Exploration of Activation-aware Weight Quantization for LLM Compression and Acceleration

Jerry Li
MIT EECS
jerryli@mit.edu

Ronald Xu
MIT EECS
ronaldxu@mit.edu

Abstract

In this work, we examine strategies for effective compression of Large Language Models (LLMs), which, despite their excellent abilities, require significant computational resources and energy to run. In order to democratize this Artificial Intelligence, we investigate 4-bit weight and 4-bit activation quantization (W4A4), an extreme level of quantization which previous works have struggled to achieve. Utilizing inspirations from previous works such as Activation-aware Weight Quantization (AWQ) and SmoothQuant, we explore two potential methods for achieving W4A4 quantization: group quantization with AWQ as well as a mixed-precision approach. We evaluate the effectiveness of these methods by testing quantized versions of Facebook's OPT models on the WikiText-2 dataset and discuss their trade-offs. The results show that group quantization with AWQ narrows the perplexity gap between the FP16 and W4A4 models for small group sizes, while the mixed-precision method struggles to recover perplexity. Finally, we suggest some possible avenues of future research.

1 Introduction

Recently, Artificial Intelligence (AI) in the form of Large Language Models (LLMs) such as LLaMA[4] and OPT[7] have been revolutionary in their comprehension and contextual understanding abilities. However, we see in the following figure that the size of these language models is as great as their knowledge.

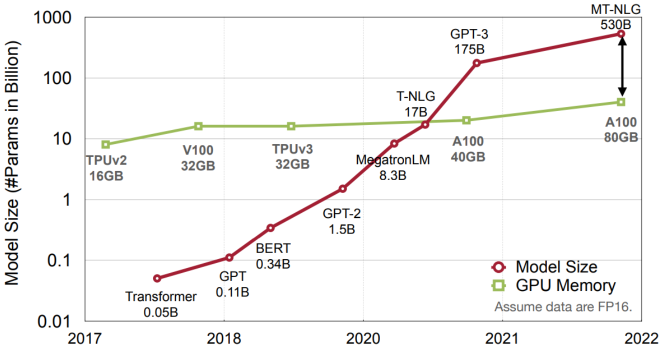

Figure 1: LLMs' size is increasing faster than the capabilities of hardware. Graph adapted from SmoothQuant[5].

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Therefore, to unlock the power of LLMs, one needs significant computing resources and electricity. In order to democratize AI, it is imperative for model compression and optimization strategies to be developed, enabling LLMs to run efficiently on lower-powered and less resource-intensive machines.

AWQ (Activation-aware Weight Quantization) [2] has already successfully achieved lossless quantization by quantizing model weights to 4 bits while leaving activations full-precision. Other works such as SmoothQuant [5] have worked toward quantizing both weights and activations, achieving strong performance for W8A8 (8-bit weights, 8-bit activations) quantization. For our project, we will investigate 4-bit quantization for activations as well, to attempt to achieve strong W4A4 quantization for LLM applications.

1.1 Related work

In the past, many different strategies have been used to compress neural networks, most notably pruning and quantization[1]. Pruning is the process of removing unnecessary connections or neurons to reduce model size while quantization reduces the memory footprint and computational requirements of large language models by representing model parameters in a lower precision. Thus, quantization effectively optimizes models without removing any of their underlying components. In this work, we focus on pushing the boundaries of quantization to low bit widths. A key difficulty in the quantization of LLMs is the presence of outliers in activations. Notably, recent works such as Activation-aware Weight Quantization (AWQ) [2] and SmoothQuant [5] implement strategies to address this problem, and they have achieved strong results for low-bit weight-only quantization and W8A8 quantization, respectively.

1.1.1 Activation-aware Weight Quantization

As mentioned above, Lin et al.'s method of AWQ achieves strong results for W4A16 (4-bit weight, 16-bit activation) compression on state of the art language models [2]. AWQ works by first identifying salient weight channels - weight channels that significantly affect an LLM's outputs. They examine an LLM's activations on a calibration set and find the channels with largest magnitude - they then denote the corresponding weight channels as salient. Then, to protect these channels, they multiply them by a scaling factor prior to quantization. A grid search is utilized to find the optimal scaling factor.

