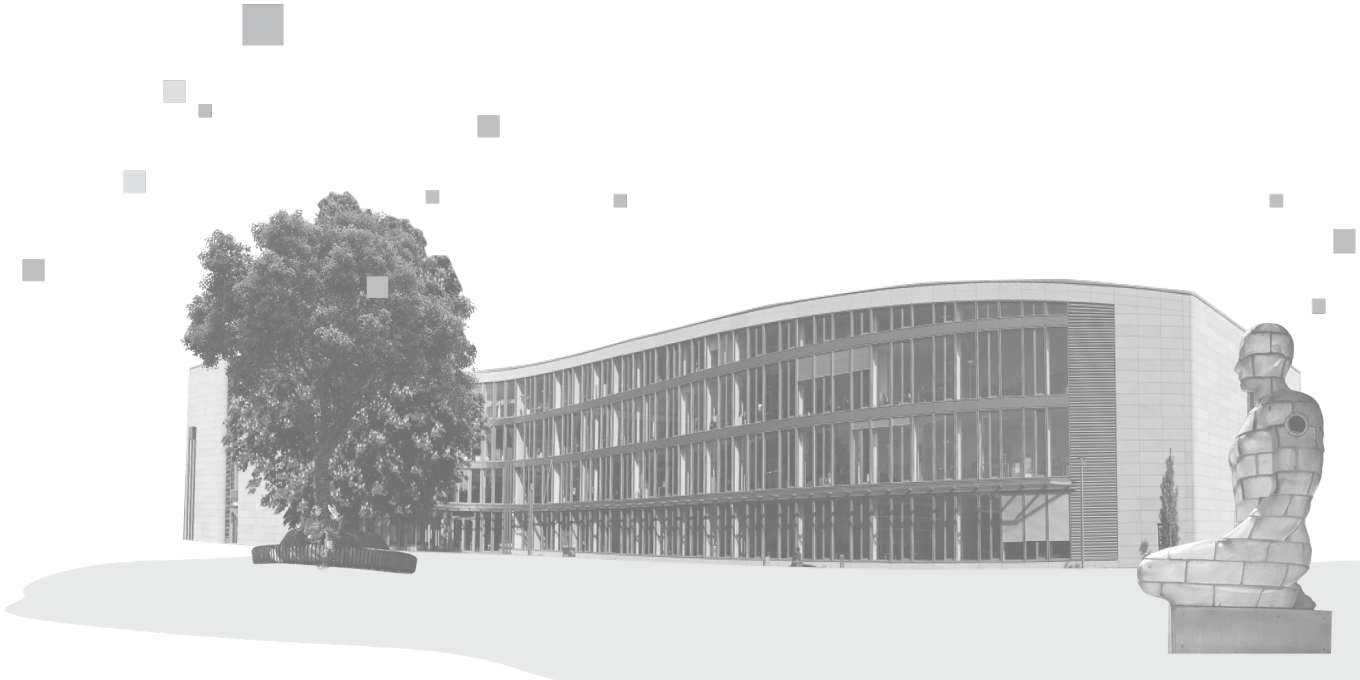
Open-Source Tools for Local LLM

PD Dr. Haojin Yang,

Multimedia and Machine Learning Group,

Hasso Plattner Institute



Design IT.
Create Knowledge.

