No-Op-Aware Training and Quantization Framework for Outlier Robust Transformer based Language Models

Sameed A. Khan, ASM H. Kabir

***Abstract*—This paper introduces a no-op-aware training and quantization framework for transformer-based language models that improves robustness to activation outliers while enabling efficient low-precision deployment. We modify an OPT-12L12H model with No-Op-Aware Attention Training (NOAT), combining conditional per-head gating with a Softmax1-based attention activation to suppress extreme attention during training. The model is trained and then quantized with two schemes: standard 8-bit uniform quantization and GPTQ-based post-training quantization. Experimental evaluation shows that the NOAT-trained, GPTQ-quantized model not only preserves but slightly improves perplexity to 10.68 compared to the full-precision 10.96. The paper also shows that GPTQ applied to the NOAT model closely matches the statistical structure of the full-precision activations by maintaining kurtosis, whereas uniform quantization exhibits heavier tails, indicating higher presence of outliers. Stabilizing attention activations during training substantially enhances the effectiveness of downstream quantization, narrowing the gap between model efficiency and accuracy and enabling more reliable deployment of large language models on low resource hardware.**

***Keywords*— Quantization, Outliers, No-Op-Aware Attention Training, Activations.**

I. Introduction

In recent years, Large language models (LLMs) [1] and other deep learning models have evolved into important components that are used in the modern artificial intelligence systems. Some of the useful applications of such models include machine translations, summarization, reasoning, text generations, information retrieval and automations. The growth of these models in terms of depth, scale, and capabilities also leads to demand for more computational power. In addition to computational power, training and deploying these models requires more memory bandwidth and energy. We also see specialized hardware as well being developed to cater such models in the past few years. However, even with and without these specialized hardware, the higher demand for energy, memory bandwidth, and computational power creates barriers for widespread and sustainable deployment for large deep learning models.

S. A. Khan, Innopolis University, Innopolis, Republic of Tatarstan, Russia, 420500 (e-mail: sameedkhandurrani@gmail.com)

ASM. H. Kabir, Moscow Institute of Physics and Technology (National Research University), Moscow Oblast, Russia, 141701 (e-mail: humaun.kabir@phystech.edu)

We have some algorithmic solutions for efficient training and deployment of large deep learning models. Model quantization and compression remains to be the most effective algorithmic solutions. Some other approaches include neural architecture search and model pruning.

Quantization remains to be the most promising approach since it can substantially reduce memory footprint and make inferences faster without any major architectural modifications [2]. It works mainly by converting floating point values into lower bit representations, enabling large models to run efficiently even on the edge devices and GPUs with limited resources. However it is not always an easy task to apply quantization to transformer models.

The problem arises when the transformers produce outliers in activations. These extreme values arise mainly within feed-forward network (FFN) layers and attention mechanisms [3]. These outliers distort scaling factors during quantization, causing the majority of values to collapse into narrower ranges and resulting in severe information loss and degraded model quality.

This is a big challenge and has led researchers to look for solutions to reduce presence of outliers before quantization. This implies not only focusing on post training strategies but shifting focus more towards robustness during the training phase.

Ensuring the stability of activations is increasingly becoming a crucial step in keeping a low precision model accurate. As the LLMs continue to grow in size and get adopted widely in the real world, it becomes essential to establish training methodologies that are quantization friendly.

Over time, researchers have developed some solutions to mitigate outliers. One such influential approach is SmoothQuant. It works by redistributing the activation and weight magnitudes to get rid of large outlier values. It does this by scaling each channel separately, which helps shrink the large, problematic activations. As a result, transformer layers become much easier to convert into 8-bit format without losing much accuracy.

In addition to such smoothing methods, the ongoing research continues to improve the quantization strategies including post-training and activation-aware techniques that selectively preserve salient parameters. Such research eventually leads to continued development of quantization robust model training frameworks which eventually results in the ability to deploy LLMs and other deep learning models on mainstream hardware without requiring extensive computational power or memory bound issues.

