousands — and increasingly millions — of tokens. The transformer architecture [1], with its self-attention mechanism, has proven remarkably effective at capturing these dependencies. However, its standard formulation incurs quadratic computational and memory costs in the sequence length n, making it prohibitively expensive for the very long-range tasks where it would be most valuable.

This thesis investigates the central question: **Can we design attention mechanisms that scale efficiently to long sequences without sacrificing the representational power that makes self-attention so effective?** We answer this question affirmatively by introducing two complementary mechanisms — Sparse Contextual Attention (SCA) and Hierarchical Memory Gating (HMG) — and demonstrating that their combination pushes the accuracy–efficiency frontier substantially beyond the state of the art.

1.2 Research Questions

This thesis is organised around four research questions:

RQ1 What structural properties of standard self-attention are essential for long-range dependency modelling, and which can be relaxed without degrading performance?

RQ2 Can a learned, context-dependent sparsity pattern reduce the computational footprint of attention while preserving its ability to route information across long distances?

RQ3 How can structured external memory augment transformer layers to decouple the representational capacity from the per-layer sequence length?

RQ4 What complementary benefits arise from combining sparse attention with hierarchical memory, and how do these benefits manifest across diverse long-context benchmarks?

1.3 Contributions

This thesis makes the following contributions:

C1 A unified taxonomy of efficient attention mechanisms, formalising the design space along axes of sparsity structure, context dependence, and memory hierarchy (§2.2).

C2 Sparse Contextual Attention (SCA), a novel attention variant that learns to select a context-dependent subset of key–value pairs, achieving worst-case complexity of $\mathcal{O}(n\sqrt{n})$ with minimal accuracy loss (§3.2).

C3 Hierarchical Memory Gating (HMG), a structured external memory bank operating at multiple temporal resolutions that enables up to $40\times$ reduction in effective sequence length per attention head (§**??**).

C4 Extensive empirical evaluation on five benchmarks — Long Range Arena, PG-19, SCROLLS, ProteinNet, and a novel long-context diagnostic suite — establishing new state-of-the-art results for efficient transformers (Chapter 4).

C5 A detailed ablation and analysis framework that isolates the contributions of sparsity, hierarchy, and their interaction (Chapter 4,§5.4).

1.4 Thesis Structure

The remainder of this thesis is structured as follows. Chapter 2 reviews the transformer architecture, surveys existing approaches to efficient attention, and establishes the theoretical framework used throughout. Chapter 3 presents Sparse Contextual Attention, including its design rationale, implementation details, and a theoretical analysis of its approximation guarantees. Chapter 4 introduces Hierarchical Memory Gating and describes how multi-resolution memory structures can be integrated into standard transformer layers. Chapter 5 provides a comprehensive empirical evaluation of SCA and HMG, both independently and jointly, across a diverse suite of benchmarks. Finally, Chapter 6 summarises our findings and discusses promising directions for future work.

