From Large Language Models to Small Language Models

Abstract

Large Language Models (LLMs) have demonstrated a remarkable capacity to achieve state-of-the-art performance across a wide range of natural language tasks. However, their size requires significant computational and memory resources. To address this challenge, we investigate a practical approach to reduce their size by converting them into Small Language Models (SLMs) while keeping most of their capabilities. In this work, we investigate the use of Singular Value Decomposition (SVD) combined with sensitivity analysis to guide the process. SVD helps break down the large weight matrices into smaller low-rank parts, removing repeated or less useful information. Sensitivity analysis then tells us which layers or parts of the model are more important for accuracy. Layers that are less sensitive can be compressed more, while sensitive layers are kept with higher detail. By combining these two ideas, we can shrink LLMs into smaller models that are faster, lighter, and easier to deploy, but still perform well on language tasks.