This paper introduces a new framework, No-Op-Aware Training (NOAT) and Quantization, to address these challenges by enhancing the robustness of transformer models against outliers. We propose combining NOAT with GPTQ quantization to achieve improved model performance and memory efficiency, making large-scale transformer models easier to be deployed and efficient in terms of energy on resource-constrained hardware.

A concept in this paper is the notion of a no-op-aware attention update. Here, a "no-op" does not imply that the model entirely skips the computation but it instead refers to a near-identity behavior of the attention branch. In this behavior the contextual update is minimal and the residual pathway remains dominant. This is different from the standard softmax attention where every query must distribute all of its probability mass over the available tokens. This leads to a full contextual update even when the evidence is weak. On the other hand, the NOAT formulation allows the model to express uncertainty more naturally by reducing the magnitude of the attention-induced update. The training is therefore referred to as no-op-aware, which means that during optimization the model learns not only where to attend, but also when it is preferable for attention to remain weak rather than strongly modifying the representation.

II. Related Work

A key goal in the field of machine learning is to reduce energy consumption and make models more energy-efficient. This can be achieved by reducing the number of bits involved in mathematical operations. Horowitz noted that 32-bit floating-point additions and multiplications can be up to 10 times more energy-consuming than 8-bit integer operations [4]. As a result, quantization has become one of the main techniques for model compression.

Some of the early works include K-means weight sharing for model compression, where the weights of a full-precision model are clustered. The method stores low-bit cluster indices along with a codebook of centroids [5]. This can lower storage costs by a factor of 3 or more, especially when combined with centroid fine-tuning to reduce reconstruction errors. However, such schemes, which involve codebooks, can complicate hardware implementations and are not well aligned with modern integer-only LLMs.

Uniform quantization, as proposed by [6], has been widely adopted in TensorFlow Lite, where floating-point values are mapped to integers through affine mapping. The mapping is defined as:

$$r = S(q - Z) \tag{1}$$

where,

r = floating-point weights,
S = scale (quantization parameter),
q = quantized weight as signed integer values,
Z = zero-point, the quantized integer representing r=0

This formulation underlies both symmetric weight quantization and asymmetric activation quantization, and is compatible with integer-only kernels.

Per-channel quantization [7] assigns a separate scale (and sometimes zero-point) for each output channel. This helps match the dynamic ranges across channels, improving accuracy compared to per-tensor scaling [8]. Similarly, per-vector schemes such as VS-Quant [9] propose assigning scale factors to short subvectors. This method trades a small overhead in scale storage for tighter local dynamic range.

In addition, the MX / Block Data Representation (BDR-MX) formats [10] share "micro-exponents" across blocks of values. It offers flexible low-precision floating-like formats such as MX4, MX6, and MX9. This approach maintains a balance between precision, hardware complexity, and feasibility for LLMs.

By contrast, the framework we propose focuses on the root cause of the problem—outlier activations during the training phase. No-Op-Aware Attention Training (NOAT) integrates conditional per-head gating and Softmax1 activation, which reduces the production of extreme values in attention outputs. This simplifies range estimation for both uniform quantization and GPTQ. In other words, we aim to improve the default behavior of standard post-training quantization pipelines rather than designing more complex range estimators.

Our approach complements both quantization-aware training (QAT) and memory-driven precision allocation. We do not modify bit-width per layer, nor do we simulate quantization during the training loop. Instead, we introduce architectural modifications such as per-head gating and Softmax1 to regularize attention activations in an OPT-12L12H model during standard full-precision training. Afterward, we apply PTQ (uniform) and GPTQ quantization. While NOAT could be combined with QAT or mixed-precision schemes, this is outside the scope of the current study.