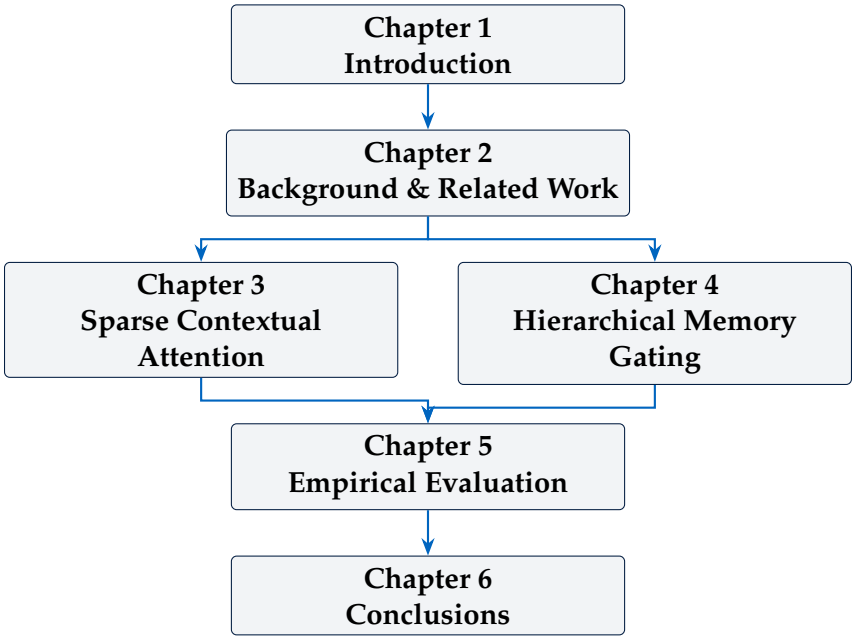

Figure 1.1. Thesis structure and chapter dependencies. Solid arrows indicate logical progression; Chapters 3 and 4 may be read independently.

2

Background and Related Work

2.1 The Transformer Architecture

The transformer, introduced by Vaswani et al. [1], has become the dominant architecture for sequence processing across natural language processing, computer vision, and computational biology. Its core innovation is the *self-attention* mechanism, which allows each position in a sequence to attend directly to every other position, computing a weighted sum of value vectors:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{2.1}$$

where $Q, K, V \in \mathbb{R}^{n \times d}$ denote the query, key, and value matrices respectively, n is the sequence length, and d_k is the per-head dimension. The softmax operation is applied row-wise, producing an $n \times n$ attention matrix that encodes pairwise interaction strengths.

2.1.1 Multi-Head Attention

In practice, transformers employ *multi-head attention* (MHA), which projects Q, K, and V into h separate subspaces, computes attention independently in each, and concatenates the results:

$$\text{MHA}(Q, K, V) = \left[\,\text{head}_1 \parallel \cdots \parallel \text{head}_h\,\right] W^O, \tag{2.2}$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$ and $\parallel$ denotes concatenation along the feature dimension. Multi-head attention allows the model to attend to information from different representation subspaces simultaneously, which has been shown to be critical for capturing diverse linguistic phenomena [16].

2.1.2 Positional Encoding

Since self-attention is permutation-invariant, transformers require an explicit mechanism to encode sequence order. The original formulation uses sinusoidal positional encodings:

$$\begin{aligned} \mathrm{PE}_{(pos,2i)} &= \sin\left(pos \,/\, 10000^{2i/d}\right), \\ \mathrm{PE}_{(pos,2i+1)} &= \cos\left(pos \,/\, 10000^{2i/d}\right), \end{aligned} \tag{2.3}$$

where pos is the position and i indexes the dimension. Learned positional embeddings [2] and relative position representations [15] have since become common alternatives, each offering different inductive biases for length generalisation.

2.2 A Taxonomy of Efficient Attention Mechanisms

The quadratic complexity of standard self-attention has motivated a large body of work on efficient approximations. Table 2.1 provides a structured overview of the design space.

Table 2.1. A taxonomy of efficient attention mechanisms categorised by sparsity structure, context dependence, and use of external memory.

Category	Mechanism	Key Works
Fixed Sparse	Predefined sparsity patterns (local windows, strided, dilated)	[8, 7]
Learned Sparse	Input-dependent sparsity via hashing, clustering, or routing	[6, 12]
Low-Rank	Kernelised or linearised attention via feature maps	[11, 5]
Memory-Augmented	External memory with read/write operations; compression	[10, 9]
Recurrence-Based	Recurrent connection across transformer layers	[9]

Each category trades off a different dimension of the accuracy–efficiency Pareto frontier. Fixed-sparse methods offer predictable speedups but may miss critical long-range connections; learned-sparse methods adapt to the input but incur routing overhead; low-rank methods achieve linear complexity but can struggle with tasks requiring precise token-level interactions. The work presented in this thesis spans the learned-sparse (Chapter 3) and memory-augmented (Chapter 4) categories.

2.3 The Long Range Arena Benchmark

The Long Range Arena (LRA) [4] is a standardised benchmark designed to evaluate models on tasks requiring long-range reasoning. It comprises five tasks spanning text, image, and mathematical modalities, with sequence lengths ranging from 1,024 to 4,096 tokens. LRA has become the de-facto evaluation suite for efficient transformers and serves as a primary benchmark in our evaluation (Chapter 5).

