Yao Yao Wang Quantization

GPTQ Deep Dive: The Hessian Matrix Method 87fbb90029 Why a 70B model needs 140GB 87fbb90029 Ternary Frontier: BitNet b1.58 (1.58-Bit Weights) 87fbb90029 Intro 87fbb90029 VQ-VLA: Improving Vision-Language-Action Models via Scaling Vector-Quantized Action Tokenizers - VQ-VLA: Improving Vision-Language-Action Models via Scaling Vector-Quantized Action Tokenizers 3 minutes, 44 seconds - [ICCV 2025] VQ-VLA is an innovative vector **quantization**, based action tokenizer built upon the largest-scale action trajectory ... 87fbb90029 Type 1: Post-Training Quantization (PTQ) 87fbb90029 Keyboard shortcuts 87fbb90029 What 87fbb90029 □□□ - Wang Yi Liu Yao Yao (Remix) - □□□ - Wang Yi Liu Yao Yao (Remix) 5 minutes, 2 seconds - Subscribe for more videos! □ My favorite collections of songs: Chinese, English, Japanese, Korean, and Vietnamese (these will ... 87fbb90029 Spherical Videos 87fbb90029 The trade-off: smaller but still useful 87fbb90029 Calculate Memory for Model 87fbb90029 The 8-Bit Fallacy: Half Memory, Zero Speedup 87fbb90029 When 87fbb90029 Choosing the Right Quant 87fbb90029 Where LLMs actually live 87fbb90029 What is GGUF? 87fbb90029 Round to Nearest (RTN): The Simplest Quantization Baseline 87fbb90029 Quantization Explained 87fbb90029 Fixed point arithmetic 87fbb90029 Who should quantize a model? 87fbb90029 The Verdict: The 4-Bit Pareto Sweet Spot 87fbb90029 Understanding Double Quantization for LLMs - Understanding Double Quantization for LLMs 4 minutes, 24 seconds - link to full course: https://www.udemy.com/course/fine-tune-deploy-llms-with-qlora-on-sagemaker-streamlit/? 87fbb90029 Outro 87fbb90029 Why Quantize? The VRAM \u0026 Accuracy Paradox 87fbb90029 What even are these files? 87fbb90029 Quantization Explained: Run Bigger LLMs on Smaller Hardware - Quantization Explained: Run Bigger LLMs on Smaller Hardware 6 minutes, 42 seconds - Quantization, is how you fit a model that \"shouldn't\" run on your board onto your board — and it's simpler than the jargon makes it ... 87fbb90029 What is a quant (or Quantization) 87fbb90029 Type 2: Quantization-Aware Training (QAT) 87fbb90029 Quantization Explained: How to Run Large AI Models on Small Devices - Quantization Explained: How to Run Large AI Models on Small Devices 4 minutes, 5 seconds - Ever wondered how massive Large Language Models (LLMs) can run on your laptop or phone? The secret is **Quantization**,! 87fbb90029 General 87fbb90029 LLM Quantization Explained - LLM Quantization Explained 4 minutes, 18 seconds - LLM **quantization**, is how a 70B model that needs 140GB of memory gets small enough to run on a normal GPU. Every model you ... 87fbb90029 Subtitles and closed captions 87fbb90029 Data Types: The Building Blocks 87fbb90029 Yao Wang \"Building multi-subject neural speech decoders from intracranial EEG signals\" - Yao Wang \"Building multi-subject neural speech decoders from intracranial EEG signals\" 36 minutes - All right So our next speaker is **Yao Wang**, from NYU She has been doing some really cool work in decoding speech from uh ... 87fbb90029 The 3-Bit Catch: When Sophisticated Math Fails 87fbb90029 PTQ vs QAT: Post-Training vs Quantization-Aware Training 87fbb90029 Pushing the Limits 87fbb90029 AI Model Quantization: The Complete Guide — FP32 to Q4_K_M - AI Model Quantization: The Complete Guide — FP32 to Q4_K_M 4 minutes, 47 seconds - Everything about **quantization**, for local AI inference. FP32, FP16, INT8, INT4, GGUF, GPTQ, AWQ explained with real benchmarks. 87fbb90029 GGUF \u0026 K-Quants: The Format That Saved Local AI 87fbb90029 □ LLM Quantization Explained: FP32, FP16, INT8, INT4, GPTQ, AWQ \u0026 GGUF - □ LLM Quantization Explained: FP32, FP16, INT8, INT4, GPTQ, AWQ \u0026 GGUF 20 minutes - LLM **Quantization**, Explained: FP32, FP16, INT8, INT4, GPTQ, AWQ \u0026 GGUF What exactly is LLM **quantization**,, and why is it so ... 87fbb90029 3 Ways Quantization is done 87fbb90029 Matrix multiplications 87fbb90029 360-Degree Video Streaming (Yao Wang) - 360-Degree Video Streaming (Yao Wang) 54 minutes - ACM MMSys 2020 keynotes. 87fbb90029 How memory is used 87fbb90029 How LLMs survive in low precision | Quantization Fundamentals - How LLMs survive in low precision | Quantization Fundamentals 20 minutes - In this video, we discuss the fundamentals of model **quantization**,, the technique that allows us to run inference on massive LLMs ... 87fbb90029 The Cliff: Why Models Collapse Below 4-Bits 87fbb90029 Calculate the KV Cache size (Context window memory) 87fbb90029 Chien-Yao Wang: From real-time object detectorto real-time generalist model #ICBS2026 - Chien-Yao Wang: From real-time object detectorto real-time generalist model #ICBS2026 1 hour, 1 minute - ... in method such as nnf vector **quantization**, or PCA the model is often described by a simple equation the learned presentation is ... 87fbb90029 How Do We Get MASSIVE Model To Run On Device? Quantization Explained. - How Do We Get MASSIVE Model To Run On Device? Quantization Explained. 26 minutes - Every time I do a video about a model I get a comment saying \"Well you never said what it takes to run it!\" Well since I am not ... 87fbb90029 AWQ: Activation-Aware Weight Quantization (Salient Weights) 87fbb90029 The Trade-Off 87fbb90029 Search filters 87fbb90029 Who is making these files? 87fbb90029 How It Works: Mapping 87fbb90029 Outro 87fbb90029 Crash Course is over 87fbb90029 What is Quantization? 87fbb90029 GGUF vs AWQ vs GPTQ: LLM Quantization Methods Explained - GGUF vs AWQ vs GPTQ: LLM Quantization Methods Explained 9 minutes, 48 seconds - Quantization, is often sold as \"smaller = cheaper = better.\" But the three things people conflate—VRAM savings, inference speed, ... 87fbb90029 What is Quantization? 87fbb90029 AWQ for LLM Quantization - AWQ for LLM Quantization 20 minutes - Large language models (LLMs) have shown excellent performance on various tasks, but the astronomical model size raises the ... 87fbb90029 Why Do We Need It? 87fbb90029 Looking at a real example 87fbb90029 All about Quants and How They Work] 87fbb90029 About Perplexity 87fbb90029 Summary: Key Takeaways 87fbb90029 How 87fbb90029 Playback 87fbb90029 Quantized LLM Training at Scale with ZeRO++ // Guanhua Wang // AI in Production 2025 - Quantized LLM Training at Scale with ZeRO++ // Guanhua Wang // AI in Production 2025 24 minutes - Huge shout out to our sponsors @rafaysystems7900, @humanloop8511, arcade.dev, @filopedraz, @Deepgram. //Abstract ... 87fbb90029