The Open Pretrained Transformer (OPT) family provides decoder-only transformer architectures trained with a causal language modeling objective on large-scale web corpora. These models were designed as reproducible and transparent counterparts to GPT-style LLMs. OPT allows modifications to public training configurations, optimizer settings, and hyperparameters. The OPT architecture follows a standard transformer decoder stack with multi-head self-attention, feed-forward networks, learned token and positional embeddings, and layer normalization. It is trained using Adam-based optimizers, learning-rate warm up, and cosine or linear decay schedules [11].

OPT has gained popularity in quantization research due to its wide range of parameter scales, from small models to those with tens of billions of parameters. This makes OPT models suitable for systematic studies on how quantization errors scale with model size. Several quantization works, including those focused on activation smoothing techniques, use OPT checkpoints as primary benchmarks, particularly in 8-bit and 4-bit settings.

Prior research on transformer quantization has focused on refining numeric formats, improving range estimations, and training with STE-based quantization-aware training (QAT). Our work extends this body of research by demonstrating that No-Op-Aware Attention Training (NOAT), combined with per-head gated attention and Softmax1 activation, can be trained from scratch on an OPT-12L12H model. This

framework produces activation distributions that are more robust to low-bit quantization, resulting in GPTQ models that outperform both full-precision and uniformly quantized baselines in perplexity, while preserving the statistical structure of activations.

III. DATASET

The dataset used in this work is wikitext-103-raw-v1, this is a configuration of widely known WikiText dataset. WikiText is a vast language modelling dataset that has been constructed from verified Good and Featured articles on Wikipedia. Since it is derived from high quality articles, the dataset contains grammatically correct and well structured english text.

WikiText was initially introduced as an alternative to Penn Treebank (PTB) dataset and is considered an improved alternative since PTB requires heavy preprocessing and contains relatively limited vocabulary.

The wikitext-103-raw-v1 is over 110 times larger than PTB dataset and has following features:

- It preserves original casing, punctuation, numbers and the natural structure of the original articles from Wikipedia.
- It contains a much larger and more realistic vocabulary including the original tokens without replacing out-of-vocabulary words.
- It is suitable for character-level or subword-level modeling.
- The training set contains roughly 1.8 million examples and the validation set contains around 4.36 thousand examples as shown in the table I.

Table I. Counts of instances in dataset used

Dataset	Training Set	Test Set	Validation Set
wikitext-103-raw-v1	1,801,350	4,358	3,760

The total generated dataset size is approximated to be 549 MB, while the total disk usage (including downloaded files) is around 741 MB.

Since the dataset is large scale, high quality natural language data it supports training and evaluating LLM that we used in this work. Also the preserving nature of punctuation, casing, and numeric tokens makes it more realistic for language modeling.

IV. METHODOLOGY

A. About the Model

The LLM model used for this experiment is OPT-12L12H. Some features of this model are defined in the table II.

Table II. Parameters of OPT-12L12H

Parameter	Value
Maximum Sequence Length	512
Feed Forward Network Dimensions	3072
Dimensions of Layers	768
Number of Hidden Layers	12
Number of Attention Heads	12
Standard Deviation of Initializer for Weight Matrices	0.006
Dropout	0.1

There are three major reasons to opt for OPT-12L12H. First, it is found sufficiently expressive to show meaningful attention behavior for transformer models, including how activation outlier is relevant to quantization and how multi head interact. Secondly, it is computationally feasible for quantization experiments while enabling studying effect of proposed NOAT modifications. Lastly, OPT is a well established decoder only transformer with documented architectural configurations, providing a transparent, reproducible testbed. Thus, OPT-12L12H allows for a practical balance between representational relevance, interpretability, and experimental feasibility.

B. Training the Model

The OPT-12L12H model was trained with a causal language modeling objective, using sequences tokenized up to 2,048 tokens and cropped into training blocks of length 512. The model was optimized with AdamW using a learning rate of 0.0004, a total of 125,000 training steps with 2,000 warmup steps, gradient accumulation of 4, per-device batch size of 8 for both training and evaluation, weight decay of 0.1, and a maximum gradient norm of 1.0. In addition to the standard OPT architecture, the attention layers were modified to use conditional per-head gating (initialized at 0.25) together with a Softmax1-based attention activation.