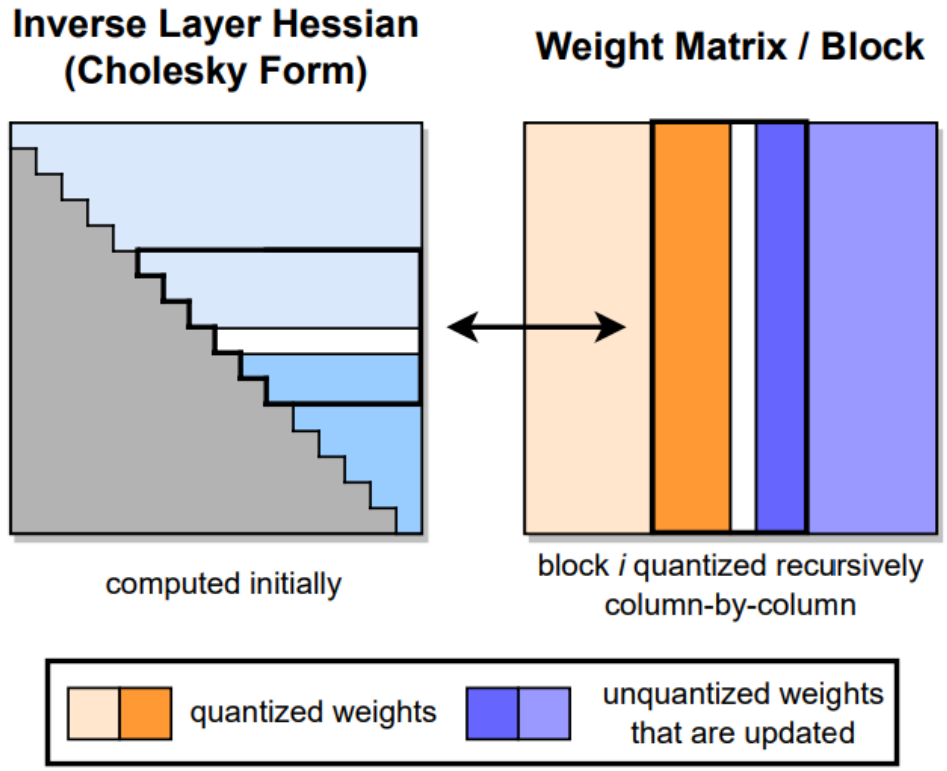
www.hpi.de

Introduction

- Large Language Models (LLMs) require significant computational resources
- Quantization reduces model size while maintaining performance
- We'll explore three quantization approaches: GPTQ, GGUF, and BitNet
- GPTQ compresses models from 32-bit to 4-bit precision (8x smaller)
- GGUF enables flexible quantization for limited hardware
- BitNet represents the future with extreme 1-bit quantization

GPTQ - Making Models Smaller with Smart Compression

- GPTQ compresses models from 32-bit to 4-bit precision (8x smaller)
- Uses layer-wise quantization with inverse-Hessian matrices
- Preserves accuracy by intelligently redistributing quantization errors
- Designed for running full models on GPU
- Maintains model performance while significantly reducing size
- Enables faster inference and lower memory requirements



Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2022). Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

GGUF - Flexible Quantization for Limited Hardware

- GGUF (GPT-Generated Unified Format) allows offloading layers to CPU when VRAM is limited
- Uses block-wise quantization with 'super' blocks and 'sub' blocks
- Supports various quantization levels (2-bit, 4-bit, 6-bit)
- Creates a hierarchical quantization scheme
- Ideal for running models on hardware with limited GPU memory
- Provides flexible balance between model size and performance

