Efficient Parameter Optimization for Compact Language Models

Youngjun Yu
Pohang University of Science and Technology (POSTECH)
Department of Computer Science and Engineering
colin31472@postech.ac.kr

Abstract

This project aims to develop a high-performance, compact language model optimized for mobile and low-resource devices by reducing parameters to under one billion without sacrificing effectiveness. Inspired by MobileLLM, the research explores various activation functions, such as ReLU variants and GLU-based methods, and investigates the trade-offs between model depth and width to maximize parameter efficiency. Additionally, techniques like embedding and layer sharing will be evaluated to further enhance performance. The ultimate goal is to enable real-time applications, such as translation and voice assistants, on mobile devices by eliminating cloud-dependence and latency, thereby improving accessibility while minimizing energy consumption and environmental impact.