AYI NEDJIMI Consultants

AWQ Quantization : Optimiser les LLM en INT4 sans perte

8 mai 2026 • Mis à jour le 17 mai 2026 • 25 min de lecture • 5098 mots • 89 vues •

AWQ (Activation-aware Weight Quantization) est devenue la technique de référence pour compresser les LLM en INT4 sans perte de qualité. Guide complet : algorithme, comparatif GPTQ/SmoothQuant/SpQR, implémentation AutoAWQ, déploiement vLLM/TensorRT-LLM, benchmarks Llama 3.1 70B, Mixtral, Qwen 2.5, DeepSeek-V3 et workflow pratique pour quantifier un modèle frontier sur un seul GPU H100.

L'**AWQ quantization** (Activation-aware Weight Quantization) s'impose en 2026 comme la technique de référence pour compresser les grands modèles de langage en **INT4** sans dégradation de la qualité. Conçue par les équipes du MIT Han Lab, A

intuition simple mais redoutablement efficace : toutes les pondérations d'un LLM ne sont pas égales devant l'inférence. En identifiant les canaux d'activation saillants (les fameux 1% de poids qui portent l'essentiel de la précision), AWQ préserve sélectivement leur dynamique tout en quantifiant agressivement le reste. Le résultat est spectaculaire : un Llama 3.1 70B passe de 140 Go en FP16 à 35 Go en INT4, tient sur un seul GPU H100 80 Go, et conserve plus de 99% de la perplexité d'origine sur WikiText-2. Pour les RSSI, architectes IA et développeurs cherchant à déployer des LLM *on-premise* à coût contenu, AWQ change la donne. Cet article démonte le mécanisme algorithmique, compare les alternatives (GPTQ, SmoothQuant, SpQR), détaille le workflow de quantization avec llm-awq et AutoAWQ, et expose les benchmarks réels mesurés sur Llama 3.x, Mistral, Mixtral, Qwen 2.5 et DeepSeek-V3.

À RETENIR

Points clés à retenir

Compression 4x sans perte significative : AWQ quantifie les poids d'un LLM de FP16 à INT4 en préservant les canaux saillants identifiés via les statistiques d'activation, avec une perte de perplexité typique inférieure à 0,5%.

Hardware-friendly : contrairement aux schémas mixed-precision complexes, AWQ produit des poids INT4 uniformes parfaitement compatibles avec les kernels GPU (vLLM, TensorRT-LLM, ExLlamaV2) et les Tensor Cores INT4.

Réponse sous 24h

Devis gratuit →