1.1.2 SmoothQuant

SmoothQuant [5] is another recent quantization strategy which instead targets quantization of both weights and activations. Their key observation was that activations are difficult to quantize whereas weights are uniform and much easier to quantize. SmoothQuant works by scaling the activations by a factor while simultaneously scaling the weights by the inverse. This preserves the numerical outputs while reducing the size of the activations and increasing the size of the weights. This strategy assumes that the quanitzation range for activations is larger than that of the weights. Thus, smoothing reduces the quantization range for activations and enables more effective bits per value. Importantly, the weights are not too difficult to quantize after this transformation.

2 Method

The objective of this work is to experiment with methods to attempt to achieve W4A4 (4-bit weight, 4-bit activation) quantization while preserving model performance. We iterate on the observations and strategies from AWQ (and in part, SmoothQuant) to push the quantization of LLMs to lower bit widths. In particular, we implement two methods - a group quantization with AWQ approach and a mixed-precision approach - to quantize Facebook's OPT-1.3b, OPT-2.7b, and OPT-6.7b models. We evaluate the quantized models using their perplexity on the WikiText-2 dataset (CITE) and compare against the perplexity of the original FP16 model. All experiments are run on a single NVIDIA A100 80GB GPU.

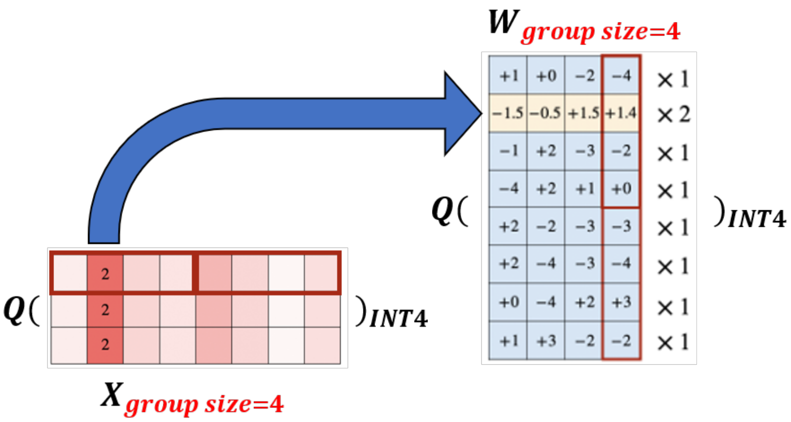

Figure 2: Protect salient weight and activation channels via scaling during W4A4 quantization with group quantization.

2.1 W4A4 quantization with group quantization and AWQ

The first approach is using W4A4 group quantization along with AWQ, as demonstrated in Figure 2. In other words, we first collect a small calibration set from the Pile dataset and pass it through the model to obtain average activation magnitudes. We then use grid search to identify the best "scale" α which minimizes the following objective.

$$s = s_X^{\alpha} \qquad \alpha^* = \text{arg min}_{\alpha} L(s_X^{\alpha}) \qquad L(s) = \left\| Q(W \cdot s)(s^{-1} \cdot X) - WX \right\|$$

After scaling the weights and activations appropriately, we now differ from the original AWQ paper by quantizing both weights *and activations* using 4-bit group quantization. Note that we quantize activations dymanically rather than statically (i.e. computing scales/shifts based on the incoming activations, rather than setting them using a calibration set), as this is needed for best results in such a low-bit quantization setting, despite a tradeoff in inference efficiency. We experiment with several group sizes - 32, 64, 128, and no grouping - and use vanilla round-to-nearest quantization (RTN) as the baseline.

2.2 Mixed-precision W4A4 quantization

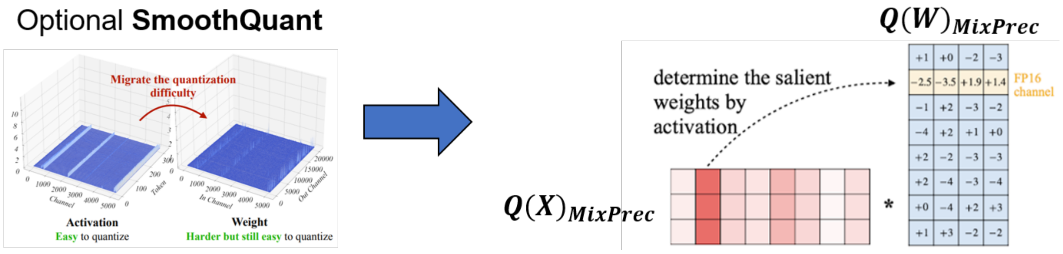

Figure 3: Preserve salient weight and activation channels in a mixed-precision format during W4A4 quantization.

