Efficient Attention Mechanisms for Long-Context Transformer Models

Your Name

Abstract

This dissertation studies sparse and learned-routing attention mechanisms for transformer models operating on long sequences. We show that routing quality, rather than sparsity ratio alone, governs the trade-off between computational efficiency and model accuracy, and we characterize how this trade-off shifts with model scale. Across model sizes from 125M to 2.7B parameters, we find that larger models tolerate more aggressive sparsification, narrowing the gap to dense attention as scale increases. Our findings inform the design of compute-optimal long-context architectures and motivate treating sparsity ratio as a scale-dependent hyperparameter rather than a fixed architectural choice.

Acknowledgments

I am deeply grateful to my advisor, Mara Okafor, for her patient guidance and for the countless whiteboard sessions that shaped this work. Thanks also to my committee members, Elena Vasquez and David Torres, for their careful feedback at every stage. I thank my labmates for years of shared debugging sessions and honest criticism, and the Fairview systems team for keeping the cluster running through more deadline-week jobs than I'd like to admit.

Finally, thanks to my family for their unwavering support throughout this long process.

Dedication

To my parents, who never once asked when I would finish.