CNN-GPTQ: 4-bit Quantization for CNN Architectures

Saathvik Dirisala
sdirisala@ucsd.edu

Rayan Saab
rsaab@ucsd.edu

Alex Cloninger
acloninger@ucsd.edu

Abstract

CNNs are a type of Deep Neural Network that, like ANNs, have fully-connected layers but are specifically designed to process image data represented as matrices. To capture valuable information from images, CNNs use convolutional layers to preprocess the input, which is then passed into the fully-connected layers. This feature-extraction process involves learning weights for filters within the convolutional layers. Since images are often 3-dimensional, and some CNNs are particularly deep, these convolutional operations can demand substantial computational power, posing challenges for real-time systems with limited RAM and processing capacity. With advancements in quantization for large language models (LLMs), algorithms have emerged that efficiently reduce neural network precision from 32-bit to as low as 4-bit, retaining most of the model's capabilities while enabling inference on devices with restricted GPU access. One promising approach for compressing deep CNNs is GPTQ, a 4-bit quantization method for LLMs that incorporates weight pruning from the Optimal Brain Surgeon algorithm . Because GPTQ redistributes losses at each layer independently, it can be adapted effectively to compress CNNs as well.

Code: https://github.com/saathvikpd/DSC180AB_Capstone

1 Introduction

In the past decade, neural networks and the efficient leveraging of computation resources have been thoroughly researched to such an extent that model sizes have exponentially grown. LLMs, being the largest in the DNN family, have billions and trillions of weights. These massive models require a lot of compute power even for inference, necessitating techniques that help compress them with minimal loss in general knowledge and instruction-based task performance. And one of many techniques that has gained traction is quantization

Quantization is a model compression technique that reduces precision in neural network parameters, typically from 32-bit floating points to lower-bit integers, such as 8-bit or even 4-bit. It has two primary approaches: Quantization-Aware Training (QAT) and Post-Training Quantization (PTQ), each with distinct advantages and trade-offs.

Quantization-Aware Training (QAT) integrates quantization during training by simulating lower-precision operations, optimizing for both weight values and activation distributions to minimize loss. QAT offers high accuracy retention, especially for sensitive tasks, but it is computationally expensive due to its iterative adjustments. With the advent of LLMs, model sizes have been skyrocketing, and training of these models requires a massive amount of GPU compute power and memory—both resources that are very expensive.

Post-Training Quantization (PTQ), in contrast, applies quantization after a model is fully trained, often using minimal calibration data to adjust activations. PTQ is efficient, requiring significantly less computational power since it avoids retraining. However, it may lead to accuracy degradation, particularly in deep or highly nonlinear networks, as it lacks layer-wise optimization.

Optimal Brain Surgeon (OBS) lays the foundation for modern pruning techniques, using second-order derivatives to assess the impact of pruning each weight on model loss. Originally developed for simpler networks, OBS's influence extends to CNNs and transformers, informing layer-by-layer pruning to optimize size and accuracy. In CNNs, where spatial and hierarchical patterns play a key role, OBS-derived methods could allow for targeted pruning across layers while maintaining accuracy in critical feature-detection regions. This strategy has been beneficial for CNN deployment on hardware-constrained platforms, preserving performance while enabling efficient scaling for large datasets and intricate tasks like object detection (Hassibi and Stork 1992).

The **Optimal Brain Quantizer (OBQ)** is a post-training quantization framework that unifies quantization and pruning, extending the Optimal Brain Surgeon (OBS) framework to handle modern deep learning models efficiently. OBQ minimizes accuracy loss by applying the Optimal Brain Quantizer (OBQ) approach, a per-layer quantization and pruning method that accounts for second-order information. Rather than pruning weights, OBQ aims to quantize weights. The losses created by both, pruning and quantization, are treated the same. For CNN models, which often have many parameters in early layers, OBQ's layer-wise approach could help prioritize critical layers while simplifying less impactful ones. This targeted approach may allow CNNs to maintain visual task accuracy while achieving

aggressive compression, potentially offering substantial computational savings, especially in high-resolution image processing tasks and real-time applications such as video analysis (Frantar, Pal Singh and Alistarh 2022).

The **GPTQ** approach, built on the OBQ algorithm, is designed for transformer models, quantizing LLMs to 3 or 4 bits with minimal accuracy loss and significant speed improvements. This method relies on second-order information to optimize compression by selecting weights that yield minimal loss. While OBQ defines a specific order to quantize weights—prioritizes the weights with lower impact on loss—GPTQ recognizes that the algorithm is still effective when weights are quantized in a random order. Furthermore, GPTQ also proposes a highly parallelized approach that minimizes redundant computations and updates multiple neurons in a single iteration (Frantar et al. 2022).

While GPTQ focuses on transformers, the reliance on second-order derivatives and layer-wise quantization strategies could inspire similar adaptations for CNNs, particularly in reducing resource-intensive computations and enabling efficient deployment on low-power devices. For CNNs, GPTQ might need re-optimization to account for convolutional layers' unique spatial and hierarchical characteristics, but it holds promise for significant memory and speed gains in CNN applications on edge devices and embedded systems.

2 Methods

3 Results

4 Discussion

5 Conclusion

References

Frantar, Elias, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. "GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers." In *arXiv preprint*. [Link]

Frantar, Elias, Sidak Pal Singh, and Dan Alistarh. 2022. "Optimal Brain Compression: A Framework for Accurate Post-Training Quantization and Pruning." In *Advances in Neural Information Processing Systems*. [Link]

Hassibi, Babak, and David G. Stork. 1992. "Second order derivatives for network pruning: Optimal Brain Surgeon." In *Advances in neural information processing systems*.