THE UNIVERSITY of EDINBURGH

Dr Pasquale Minervini
Lecturer at the School of Informatics
The University of Edinburgh
Edinburgh EH8 9AB, United Kingdom
p.minervini@ed.ac.uk

14 April 2025

This thesis, titled “Efficient Large Language Models with Conditional Computation” and authored by Sebastian Jaszczur, analyses several methods for improving the efficiency of LLMs via Conditional Computation. It covers several approaches and includes the development of new sparse Transformer layers (Sparse Feed-Forward and Sparse QKV); the introduction of fine-grained Mixture of Experts (MoE) and corresponding scaling laws; a continuous MoE variant called Mixture of Tokens (MoT); a new method (SPLiCe) for improving long-context utilisation via structured data packing during fine-tuning; and the integration of MoE with the Mamba SSM architecture. Overall, this work has a broad breadth -- it spans architectural extensions/modifications; theoretical analyses of scaling properties; and techniques to improve training stability and long-context handling in LLMs.

Ch. 1 introduces the problem of the computational cost associated with training and deploying LLMs; it emphasises that high computational costs associated with LLMs are a critical bottleneck and states the goal of this thesis, namely using conditional computation to improve their efficiency.

Ch. 2 introduces new sparse variants for layers in Transformer-based architecture. Specifically, it proposes a new Sparse Feed-Forward layer that uses dynamic sparsity to activate only a portion of the parameters for each token; and a Sparse QKV layer that aims at reducing the quadratic complexity of self-attention mechanisms. These sparse layer variants are used to create a new Transformer variant, namely "Terraformer", which yields significant efficiency improvements during inference (e.g. up to ~20x for a 17B model) without significantly decreasing downstream performance. Here, the focus was mainly on inference efficiency rather than pre-training efficiency.

Potential issue: in Tab. 2.6, Transformer-based architectures are compared with Terraformer-based ones; the former was trained for ~500k steps while the latter was on ~125-175k. Due to a change in batch size, Terraformer with 175k steps was trained on significantly more training samples -- is this an issue in the evaluation? What happens if you use the same amounts of pre-training data? Very small issue: 2.3.1 says that argmax is not differentiable while it is (it's piece-wise constant, which doesn't work well with back-propagation since gradients are zero almost everywhere).

Ch. 3 introduces "granularity" in MoE models, where each expert is, in turn, represented by a set of sub-experts. The author devises new scaling laws that incorporate granularity, model size, and number of training steps; and shows that, with optimal granularity settings, MoE-based models can ~always yield more accurate results than dense Transformers under any compute budget (e.g., by yielding ~40x savings at 10^25 FLOPs).

Ch. 4 proposes a new continuous variant of MoE called Mixture of Tokens (MoT) -- rather than using discrete routing (as most MoE architectures), MoTs create a weighted combination of tokens across examples. Using a continuous gating function rather than a discrete decision step helps avoid the training difficulties that stem from discrete routing schemes while maintaining the efficiency benefits of MoE architectures. The author shows that MoTs can achieve performance comparable to sparse MoE variants while showing significantly better stability in low-precision training environments. Furthermore, this chapter also introduces "transition tuning", a technique that enables converting a MoT model to a sparse MoE.

Ch. 5 addresses the challenge of long-context windows in LLMs; the author proposes SPLiCe (Structured Packing for Long Context), a method that organises the training data by collating semantically related documents into a single context. Experiments on varying model sizes show that fine-tuning with SPLiCe on 2B-6B tokens significantly improves accuracy on tasks requiring processing very long contexts, such as open-book open-domain QA, in-context learning, and needle-in-a-haystack type of retrieval problems.

Ch. 6 explores the integration of MoEs with Mamba, a recent SSM architecture designed to efficiently process very long sequences. Specifically, this chapter introduces MoE-Mamba, which interleaves Mamba and MoE layers, and shows that this can significantly produce more accurate results than both Mamba and Transformer-based MoE architectures -- in some experiments, MoE-Mamba more than halves the number of training steps in comparison with Mamba to reach the same performance levels.

The work in this thesis provides a great contribution to NLP via the analysis of several conditional computation approaches; it is very practically relevant, as it offers several novel and effective approaches for scaling LLMs while reducing their computational requirements. I'm really impressed by the author's comprehensive treatment of MoE-based architectures and the introduction of MoE granularity as a hyper-parameter, revising previous assumptions about MoE scaling. Each contribution in this thesis was validated on multiple scales, analysing trade-offs and limitations. I really loved the adoption and extension of Mamba in Ch. 6 since we are likely to get back to ~recurrent architectures at some point to handle the ever-increasing context lengths (e.g., caused by LLM-based agentic systems). Overall, the thesis is written in a clear and well-organised manner; it follows a logical structure that aids the reader in understanding a significant body of work. **I deem the thesis sufficient to grant a PhD**.

Sincerely,
Pasquale Minervini