Title:
Hypothesis: Efficient Scaling and Specialization of Language Models via Layer Reuse and Domain-Focused Training

Abstract:
Large language models (LLMs) have achieved remarkable performance across a wide range of tasks, but their increasing scale leads to substantial computational and resource demands. In this paper, we hypothesize that similar or even improved performance may be achieved more efficiently through three interrelated strategies: (1) initializing larger models by reusing layers from smaller models trained with the same hidden size, (2) reusing not only the outer layers but also the middle layers during model expansion, and (3) training medium-sized models tailored to specific domains, such as medicine, which may yield comparable results to much larger general-purpose models. These ideas, while not yet experimentally verified, suggest promising directions for making LLMs more resource-efficient, interpretable, and adaptable to specialized use cases.

1. Introduction

As language models continue to grow in size and capability, so do their requirements for training data, compute resources, and energy [1]. This raises the question of whether more efficient and targeted strategies can be developed to retain high performance without resorting to massive scale. In this work, we present a set of related hypotheses: that certain forms of architectural reuse and domain-focused training may offer more efficient paths to effective language models. Specifically, we explore the theoretical potential of reusing learned parameters from smaller models when scaling up [2], and the idea that medium-sized, domain-specific models may rival much larger models on specialized tasks [3]. These strategies have not yet been systematically tested, but they point toward a more efficient and modular approach to building future LLMs.

2. Hypothesis One: Progressive Scaling via Layer Transfer

We hypothesize that large models can be more efficiently trained by first training smaller models with the same hidden size, i.e. embedding dimensions, and then transferring selected layers to initialize deeper architectures. Typically, the first and last few transformer layers—responsible for input encoding and output generation—may capture fundamental representations that are transferable across scales [4]. Reusing these layers could reduce training time, improve convergence behavior, and lower the computational cost associated with initializing large models from scratch. While this approach aligns with the modular nature of transformer architectures [5], its actual efficacy remains to be tested in controlled settings.

3. Hypothesis Two: Middle-Layer Reuse for Model Expansion

Extending the idea of layer transfer, we further hypothesize that the middle layers of smaller models could also be reused during scaling. These layers often learn generalizable linguistic and semantic patterns that are not tied to position in the network [6]. By mapping or interpolating these layers into the center of a deeper model, it may be possible to preserve their utility in the context of a larger architecture. This hypothesis relies on the assumption that the hierarchical representation structure learned by transformers is compositional and transferable [7]. Empirical studies would be necessary to determine the degree of performance retained through such reuse, and how best to align layers between models of differing depth.

---

4. Hypothesis Three: Efficient Specialization via Medium-Sized Domain Models

A complementary hypothesis is that medium-sized models trained specifically on domain-relevant corpora can outperform or match much larger general-purpose models on in-domain tasks. For example, in the medical domain, a model trained on clinical notes, biomedical literature, and domain-specific vocabularies may learn relevant features more effectively and with fewer parameters than a general LLM trained on broad data [8]. This efficiency arises from reduced noise, stronger domain priors, and focused token distributions. While anecdotal evidence and early research suggest this could be viable [9], systematic comparisons between domain-specific medium models and general large models remain limited, and warrant further exploration.

---

5. Conclusion and Future Directions

This paper presents a set of interconnected hypotheses about how large language models might be built and trained more efficiently through progressive scaling, full-layer reuse, and domain specialization. These ideas aim to reduce resource consumption while maintaining or improving performance in both general and specialized tasks. However, these strategies are currently speculative and require empirical validation. Future work should explore experimental designs to test these hypotheses, investigate transferability limits, and identify optimal strategies for layer reuse and specialization. If confirmed, these approaches could contribute significantly to the development of more sustainable, modular, and domain-adaptable LLMs.

References

[1] T. Brown et al. Language Models are Few-Shot Learners. In *NeurIPS*, 2020.

[2] N. Houlsby et al. Parameter-Efficient Transfer Learning for NLP. arXiv:1902.00751v2, 2019.

[3] E. Bolton et al. BioMedLM: A 2.7B Parameter Language Model Trained On Biomedical Text. arXiv:2403.18421, 2024.

[4] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. What Does BERT Look at? An Analysis of BERT's Attention. In *ACL*, 2019.

[5] A. Vaswani et al. Attention is All You Need. In *NeurIPS*, 2017.

[6] B. Aken et al. How Does BERT Answer Questions? A Layer-Wise Analysis of Transformer Representations. In *CIKM*, 2019.

[7] I. Tenney et al. What do you learn from context? Probing for sentence structure in contextualized word representations. In *ICLR*, 2019.

[8] Kexin Huang, Jaan Altosaar, Rajesh Ranganath. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. In *CHIL Workshop*, 2020.

[9] K. Singhal et al. Large Language Models Encode Clinical Knowledge. *arXiv:2212.13138*, 2022.