1 Bit LLM for Hindi Language at Lower Memory Consumption

Abhi Akshat
Department of Computer Science and Engineering, Sharda University, Greater Noida, Uttar Pradesh, India

Kritika Tripathi
Department of Computer Science and Engineering, Sharda University, Greater Noida, Uttar Pradesh, India

Devanshi Malik
Department of Computer Science and Engineering, Sharda University, Greater Noida, Uttar Pradesh, India

Kusum Lata
Department of Computer Science and Engineering, Sharda University, Greater Noida, Uttar Pradesh, India

Abstract

With the ever-increasing size of Large Language models, an approach to reduce the memory consumption has become vital. Techniques like parameter scaling and instruction tuning are commonly used to improve LLM performance. Although effective, these methods greatly increase memory consumption and computational complexity, requiring more significant resources for model training and inference. Model compression methods like quantisation are becoming more popular as a solution to these problems. One such method is Bitnet, specifically Bitnet-1.58, here we quantize the weights in such a way that it takes one of the tree values, 0,1 or -1. Since there is little research on quantization of Hindi LLMs, this paper explores the Bitnet-1.58 architecture for Hindi LLM based on the LLaMA architecture. A 100 million and a 1 Billion parameter model was trained on a 2 billion token dataset. A perplexity of almost 42 and 12 was achieved respectively. Although the performance was not at par with the current state of the art model, but given the limited resources, the performance achieved was good. The model remembered the training dataset sufficiently well, although the performance in general NLP tasks was suboptimal. Given enough computational resources and large datasets, the models could perform at par with full precision models. Training was conducted on two NVIDIA A100 GPUs.

Keywords

Large Language Models, LLMs, NLP, Model Compression, Model Quantization, Bitnet, Hindi LLMs.