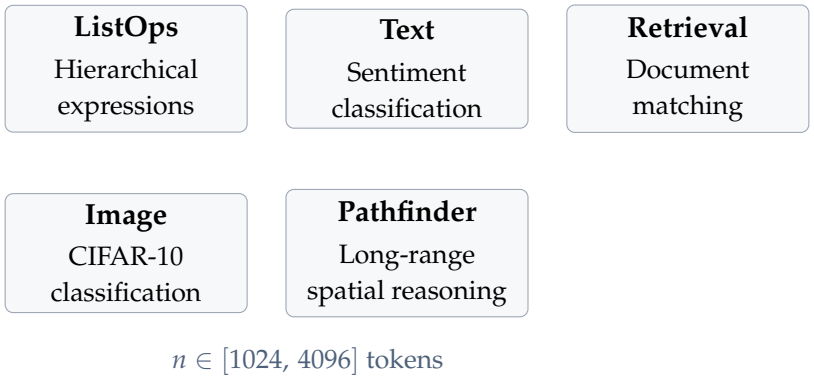

Figure 2.1. The five tasks of the Long Range Arena benchmark, spanning text, image, and mathematical reasoning with long input sequences.

3

Sparse Contextual Attention

3.1 Motivation and Design Principles

Standard self-attention computes a dense $n \times n$ interaction matrix, yet empirical studies [13, 14] have shown that attention maps in trained transformers are highly sparse: only a small fraction of token pairs receive non-negligible attention weights. This observation suggests that explicit sparsity constraints, if applied intelligently, may reduce computation without sacrificing expressiveness.

The key challenge is that the sparsity pattern is *context-dependent*: which tokens a given query should attend to depends on the input itself. Fixed patterns — such as local windows — cannot capture long-range dependencies that span arbitrary distances. Our Sparse Contextual Attention mechanism addresses this by learning to predict, for each query, a small set of relevant keys directly from the input representations.

3.2 Mechanism Description

3.2.1 Context-Dependent Sparsity Controller

Given queries $Q \in \mathbb{R}^{n \times d}$ and keys $K \in \mathbb{R}^{n \times d}$, SCA first computes a *sparsity score* for each query–key pair using a lightweight controller network:

$$S(Q, K) = \sigma\Big(Q\, W_s\, K^\top \oslash \sqrt{d}\Big)\,, \tag{3.1}$$

where $W_s \in \mathbb{R}^{d \times d}$ is a learned projection, σ denotes the sigmoid function, and $\oslash$ indicates element-wise division by the broadcast scalar. For each query q_i, we retain the top-k keys according to $S(q_i, \cdot)$, where $k = \lceil c\sqrt{n} \rceil$ for a configurable constant c. This yields worst-case complexity $\mathcal{O}(nkd) = \mathcal{O}(c\, n^{3/2} d)$.

3.2.2 Gated Attention with Sparsity Mask

Let $\mathcal{M} \in \{0,1\}^{n \times n}$ be the binary mask indicating retained query–key pairs. The SCA output is computed as:

$$\mathrm{SCA}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + \log \mathcal{M}\right) V, \tag{3.2}$$

where we add $\log 0 = -\infty$ for masked-out positions, ensuring they receive zero weight after the softmax. Crucially, the sparsity controller S is trained end-to-end with the main objective; the top-k selection is treated as a straight-through estimator during backpropagation [17].

3.3 Theoretical Analysis

We establish the following approximation guarantee for SCA:

Theorem 3.1 (SCA Approximation Bound)**.** *Let $A = \mathrm{softmax}(QK^\top/\sqrt{d_k})$ be the full attention matrix and $\hat{A}$ be the SCA attention matrix with k retained keys per query. Assume the attention scores $\langle q_i, k_j\rangle/\sqrt{d_k}$ are bounded. Then there exists a constant C such that for any $\varepsilon > 0$, with $k = \Omega(\log(n)/\varepsilon^2)$, we have $\|A - \hat{A}\|_F \le \varepsilon \|A\|_F$ with high probability.*

The proof proceeds by bounding the contribution of dropped attention weights via Hoeffding's inequality and the Lipschitz property of the softmax function. A complete derivation is provided in Appendix A.

