Local LLM Deployment

Dylan Gresham for CyberAI Club

llama.cpp

- llama.cpp GitHub
- LLM inference library in C/C++
  - Only for Meta LLaMA models

“The main goal of `llama.cpp` is to enable LLM inference with minimal setup and state-of-the-art performance on a wide variety of hardware - locally and in the cloud.”

- Provides support for running CPU only or with GPU(s) as well
  - Also automatically does CPU+GPU hybrid inference
- Provides quantization capabilities
- List of supported models on their GitHub

Quantization

- Transformations of a model's parameters from one precision to another
  - I.e., from 16-bit floating point to 8-bit floating point weights
- Goal is to reduce the size of the model in memory while still maintaining appropriate performance

$$(8.03 \times 10^9) \times 16 = 1.2848 \times 10^{11} \text{ bits} \approx 14.96 \text{ GB}$$

$$(8.03 \times 10^9) \times 8 = 6.424 \times 10^{10} \approx 7.48 \text{ GB}$$

$$(8.03 \times 10^9) \times 2 = 1.606 \times 10^{10} \approx 1.87 \text{ GB}$$

Question: What is shown in this image?

LLaVA-Next-8B 16bit: *The image shows a group of individuals engaged in what appears to be a construction or excavation activity. One person is actively digging or shoveling into a hole, while the others are standing around, observing or waiting for their turn. The setting looks like a construction site or a similar outdoor work environment. The text "Big companies" is superimposed on the image, suggesting a commentary on the scale or nature of the work being done, possibly implying that it is a large-scale operation.*

LLaVA-Next-8B AWQ-4bit: *The image shows a group of individuals engaged in what appears to be a construction or excavation activity. One person is actively digging or shoveling into a hole, while the others are standing around, observing or waiting for their turn. The setting looks like a construction site or a similar outdoor work environment. The text "Big companies" suggests a commentary on the scale or impact of large corporations, possibly in the context of the work being done or the conditions of the site.*

LLaVA-Next-8B AWQ-3bit: *The image shows a group of people gathered around a large hole in the ground, which appears to be a construction site. One person is actively engaged in digging, while the others are observing the work. The text "Big companies" suggests a commentary on the scale of the hole in comparison to the people present, emphasizing the size of the excavation. The setting looks like a construction site, possibly for a building or infrastructure development.*

LLaVA-Next-8B AWQ-2bit: *imers imers imers imers imers imers imersimersimersimers imersimers*

Fig. 2: The VQA results of LLaVA-Next-8B under different bit-width (1/5)

Live Demo

- Using llama-cpp-python for CPU and GPU inference
- PyTorch short proof-of-concept for CPU only inference
    - Utilizes the Transformers library by HuggingFace for model loading and tokenizing

References & Links

- Huang, W., Ma, X., Qin, H., Zheng, X., Lv, C., Chen, H., ... & Magno, M. (2024). *How good are low-bit quantized llama3 models? an empirical study*. arXiv preprint arXiv:2404.14047.
- PyTorch
- llama.cpp
- llama-cpp-python
- HuggingFace
- HuggingFace Transformers
- GitHub repository for all code, notes, and slides used in this presentation