The parameters for the model along with their counts are mentioned in the table III.

Table III. Parameter counts for OPT-12L12H model

Component	Parameter Count
Embeddings	39,003,648
Decoder	85,065,360
Head	38,608,896
Total (pre-training)	162,677,904
Total (decoder only)	124,069,008

The activation for the model was modified so that it uses Softmax1 instead. The Softmax1 is given by:

$$Softmax1(z_i) = \frac{e^{z_i}}{1+\sum_i e^{z_j}} \quad (2)$$

The training loss dropped to 2.109 after reaching the 125000th training step. Drop of loss is visualized in the figure 1.

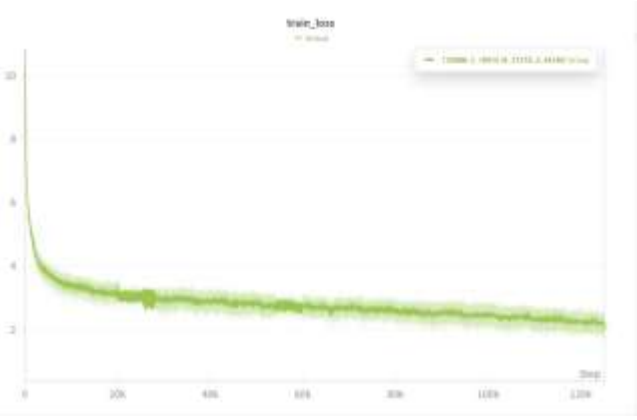

Fig. 1. Training Loss over Steps

The model was trained in segments on A100 GPU with 80GB RAM. Throughout the process, the average consumption of GPU stayed around ~73%. These stats of the GPU utilization are shown in figure 2.

Fig. 2. GPU utilization (%) measured over wall time during training

C. Motivation for No-Op-Aware Attention Training (NOAT)

The standard attention mechanisms of transformers may produce disproportionately large attention values in certain heads. This leads to heavy tailed activation distributions and poor behavior under low-precision quantization. Especially in case when the attention mechanism is forced to normalize probability mass under contextual tokens, even the evidence that are vague or weak are also able to generate a contextual update with full strength.

No-Op-Aware Attention Training (NOAT) is introduced to address this issue, where in the proposed settings a "no-op" refers to a near identity attention action. In "no-op", the contribution to hidden representation becomes negligible even while the attention branch is still computed. The model learns that it is best not to intervene in some context and for some tokens unlike the conventional attention, where the model produce a fully normalized contextual mixture.

Since the training makes the attention mechanism sensitive to this weak-update regime during optimization, it is termed as no-op-aware. Combining conditional per-head gating with attention normalization based on Softmax1 allows the model to reduce attention distributions that are overconfident and suppress unstable head outputs. Therefore, NOAT improves the compatibility of the model with subsequent quantization by reducing formation of activation outliers during the training.

D. Conditional Per-Head Gating

The gating values were set to 0.25 to avoid aggressive amplification and avoid sharp activation spikes. With the progress of training, the gating parameters adapt based on gradient updates increasing the influence of useful heads and suppressing unstable ones.

Mathematically, if H(i) represents the output of attention head i, and g(i) is the learnable gate parameter, then the modified attention output becomes:

$$\widetilde{H}_i = g_i \cdot H_i \quad (3)$$

The contribution of head negligible as the learned gate value approaches zero for a head, leading the corresponding attention pathway to behave as a no-op, while the larger gate values recover standard head participation.

E. Softmax1 Activation Based Attention Stabilization

The Softmax1 variant was used to constrain extreme probabilities. It works by preventing any token from dominating the attention distribution excessively by slightly regularizing the denominator term.

Softmax1 reduces kurtosis of activation distributions across attention layers. As a result, the attention outputs have smooth distribution and it makes them more compatible with low-bit quantization.