The second approach is to use mixed-precision quantization, again inspired by an observation from AWQ. Rather than scaling the salient channels, we can protect them directly by preserving them as FP16. After identifying salient weight channels based on the corresponding activation outliers, we preserve both the salient weights and corresponding activations using FP16, while quantizing the rest to 4-bit using per-channel (for the weights) and per-token (for the activations) quantization.

Note that we do not use group quantization here, as this would be complex to implement and deploy efficiently in a mixed-precision setting. Also, due to poor perplexities from applying only the mixed-

Table 1: W4A4 Quantization with Group Quantization and AWQ

OPT/PPL ↓		**1.3B**	**2.7B**	**6.7B**
FP16	-	14.47	12.36	10.67
W4A4g32	RTN	**19.85**	12.89	10.98
	AWQ	21.39	**12.81**	**10.91**
W4A4g64	RTN	25.46	13.61	11.23
	AWQ	**24.56**	**13.13**	**11.10**
W4A4g128	RTN	35.12	15.01	12.02
	AWQ	**25.53**	**13.54**	**11.46**
W4A4	RTN	42482	14178	443080
	AWQ	**716.40**	**947.35**	**2533.81**

precision approach (as further described in 3), we further experiment with using SmoothQuant to scale the weights and activations prior to the FP16 protection and W4A4 quantization. We use the precomputed activation scales from the SmoothQuant paper for each of the corresponding OPT models.

3 Results and Discussion

3.1 W4A4 quantization with group quantization and AWQ

Overall, using the AWQ algorithm as a tuning mechanism for W4A4 quantization yields better performance than the naive method of round-to-nearest (RTN) quantization, with the effect becoming more significant for larger group sizes. For example, AWQ is able to close the perplexity gap between the FP16 and W4A4g128 quantized OPT-1.3B models by 46.4%. In addition, it is clear that using group quantization improves the performance significantly compared to no group quantization. Looking at Table 1, we see that the perplexity values for the models quantized via group quantization (g32, g64, and g128) are much better than the plain W4A4 quantized model. This is logical because the smaller the group size, the more unique quantization parameters can be applied. However, there is a trade off between having a smaller group size and the effective bit width, with small group sizes such as 32 and 64 incurring a significant number of extra bits and not “truly" yielding W4A4 quantization. Thus, we focus on the g128 result which is most commonly reported in the literature; we find that AWQ is able to bring the W4A4 quantized model perplexity to within 1.18 and 0.79 for the OPT-2.7b and OPT-6.7b models, respectively, but performs poorly on OPT-1.7b. A very small group size of 32 is required to bring the quantized model perplexities within 0.5 of the FP16 models for OPT-2.7b and OPT-6.7b.

3.2 Mixed-precision W4A4 quantization

Overall, the mixed precision strategy for W4A4 quantization is unable to recover the perplexity with reasonable protection ratios. Applying SmoothQuant prior to salient channel protection and quantization improves the performance; however, we cannot recover the perplexity unless we preserve around 95% of the salient activations and weights. As illustrated in Table 2, using mixed-precision W4A4 quantization with SmoothQuant on OPT-6.7B while preserving 10% of the salient activations and weights yields a perplexity of 891 - nowhere near the FP16 model. Only with preservation ratio of 95% do we see the perplexity return to within 0.5 of that of the original model - which can hardly be considered W4A4 quantization.

