Model Quantization

Deploying models that are performant (obviously statistically but in this context I primarily mean ***computationally***) is challenging when you are working with large models such as BERT etc.

Quantization involves compressing model weights into smaller, more efficient representations. Weights are normally stored as 32 bit floating point numbers but they can be compressed into 8 bit integers with a very small amount of performance loss.

This article talks about how to do quantization effectively (mirror).

Quantization with Optimum and OpenVino

Openvino is an open source framework from Intel that provides quantization and x86 CPU support for torch and huggingface transformers.

Revision #6
Created 21 November 2022 13:36:45 by James
Updated 22 February 2024 09:48:42 by James