F. Uniform Quantization

The model was quantized to int-8. For quantization of weights, we used symmetric quantization. This is given by:

$$q_w = round\left(\frac{w}{s}\right), and\ s = round\left(\frac{\max|w|}{2^{b-1}-1}\right) \quad (4)$$

where,

w: full-precision weight
q_w: quantized integer weight
s: symmetric scale factor
b: bitwidth (e.g., 8 for int8)

For quantization of activations we used asymmetric quantization instead. The asymmetric quantization is defined by:

$$q_a = round\left(\frac{a-a_{\min}}{s}\right) and\ s = round\left(\frac{a_{\max}-a_{\min}}{2^b-1}\right) \quad (5)$$

where,

w: full-precision weight
q_a: quantized integer activation
a_{min}, a_{max}: observed min and max activation values
s: symmetric scale factor
b: bitwidth (e.g., 8 for int8)

G. Reproducibility and Implementation Details

To improve the reproducibility, we release the source code, configuration files, and evaluation scripts for the proposed framework. All experiments were implemented using Hugging Face Transformers using PyTorch along with Accelerate. NVIDIA A100 GPUs with 80 GB memory was used for training and evaluation. The experimental configurations are summarized in Table IV.

Table IV. Experimental Configurations

Parameter	Value
Model	OPT-12L12H
Tokenizer	Facebook/opt-350m
Objective	Causal language modeling
Optimizer	AdamW
Learning rate	0.0004
Warmup steps	2000
Total training steps	125000
Hardware	NVIDIA A100 80 GB

V. Results and Analysis

The trained model and the quantized models were evaluated thoroughly. To make sure that the comparison is fair, the models were evaluated under similar conditions with the same evaluation split and preprocessing pipeline.

We ran the model on the evaluation dataset. For the evaluation purposes also we used A100 GPU with 80GB memory. The use of high-memory GPU makes sure the evaluation was not bottlenecked by memory constraints.

The GPTQ quantized model trained with a no-op-aware approach resulted in better perplexity than uniform quantized model or full precision model. This indicates the proposed technique improves quantization while preserving the accuracy of the model.

Table V. Perplexity comparison in FP and Quantized models

Full Precision	GPTQ Quantized	Uniform Quantized
10.9646	10.6874	13.5917

Similarly, we did a comparison of Kurtosis of each model as well. The kurtosis of a model is given by:

$$\textbf{Kurtosis}(x) = \frac{\mathbb{E}[(x-\mu)^4]}{\alpha^4} \quad (6)$$

where,

$\mu = E[x]$ is the mean,

$\alpha 2 = E[(x - \mu)2]$ is the variance.

The higher kurtosis values suggest heavier tails and more likelihood of extreme activation values that negatively impact the low precision quantization.

Table VI. Kurtosis comparison in FP and Quantized models

Full Precision	GPTQ Quantized	Uniform Quantized
6.717	6.713	6.836

GPTQ quantized model has a kurtosis very close to full precision model. It suggests that the statistical structure of activations is largely preserved after quantization. However, on the other hand, the uniform quantized model show higher kurtosis suggesting that there is a greater presence of extreme values.

Overall, these results support the hypothesis that no-op-aware training combined with GPTQ quantization makes the model more robust to outliers by stabilizing the activation distribution during the training phase.

VI. Conclusion

In conclusion, we presented a No-Op-Aware Training and Quantization framework that improves the robustness of transformer-based language models against the outliers present in activations, while allowing deployment in low precision bits. By using the No-Op-Aware Attention Training (NOAT) with conditional per-head gating, along with Softmax1 attention, the attention activations are stabilized during the training phase. These stabilized attention activations lead to the reduced formation of outliers that degrade the quantization performance.

The experiments demonstrate that catering the outliers at training significantly improves quantization. The NOAT-trained model achieves lower perplexity when quantized with GPTQ, as compared to full precision model and the uniformly quantized model. Specifically, the GPTQ-quantized model achieved a perplexity of 10.68 compared to 10.96 for full precision and 13.57 for uniform quantization.