BitNet - The Future of 1-bit Models

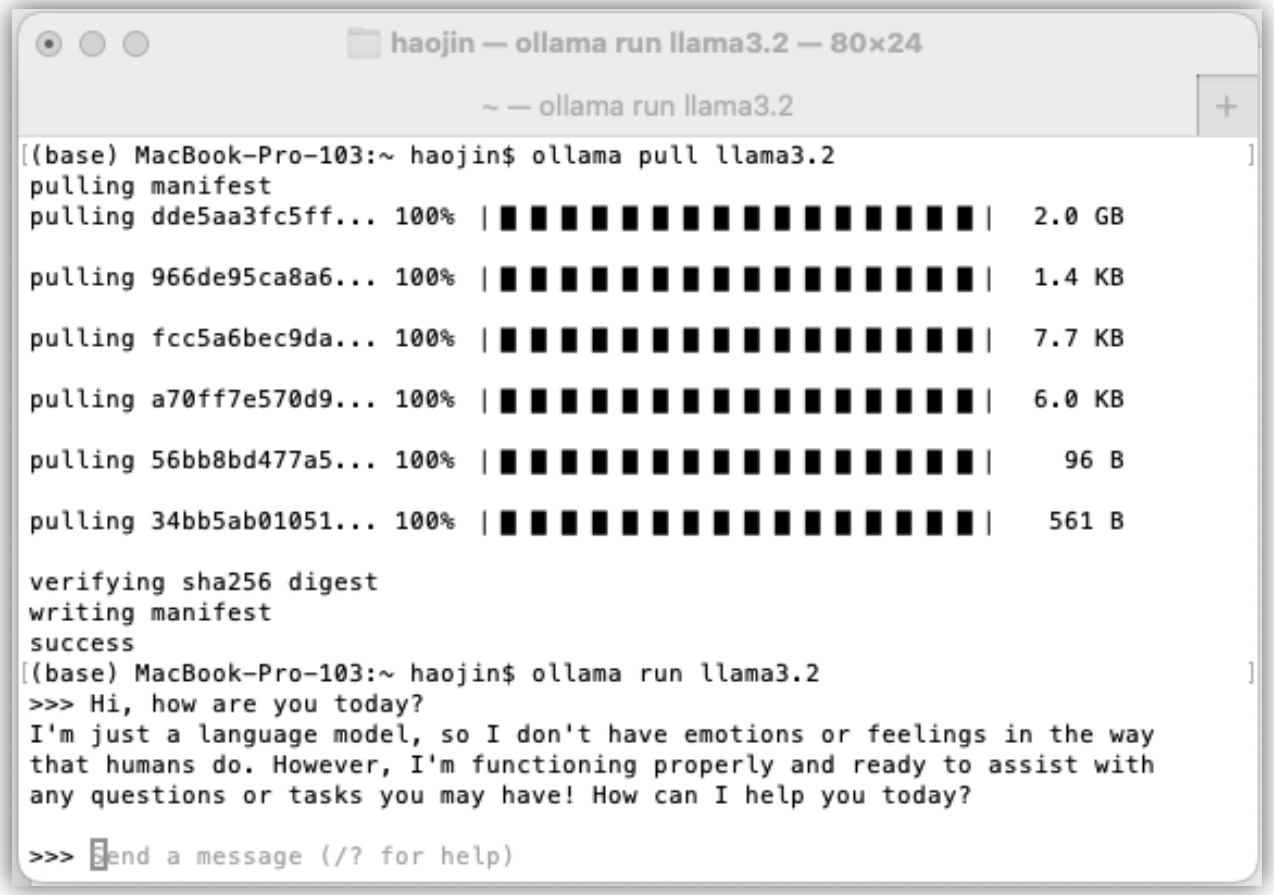
- BitNet represents weights using only 1-bit (-1 or 1)
- BitNet 1.58 adds a third value (0) for better performance
- Replaces standard linear layers with 'BitLinear' layers
- Eliminates multiplication operations, using only addition and subtraction
- Performance gap decreases as models get larger (>30B parameters)
- Offers extreme model compression with manageable performance tradeoffs

microsoft/**BitNet**

Official inference framework for 1-bit LLMs

Ollama - Running LLMs Locally Made Easy

- Ollama is a tool for running LLMs locally in a plug-and-play manner
- Based on GGUF format for efficient local deployment
- Simple commands to install and run models
- Supports web-based UI for easy interaction
- Makes powerful AI accessible without specialized hardware
- Integrates with various applications and development workflows



Summary

- GPTQ: 4-bit precision, best for full GPU deployment
- GGUF: Variable precision, best for mixed CPU/GPU deployment
- BitNet: 1-bit precision, most efficient but with some performance tradeoff
- Ollama: Makes local deployment easy using GGUF format