Pipeline for efficient quantization of large language models for resource-constrained deployment

2025 IEEE AFRICON, Polokwane, South Africa, 10-12 December 2025

Chiwewe, Tapiwa M
Council for Scientific and Industrial Research (CSIR)
Meiring Naude Drive, Pretoria, 0184
Email: Tchiwewe@csir.co.za

Recent years have seen breakthroughs in large language models (LLMs), such as the GPT and LLaMA family of models that have transformed natural language processing. Despite this, their considerable computational and memory requirements inhibit their deployment in edge and mobile environments. In this paper we introduce a modular quantization pipeline that reduces the memory footprint of LLMs while preserving core performance. We evaluate the basic and advanced quantization techniques, including Absolute Maximum Quantization, Zero-Point Quantization, GPTQ, and NF4, and make use of popular toolkits that include bitsandbytes and AutoGPTQ. Experimental results on representative tasks show that 4- and 8- bit quantized models can be run on commodity GPUs and CPUs with acceptable quality loss. Our experiments demonstrated up to 8x compression, making our pipeline suitable for LLM deployment in both edge and industrial scenarios.