The statistical analysis of distribution is also in favor of these findings. The values of kurtosis for the GPTQ quantized model are similar to those of full precision model. It suggests that the NOAT training combined with GPTQ quantization mitigates the outliers better than uniform quantization that reflects greater distortion and reduced information fidelity.

The results suggest that the outliers can be catered not only during the quantization process but through architectural modifications during the training process. The proposed framework reduces the gap between model efficiency and accuracy, making it feasible to deploy transformer models on resource-constrained hardware without prominent performance loss.

In the future, this approach can be extended to heavier models, transformer models with different architectures, and for different bit precision (e.g., 4-bit). Future work may also include evaluating performance of tasks other than language modelling. The no-op-aware training method can be combined with other smoothing techniques to further improve scalability and energy efficiency.

In summary, no-op-aware training in combination with GPTQ quantization is a practical solution to outlier free, quantization friendly deployment of LLMs in the real world applications ensuring better accuracy.

Code Availability Statement - The implementation, experiment configurations, and evaluation scripts used in this work are publicly available at: https://github.com/SameedAhmedKhan/noat-opt

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008. [Online]. Available: https://arxiv.org/abs/1706.03762

[2] S. A. Khan, S. Shulepina, D. Shulepin, and R. A. Lukmanov, “Review of algorithmic solutions for deployment of neural networks on lite devices,” Computer Research and Modeling, vol. 16, no. 7, pp. 1601–1619, 2024. [Online]. Available: http://crm-en.ics.org.ru/journal/article/3557/

[3] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in Proc. Int. Conf. Learning Representations (ICLR), 2015.

[4] M. Horowitz, “Computing’s energy problem (and what we can do about it),” in Proc. 2014 IEEE Int. Solid-State Circuits Conf. (ISSCC), 2014, pp. 10–14, doi:10.1109/ISSCC.2014.6757323.

[5] S. Han, H. Mao, and W. J. Dally, “Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding,” arXiv preprint, 2016. [Online]. Available: https://arxiv.org/abs/1510.00149.

[6] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. G. Howard, H. Adam, and D. Kalenichenko, “Quantization and training of neural networks for efficient integer-arithmetic-only inference,” CoRR, 2017. [Online]. Available: http://arxiv.org/abs/1712.05877

[7] R. Krishnamoorthi, “Quantizing deep convolutional networks for efficient inference: A whitepaper," CoRR, 2018. [Online]. Available: http://arxiv.org/abs/1806.08342

[8] M. Nagel, M. van Baalen, T. Blankevoort, and M. Welling, “Data-free quantization through weight equalization and bias correction,” CoRR, 2019. [Online]. Available: http://arxiv.org/abs/1906.04721

[9] S. Dai, R. Venkatesan, H. Ren, B. Zimmer, W. J. Dally, and B. Khailany, “VS-Quant: Per-vector scaled quantization for accurate

low-precision neural network inference," CoRR, 2021. [Online]. Available: https://arxiv.org/abs/2102.04503

[10] B. Rouhani, R. Zhao, V. Elango, R. Shafipour, M. Hall, M. Mesmakhosroshahi, A. More, L. Melnick, M. Golub, G. Varatkar, et al., "With shared microexponents, a little shifting goes a long way," arXiv preprint, 2023. [Online]. Available: https://arxiv.org/abs/2302.08007

[11] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al., "OPT: Open pre-trained transformer language models," arXiv preprint, 2022. [Online]. Available: https://arxiv.org/abs/2205.01068

Sameed A. Khan, PhD researcher, Faculty of Computer Science and Engineering, Innopolis Univeristy, Innopolis, Russia. e-mail : sameedkhandurrani@gmail.com

A S M H. Kabir, PhD Student, Department of Intelligent Information Systems and Technologies, Moscow Institute of Physics and Technology, Moscow, Russia. e-mail : humaun.kabir@phystech.edu