Literature Review: Accurate Post-Training Quantization For Generative Pre-Trained Transformers

Ludek Cizinsky | 377297 | ludek.cizinsky@epfl.ch
MLP

1 Summary

The landscape of large language models (LLMs) is dominated by the trend towards increasingly larger sizes, with current models ranging from 1 billion to 540 billion parameters, exemplified by PALM (Chowdhery et al., 2022). This approach presents several challenges: only a select few organizations can manage the costs of deploying such large models, the inference costs are high, and the environmental impact is substantial.

These factors drive research into methods for compressing these large models. The paper under review, *GPTQ: Accurate Post-Training Quantization for Generative Pre-Trained Transformers*, represents a significant advancement in LLM quantization that is supported natively by Hugging Face which makes easily integrable as part of the LLM finetuning pipeline.

Quantization techniques are primarily categorized into two types: Quantization Aware Training (QAT) and Post-training Quantization (PTQ). QAT often necessitates expensive retraining for models that, even post-quantization, possess billions of parameters and are trained on vast datasets comprising trillions of tokens. Although there are numerous efforts in PTQ, these typically struggle to achieve high bit rate compression (e.g., 4 bits) without a substantial loss in performance (Frantar et al., 2023a).

GPTQ addresses this issue by providing an efficient post-training quantization framework capable of compressing a 175-billion-parameter model into 4 bits without significant performance loss, all within four hours. This is a notable improvement over existing state-of-the-art (SOTA) methods that may take weeks to achieve similar results (Frantar et al., 2023a). Additionally, GPTQ is size agnostic, functioning effectively across both large and smaller models, and significantly speeds up inference times compare to full-precision models.

GPTQ achieves these results through layer-wise quantization, where each layer is represented as a matrix. Given a layer W and input X, the objective is to find a quantized layer $\hat{W}$ that minimizes the mean squared error (MSE) between the full-precision and quantized outputs. This approach was previously attempted in the Optimal Brain Quantization (Frantar et al., 2023b), but it scaled poorly with larger models. GPTQ builds on this foundational method and introduces several key enhancements: (1) an improved layer-wise quantization algorithm that increases efficiency, (2) optimized memory usage through lazy batch updates, and (3) robust handling of numerical errors, which become more probable as model size increases. For full details, please refer to the original paper (Frantar et al., 2023a).

Despite the contributions, GPTQ has its own limitations. It quantizes only weights, not activations, and is restricted to generative transformer-based models. Moreover, while it accelerates the movement of weights in GPU memory, the computational costs during inference remain unchanged as GPTQ dequantizes the weights back to their original precision at the inference stage.

In my personal opinion, GPTQ presents a significant advancement in LLM quantization, offering a practical solution for compressing large models without sacrificing performance proven across a range of model size and benchmark datasets - significant improvement compare to the current SOTA methods. In addition, I find important that the method is accessible via Hugging Face's Transformers library, making it easily integrable into existing workflows.

2 Strengths

Improving Baseline. GPTQ demonstrates superior performance across all model sizes in terms of perplexity over the baseline RTN method which is a simple technique that rounds weights to the nearest n-bit integer. This is functionally equivalent to the state-of-the-art LLM.int8() method (Dettmers et al., 2022). GPTQ notably excels in 3-bit compression, where RTN shows significant performance deterioration. This pattern is consistent across models ranging from 125M to 175B parameters. For 4-bit precision, the improvement in perplexity over RTN varies, with differences ranging from marginal (approximately 1 point) to substantial (up to 33 points), depending on the model size. Additionally, the authors have verified that GPTQ also achieves similar results on standard NLP benchmarks including LAMBADA (Paperno et al., 2016), PIQA (Bisk et al., 2020), and ARC (Clark et al., 2018), further validating its efficacy.

Model Size Agnosticism. GPTQ is effective across a broad spectrum of model sizes, showing competitive results even on smaller models, where it rivals methods specifically designed for such sizes. This futher demonstrates GPTQ's versatility and broad applicability which is crucial in today's landscape of LLMs with various sizes.

Significant Runtime Improvement. GPTQ can quantize models with billions of parameters in minutes on NVIDIA A100 GPUs. This is in stark contrast to reference PTQ method requiring up to three hours for similar tasks (Yao et al., 2022). Even for the largest models, such as BLOOM with 176B parameters (BigScience Workshop, 2022), quantization completes in approximately four hours.

Single GPU Capability for (Super) Large Models. GPTQ significantly reduces the hardware demands for running super large models such as OPT-175B (Zhang et al., 2022). For instance, using 3-bit quantization, the OPT-175B model compresses to 63 GB, fitting within the 80 GB memory of a single NVIDIA A100 GPU. This is a significant reduction from the five GPUs required for the standard 16-bit model and three for LLM.int8() (Frantar et al., 2023a).

Inference Speedup. A critical metric for industry practitioners is inference speed. GPTQ addresses this by developing a GPU kernel optimized for quantized-matrix and full-precision-vector products, significantly accelerating the decoding process. For instance, a 3-bit quantized OPT-175B model achieves a 3.24x speedup in average per-token latency compared to its full-precision counterpart on NVIDIA A100 GPUs, with even greater improvements (4.53x) observed on NVIDIA A6000 GPUs (Frantar et al., 2023a).

Paper Structure and Clarity. The paper effectively outlines the current SOTA in LLM quantization and the challenges faced by existing methods. Despite the relative complexity of GPTQ, the authors offer a clear and concise explanation of the method, which they organize into three main steps. Crucially, the authors include several ablation studies that demonstrate the anticipated improvements in quantization runtime, inference speed, and overall model performance.

3 Weaknesses

Baseline Runtime. Although the paper demonstrates that GPTQ outperforms the baseline method (RTN) across all model sizes and on various NLP benchmarks, including perplexity and accuracy on LAMBADA, ARC, and PIQA datasets, it lacks specific data on the total quantization process runtime for the baseline, which reportedly *scales well to billions of parameters* (Frantar et al., 2023a). Therefore, for the sake of completness, it would be beneficial to provide concrete measures on the baseline runtime.

Choice of LLMs. The study references two LLMs, OPT and BLOOM, chosen for being among the largest available models at the time. However, the results show a consistently smaller gap in perplexity between GPTQ and RTN for BLOOM compared to OPT. The only rationale provided is that BLOOM models are easier to quantize, suggesting that the choice of model impacts the performance difference. Thus, the paper would benefit from a more comprehensive analysis of the models to understand the impact of model choice on the quantization process of GPTQ.

Transformers Only. Today, majority of the LLMs are based on the Transformer architecture. Therefore, it is reasonable to focus on this type of models. However, for the sake of generality of the method, it would be beneficial if GPTQ was not only scale agnostic but also model agnostic.

Affordability of A100 GPUs. The practical speedups section indicates that a 3-bit precision quantized 175B-parameter model can run on a single A100 GPU. However, given the high cost of these GPUs (approximately $15K USD), their affordability remains a barrier for many researchers. While the paper advances the accessibility of large models, further improvements are necessary to make them more universally accessible.

References

BigScience Workshop. 2022. BLOOM (revision 4ab0472).

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm.int8(): 8-bit matrix multiplication for transformers at scale.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023a. Gptq: Accurate post-training quantization for generative pre-trained transformers.

Elias Frantar, Sidak Pal Singh, and Dan Alistarh. 2023b. Optimal brain compression: A framework for accurate post-training quantization and pruning.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernandez. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. 2022. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.