To understand why the perplexity is unable to be recovered despite the AWQ paper finding strong results for this approach under low-bit weight-only quantization, we examine two different quantization settings: low-bit weight-only quantization, and weight-and-activation quantization. As displayed in Figure 4, we see that for configurations where the weights are quantized to low bits and the activations are minimally quantized (W4A16, W4A8, and W3A8), the preservation of salient weights has a significant effect across the board; the blue (0% FP16) line is much higher than the orange (1% FP16)

Table 2: Mixed-precision W4A4 Quantization with and without SmoothQuant (SQ)

OPT/PPL ↓		**1.3B**	**2.7B**	**6.7B**
FP16	-	14.47	12.36	10.67
W4A4	No SQ	42482	14178	443080
	SQ	**234.00**	**1415**	**2393**
W4A4 1% FP16	No SQ	12653	27118	417241
	SQ	**125.52**	**847.54**	**2336**
W4A4 10% FP16	No SQ	16442	19484	47455
	SQ	**89.22**	**743.69**	**891**
W4A4 **95%** FP16	No SQ	**23.52**	12.57	10.82
	SQ	24.03	**12.49**	**10.76**

and green (10% FP16) lines. We observe near lossless quantization for W4A16 and W4A8 using the mixed precision approach; even the W3A8 model performs modestly.

On the other hand, when the compression configuration quantizes activations significantly as well, we see that the accuracy improvement from preserving salient weights and activations is minimal; the blue, orange, and green lines are very close together, and even the W6A6 mixed-precision model struggles. Therefore, we find that protecting salient weight and activation channels is useful in a low-bit, weight-only quantization setting (reaffirming the results from AWQ), but is ineffective under an extreme weight-and-activation quantization setting.

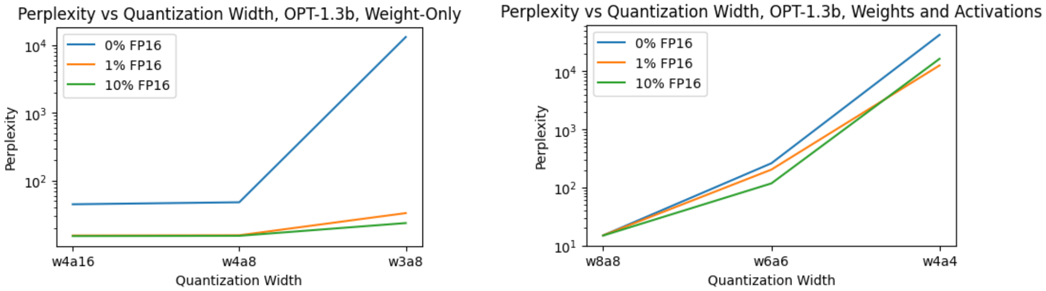

Figure 4: When doing weight-only quantization, preserving salient weights has a much greater effect on performance compared to preserving salient activations and weights during weight-and-activation quantization.

4 Conclusion and Future Work

Our work demonstrates that though methods like AWQ and SmoothQuant are simple to implement and efficient to deploy, they may be approacing their limits in terms of quantized model accuracy. Both of our approaches to W4A4 quantization evinced clear limiting factors; AWQ required a small group size to achieve decent performance (group size of 32 for perplexity difference less than 0.5), while the mixed-precision appraoch was unable to recover model perplexity entirely, even when used with SmoothQuant.

More recent works such as RPTQ [6] and OmniQuant [3] have demonstrated the ability to achieve lossless W4A4 quantization, but these methods too have drawbacks. RPTQ relies on clustering activations and quantizing along the inner channel dimension, which is inefficient when deployed. OmniQuant builds on SmoothQuant by learning shifts, scales, and clipping quantization parameters to minimize the reconstruction loss at each layer. However, this requires backpropagation and hence additional training overhead. Nonetheless, these works may offer valuable inspiration toward achieving more simple and efficieny W4A4 quantization strategies, and we encourage future investigation in this direction.

References

[1] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[2] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. Awq: Activation-aware weight quantization for llm compression and acceleration. *arXiv*, 2023.

[3] Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for large language models, 2023.

[4] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[5] Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[6] Zhihang Yuan, Lin Niu, Jiawei Liu, Wenyu Liu, Xinggang Wang, Yuzhang Shang, Guangyu Sun, Qiang Wu, Jiaxiang Wu, and Bingzhe Wu. Rptq: Reorder-based post-training quantization for large language models, 2023.

[7] 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. Opt: Open pre-trained transformer language models, 2022.