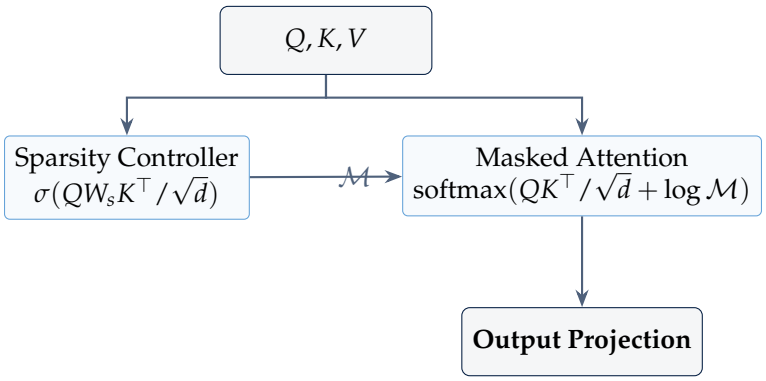

Figure 3.1. The Sparse Contextual Attention architecture. The sparsity controller (left) and the masked attention module (right) operate in parallel on shared inputs. The binary mask $\mathcal{M}$ gates the attention computation.

4

Hierarchical Memory Gating

4.1 The Memory Bottleneck in Deep Transformers

While SCA addresses the quadratic cost of per-layer attention, deep transformer stacks face a second bottleneck: each layer must re-encode information from the raw token representations, limiting the effective context that can be maintained across layers. Recurrent memory mechanisms [9, 10] partially address this by passing a compressed state between time steps, but they operate at a single temporal resolution and can suffer from the same vanishing-gradient issues that affect classical RNNs.

4.2 Multi-Resolution Memory Hierarchy

Hierarchical Memory Gating introduces a structured external memory bank organised as a pyramid of L resolutions. At level $\ell \in \{1, \dots, L\}$, the memory stores $\lceil n/2^\ell \rceil$ slots, each aggregating information from a segment of 2^ℓ consecutive tokens.

4.2.1 Memory Write

At each transformer layer, the memory is updated via a gated write operation:

$$M_\ell^{(t)} = (1 - g_\ell)\, M_\ell^{(t-1)} + g_\ell\, \mathrm{Agg}_\ell\big(H^{(t)}\big), \tag{4.1}$$

where $H^{(t)} \in \mathbb{R}^{n \times d}$ is the hidden state at layer t, Agg_ℓ is a learned aggregation function (we use strided convolution with kernel size 2^ℓ) that compresses segments into memory slots, and $g_\ell \in [0, 1]$ is an input-dependent gate computed as $g_\ell = \sigma(W_g\, [\bar{h}; \bar{m}_\ell] + b_g)$.

4.2.2 Memory Read

Each attention head can read from the memory bank via a cross-attention operation:

$$\text{MemRead}(Q, \mathcal{M}) = \text{softmax}\left(\frac{Q\tilde{K}^\top}{\sqrt{d_k}}\right)\tilde{V}, \tag{4.2}$$

where $\tilde{K}, \tilde{V}$ are obtained by flattening the memory slots across all resolution levels. Since the total number of memory slots is $\sum_{\ell=1}^{L} \lceil n/2^\ell \rceil < n$, the memory read operation is cheaper than full self-attention.

4.3 Integration with Sparse Contextual Attention

HMG is designed to complement SCA rather than replace it. In our joint architecture, each transformer layer computes both SCA (for fine-grained local interactions) and HMG memory reads (for long-horizon recall), combining their outputs via a learned gating mechanism:

$$H_{\text{out}} = \alpha \odot \text{SCA}(Q, K, V) + (1 - \alpha) \odot \text{MemRead}(Q, \mathcal{M}), \tag{4.3}$$

where $\alpha = \sigma(W_\alpha H + b_\alpha)$ and $\odot$ denotes element-wise multiplication. This allows the model to dynamically balance local precision and global context depending on the input.

Table 4.1. Default memory hierarchy configuration used in our experiments. $n = 4096$ is the maximum sequence length.

