ON THE ROLE OF WORLD REPRESENTATIONS FOR GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work has shown that pretrained neural networks develop representational structures mirroring those of the real world. However, most studies analyze only pretrained models, leaving open the question of how such structures emerge during training and how they support generalization. We address this gap using a controllable synthetic setup in which the underlying world is fixed but the data generation process can be systematically varied. With this setup, first, we show that Transformers learn the world structure in later layers and then propagate it backward. Second, we identify conditions under which the same world yields—or fails to yield—clear linear representations depending on the data generation process. Third, we probe the role of these internal representations for generalization by inducing consistent changes in the world and fine-tuning on the resulting data. Our controllable framework paves a path toward understanding the origins and functions of world representations, linking observational findings to practical insights.

1 INTRODUCTION

2 RELATED WORKS

Internal Representations. Recent work has shown that neural networks develop structured internal representations that mirror the organization of the external world (Bengio & LeCun, 2007). Studies have demonstrated that these representations emerge naturally during training, with networks learning to encode semantic relationships, hierarchical structures, and compositional patterns. Various probing techniques have been developed to analyze these representations, revealing that deeper layers tend to capture more abstract and invariant features. However, the precise mechanisms by which these representations form and how they relate to the training dynamics remain open questions. [PLACEHOLDER: Add more specific citations about representation learning, geometric structures in neural networks, and emergent properties]

Interpretability. The field of neural network interpretability has developed numerous methods to understand what models learn and how they make decisions (Hinton et al., 2006). Mechanistic interpretability approaches aim to reverse-engineer the computational structures within networks, identifying circuits and modules responsible for specific behaviors. Linear probing and other diagnostic techniques have revealed that networks often learn human-interpretable concepts in their intermediate layers. Recent advances in visualization techniques and attribution methods have provided insights into the feature detectors learned by different architectures. [PLACEHOLDER: Expand with specific interpretability methods, circuit discovery techniques, and relevant benchmarks]

Pretraining & Fine-Tuning. The pretraining-finetuning paradigm has become central to modern deep learning, particularly in natural language processing and computer vision (Goodfellow et al., 2016). Large-scale pretraining on diverse datasets enables models to learn general-purpose representations that transfer effectively to downstream tasks. Studies have shown that fine-tuning can rapidly adapt these representations while preserving much of the knowledge acquired during pretraining. The relationship between pretraining objectives, data distribution, and downstream performance remains an active area of research. Understanding how representations shift during fine-tuning and which features are preserved versus modified is crucial for improving transfer learning. [PLACE-

HOLDER: Add details about specific pretraining methods, transfer learning theory, and empirical findings about representation dynamics]

3 Setup: A Model System of Worlds

4 Results: Formation of World Representations During Pretraining

5 Results: Fine-tuning's Representational Shift Predicts Downstream Generalization

6 Discussion

References

Yoshua Bengio and Yann LeCun. Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. MIT Press, 2007.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.