Quantization of Large Language Model

Duration: 01 days (08 hours)

Generative AI models, like large language models, often exceed the capabilities of consumer-grade hardware and are expensive to run. Compressing models through methods such as quantization makes them more efficient, faster, and accessible. This allows them to run on a wide variety of devices, including smartphones, personal computers, and edge devices, and minimizes performance degradation.

The following topics will be covered:

- Quantize any open source model with linear quantization using the Quanto library.
- Get an overview of how linear quantization is implemented. This form of quantization can be applied to compress any model, including LLMs, vision models, etc.
- Apply "downcasting," another form of quantization, with the Transformers library, which enables you to load models in about half their normal size in the BFloat16 data type.
- Build and customize linear quantization functions, choosing between two "modes": asymmetric and symmetric; and three granularities: per-tensor, per-channel, and per-group quantization.
- Measure the quantization error of each of these options as you balance the performance and space tradeoffs for each option.
- Build your own quantizer in PyTorch, to quantize any open source model's dense layers from 32 bits to 8 bits.
- Go beyond 8 bits, and pack four 2-bit weights into one 8-bit integer.