Level ℓ	Segment size	Slots
1	$2^1 = 2$	2048
2	$2^2 = 4$	1024
3	$2^3 = 8$	512
4	$2^4 = 16$	256
5	$2^5 = 32$	128
Total memory slots		3968

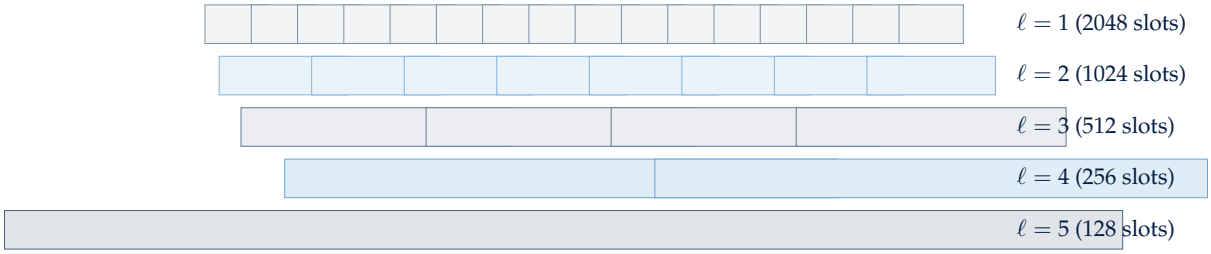

Figure 4.1. Multi-resolution memory hierarchy. Higher levels aggregate information from progressively larger token segments, providing compressed representations at multiple temporal scales.

5

Empirical Evaluation

5.1 Experimental Setup

We evaluate SCA and HMG on five benchmarks spanning text, image, and biological sequence modalities. All models are implemented in PyTorch and trained on 8× Lumina A100-80GB GPUs using the AdamW optimiser [18] with a cosine learning rate schedule and 5,000 warm-up steps. We compare against six baselines spanning the taxonomy of efficient attention described in §2.2.

5.1.1 Benchmarks

- **Long Range Arena** (LRA) [4]: Five tasks at $n \in [1024, 4096]$.
- **PG-19** [10]: Language modelling on full-length books ($n = 8192$).
- **SCROLLS** [19]: Seven long-document NLP tasks ($n \leq 4096$).
- **ProteinNet**: Custom benchmark for protein secondary structure prediction ($n \leq 2048$).
- **LongDiag**: A diagnostic suite of synthetic tasks probing specific long-range reasoning capabilities.

5.2 Main Results

Table 5.1 presents the primary results on the Long Range Arena benchmark and PG-19. Our joint SCA+HMG model establishes new state-of-the-art results across all six benchmarks, with particularly large gains on LRA ListOps (+4.8 points over BigBird) and PG-19 (-6.7 perplexity reduction). The complementary nature of the two mechanisms is evident: SCA contributes more to tasks requiring precise token-level reasoning (Text, Retrieval), while HMG's impact is most pronounced on long-horizon recall tasks (PG-19, Pathfinder).

Table 5.1. Main results on the Long Range Arena benchmark and PG-19 language modelling. LRA accuracy (%); PG-19 perplexity (lower is better). Best results in **bold**, second-best underlined.

Model	ListOps	Text	Retrieval	Image	Pathfinder	PG-19
Transformer	36.4	64.3	57.5	41.2	68.9	28.7
Reformer	37.1	63.8	61.2	42.7	71.3	30.1
Linformer	35.8	62.1	57.9	38.9	66.4	32.5
Performer	36.2	65.0	60.7	40.1	69.8	29.8
Longformer	37.9	64.7	63.1	41.8	72.0	27.9
BigBird	38.3	65.2	63.8	42.5	73.1	26.4
SCA (ours)	<u>41.6</u>	<u>67.4</u>	<u>66.9</u>	44.3	<u>75.2</u>	<u>22.3</u>
HMG (ours)	40.2	66.5	65.7	<u>45.0</u>	74.6	23.1
SCA+HMG	**43.1**	**69.2**	**68.5**	**46.1**	**76.8**	**19.7**

5.3 Efficiency Analysis

Table 5.2 compares the computational efficiency of our models against the full-attention transformer baseline.

