Probing Layer Redundancy in Large Language Models via Persistent Homology

蔡志強

清華大學

ABSTRACT

We propose a gradient-free criterion for identifying redundant layers in large language models, based on persistent homology. For each layer, we measure how its removal affects the topological structure of sentence-level representations at the model's output. Layers that cause little change in this structure are identified as candidates for pruning. The approach requires no backpropagation, no task-specific fine-tuning, and is applicable to any transformer-based architecture. We discuss the theoretical motivation, practical design choices, and empirical behavior of large-scale language models.