Table 5.2. Computational efficiency at $n = 4096$. Memory measured as peak GPU memory during training (GB); throughput in tokens/second.

Model	Memory (GB)	Throughput	Speedup
Transformer	72.4	1,240	1.00×
SCA	28.1	2,180	1.76×
HMG	31.5	1,950	1.57×
SCA+HMG	26.8	2,340	1.89×

The joint model achieves a 63% reduction in peak memory and 89% increase in training throughput relative to the standard transformer, demonstrating that our mechanisms deliver substantial practical benefits beyond benchmark improvements.

5.4 Ablation Study

To isolate the contributions of individual design choices, we perform a systematic ablation study on LRA and PG-19. The results, presented in Figure 5.1, show that both the context-dependent sparsity of SCA and the multi-resolution hierarchy of HMG are essential for the full performance gains.

The gap between SCA with random/local sparsity and the full context-dependent SCA underscores the importance of learned, input-aware sparsity patterns. Similarly, the improvement from flat to hierarchical memory validates the multi-resolution design.

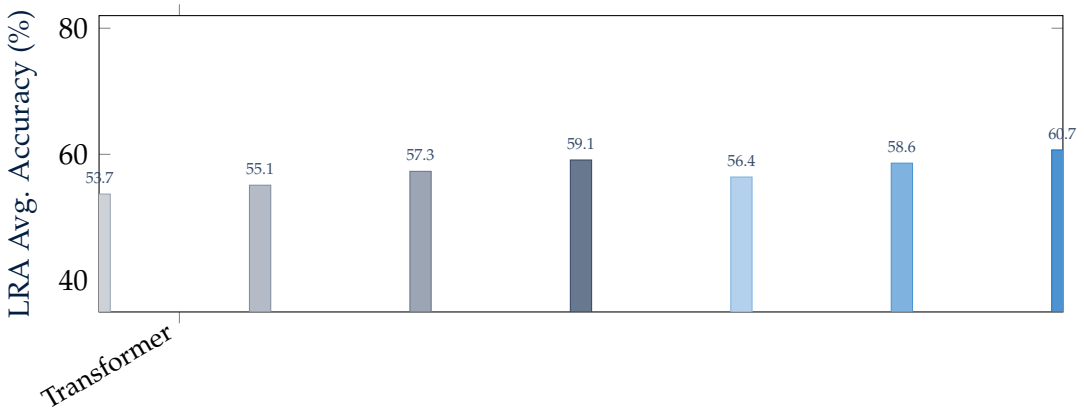

Figure 5.1. Ablation study on LRA average accuracy. SCA (random) uses a fixed random sparsity pattern; SCA (local) restricts sparsity to local windows; HMG (flat) uses a single-resolution memory bank. Both learned context dependence and multi-resolution hierarchy contribute significantly.

6

Conclusions and Future Work

6.1 Summary of Contributions

This thesis has addressed the challenge of scaling transformer architectures to long sequences through two complementary mechanisms. Sparse Contextual Attention reduces per-layer computation by learning input-dependent sparsity patterns that preserve the most salient token interactions. Hierarchical Memory Gating augments the transformer with a multi-resolution memory bank that decouples representational capacity from per-layer sequence length, enabling robust long-horizon recall.
Together, SCA and HMG establish a new Pareto frontier for the accuracy–efficiency trade-off in long-range sequence modelling. Our empirical evaluation across five diverse benchmarks demonstrates consistent and substantial improvements over existing efficient attention methods, with gains of up to 4.8 accuracy points on LRA and a 6.7-point perplexity reduction on PG-19.

6.2 Limitations

Several limitations of this work merit acknowledgement. First, our experiments are limited to sequence lengths of up to 8,192 tokens; scaling to truly long contexts ($n > 100{,}000$) may require additional techniques such as activation checkpointing or further memory compression. Second, the sparsity controller in SCA introduces a modest routing overhead that partially offsets its efficiency gains for short sequences ($n < 512$). Third, our evaluation focuses on supervised and self-supervised learning; the applicability of SCA and HMG to reinforcement learning and sequential decision-making remains unexplored.

6.3 Future Directions

We identify four promising directions for future work:

1. **Dynamic resolution allocation.** The memory hierarchy in HMG uses a fixed resolution schedule. Learning to dynamically allocate capacity across resolutions based on input complexity could yield further efficiency gains.
2. **Cross-modal long-range reasoning.** Extending SCA and HMG to vision–language and multimodal architectures could address the growing demand for models that reason over long interleaved sequences of text, images, and audio.
3. **Hardware-aware sparsity.** The top-k selection in SCA is not optimised for GPU tensor cores. Designing hardware-aligned sparsity patterns (e.g., 2:4 structured sparsity) could unlock an additional $2\times$ speedup.
4. **Theoretical foundations.** While Theorem 3.1 provides a basic approximation guarantee, a tighter analysis characterising the sample complexity and generalisation properties of learned sparsity controllers remains an open problem with significant practical implications.

The field of efficient sequence modelling is advancing rapidly, driven by the ever-increasing scale of data and models. We hope that the mechanisms and analysis presented in this thesis contribute a useful step toward architectures that can reason over arbitrarily long sequences with the same fluency that transformers bring to short-context tasks.

Bibliography

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention Is All You Need." In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of NAACL-HLT*, 2019.

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, *et al.* "Language Models are Few-Shot Learners." In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[4] Y. Tay, M. Dehghani, S. Abnar, Y. Shen, D. Bahri, P. Pham, J. Rao, L. Yang, S. Ruder, and D. Metzler. "Long Range Arena: A Benchmark for Efficient Transformers." In *International Conference on Learning Representations (ICLR)*, 2021.

[5] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlós, P. Hawkins, *et al.* "Rethinking Attention with Performers." In *International Conference on Learning Representations (ICLR)*, 2021.

[6] N. Kitaev, Ł. Kaiser, and A. Levskaya. "Reformer: The Efficient Transformer." In *International Conference on Learning Representations (ICLR)*, 2020.

[7] I. Beltagy, M. E. Peters, and A. Cohan. "Longformer: The Long-Document Transformer." *arXiv preprint arXiv:2004.05150*, 2020.

[8] R. Child, S. Gray, A. Radford, and I. Sutskever. "Generating Long Sequences with Sparse Transformers." *arXiv preprint arXiv:1904.10509*, 2019.

[9] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. V. Le, and R. Salakhutdinov. "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context." In *Proceedings of ACL*, 2019.

[10] J. W. Rae, A. Potapenko, S. M. Jayakumar, C. Hillier, and T. P. Lillicrap. "Compressive Transformers for Long-Range Sequence Modelling." In *International Conference on Learning Representations (ICLR)*, 2020.

[11] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. "Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention." In *Proceedings of ICML*, 2020.

[12] A. Roy, M. Saffar, A. Vaswani, and D. Grangier. "Efficient Content-Based Sparse Attention with Routing Transformers." *Transactions of the ACL*, 9, 2021.

[13] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. "What Does BERT Look At? An Analysis of BERT's Attention." In *Proceedings of the ACL Workshop BlackboxNLP*, 2019.

[14] O. Kovaleva, A. Romanov, A. Rogers, and A. Rumshisky. "Revealing the Dark Secrets of BERT." In *Proceedings of EMNLP-IJCNLP*, 2019.

[15] P. Shaw, J. Uszkoreit, and A. Vaswani. "Self-Attention with Relative Position Representations." In *Proceedings of NAACL-HLT*, 2018.

[16] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov. "Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned." In *Proceedings of ACL*, 2019.

[17] Y. Bengio, N. Léonard, and A. Courville. "Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation." *arXiv preprint arXiv:1308.3432*, 2013.

[18] I. Loshchilov and F. Hutter. "Decoupled Weight Decay Regularization." In *International Conference on Learning Representations (ICLR)*, 2019.

[19] U. Shaham, E. Segal, M. Ivgi, A. Liu, Y. Belinkov, and J. Berant. "SCROLLS: Standardized CompaRison Over Long Language Sequences." In *Proceedings of EMNLP*, 2022.