Q-MAMBA: TOWARDS MORE EFFICIENT MAMBA MODELS VIA POST-TRAINING QUANTIZATION

Anonymous authors
Paper under double-blind review

A APPENDIX

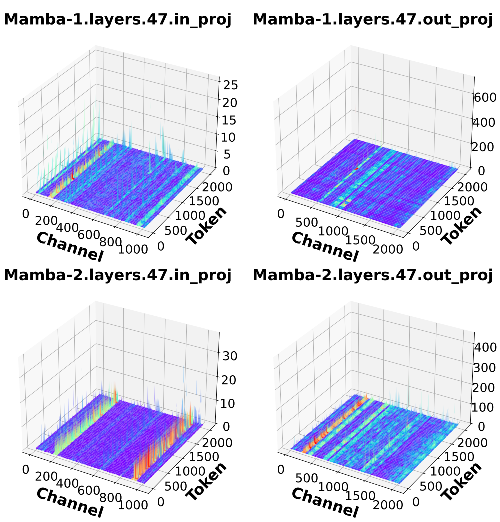

Figure 1: Visualization of inputs for linear projections. The out projection suffers from more severe outliers compared to the in projection.

A.1 PREVIOUS PTQ METHODS ON MAMBA

In Section 4, we analyze the quantization of linear projections in Mamba models. Here, we provide more detailed results about previous PTQ methods on Mamba-1 and Mamba-2 models. We will analyze the difference between Mamba-1 models and Mamba-2 models from a view of model quantization. The results presented in Table 1 indicate that Mamba2 models exhibit greater robustness to quantization compared to Mamba1 models. Further analysis in Figure 1 reveals that this improvement is largely due to the additional LayerNorm applied before the output projection in Mamba2, which helps to reduce outliers to a certain extent. Moreover, this LayerNorm simplifies the implementation of previous PTQ methods based on smoothing between weights and activations, such as SmoothQuant (Xiao et al., 2023) and AWQ (Lin et al., 2023). As a result, this paper primarily focuses on Mamba2 models, which not only feature larger state dimensions but are also more amenable to quantization.

A.2 PROOF

Theorem 1. *Assuming $u_t \sim \mathcal{N}(\mathbf{0}, \sigma \boldsymbol{I}_n)$ and A_t is a constant, $B_t, x_t = split(Wu_t)$ ($B_t \in \mathbb{R}^N$, $x_t \in \mathbb{R}^P$), the variance of states h_t can be factorized into two vectors:*

$$h_t = A_t \cdot h_{t-1} + x_t \cdot B_t^\top \tag{1}$$

$$Var[h_t] \propto \alpha \cdot \beta^T, \quad \alpha_i = ||W^x_{i,:}||_2^2 \quad and \quad \beta_i = ||W^B_{i,:}||_2^2 \tag{2}$$

Model	Method	Wikitext2	C4
Mamba1-370M	FP	14.31	17.23
	W8A8	18.95	23.04
	W8A8+SQ	16.17	19.85
	W4A16+ GPTQ	16.03	19.06
Mamba2-370M	FP	14.16	16.95
	W8A8	17.14	20.10
	W8A8+SQ	15.71	18.72
	W4A16+GPTQ	15.81	18.71

Table 1: Different PTQ methods for Mamba models. Mamba-1 models suffer much more serious outliers in output projections because of the absence of LayerNorm before it.

where $\alpha \in \mathbb{R}^P$ *and* $\beta \in \mathbb{R}^N$ *and* $W^B, W^x = split(W, dim = 0)$

Proof. Firstly, we can reformulate Equation (1) as a prefix sum:

$$h_t = \sum_i^t A_{i:t} x_i B_i^\top, \quad where \quad A_{i:t} = A_i \times A_{i+1} \times \dots A_t \tag{3}$$

Then, we can compute the mean of states h_t as follows:

$$\begin{aligned}
\mathbb{E}[h_t] &= \sum_i^t A_{i:t} \mathbb{E}[x_i B_i^\top] \\
&= \sum_i^t A_{i:t} \mathbb{E}[W^x u_i u_i^\top {W^b}^\top] \\
&= \sum_i^t A_{i:t} W^x \mathbb{E}[u_i u_i^\top] {W^b}^\top \\
&= \sum_i^t A_{i:t} \sigma W^x {W^b}^\top
\end{aligned} \tag{4}$$

After computing the mean of the states, we can similarly compute the variance of the states h_t. The equality (a) is attributed to Lemma 1.

$$\begin{aligned}
\mathrm{Var}[x_i B_i^\top] &= \mathbb{E}[(W^x u_i u_i^\top {W^b}^\top - \sigma W^x {W^b}^\top)] \\
&= \mathbb{E}[(W^x (u_i u_i^\top) {W^b}^\top)^2] - 2\sigma \cdot \mathbb{E}[W^x {W^b}^\top \odot (W^x u_i u_i^\top {W^b}^\top)] + (\sigma W^x {W^b}^\top)^2 \\
&\overset{(a)}{=} \sigma^2 \alpha \cdot \beta^\top + 2\sigma^2 \cdot (W^x W_b^\top)^2 - 2\sigma^2 \cdot (W^x W_b^\top)^2 + \sigma^2 \cdot (W^x {W^b}^\top)^2 \\
&= \sigma^2 \alpha \cdot \beta^\top + \sigma^2 \cdot (W^x {W^b}^\top)^2
\end{aligned} \tag{5}$$

Here, we assume that the second term $(W^x {W^b}^\top)^2$ is sufficiently small compared to $\alpha \cdot \beta^\top$, and then we obtain:

$$\mathrm{Var}[h_t] = \quad = (\sigma^2 \sum_i^t A_{i:t}) \cdot (\alpha \cdot \beta^\top) \tag{6}$$

□

Lemma 1. *Assuming* $z \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$*,* $w_1, w_2 \in \mathbb{R}^n$*, we have the following conclusions:*

$$\mathbb{E}[({w_1}^\top z)^2 ({w_2}^\top z)^2] = ||w_1||_2^2 \cdot ||w_2||_2^2 + 2({w_1}^\top w_2)^2 \tag{7}$$

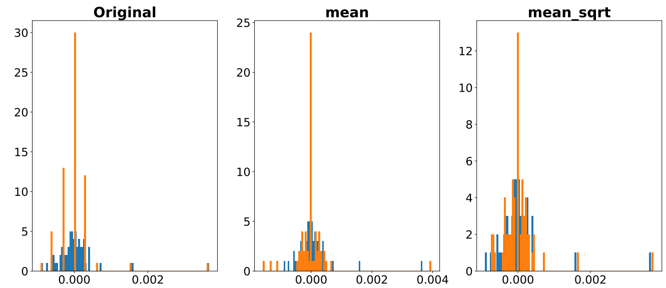

Figure 2: An illustration of how DSQ enhances performance.

Proof. Let A and B be two arbitrary symmetric matrices, we have:

$$\begin{aligned}
\mathbb{E}\left[x^\top A x \cdot x^\top B x\right] &= \mathbb{E}\left[\sum_{i,j} x_i a_{ij} x_j \sum_{k,l} x_k b_{kl} x_l\right] \\
&= \mathbb{E}\left[\sum_{i,k} a_{ii} b_{kk} x_i^2 x_k^2 + 4 \sum_{i<j} a_{ij} b_{ij} x_i^2 x_j^2\right] \\
&= \sum_{i,k} a_{ii} b_{kk} + 2 \sum_i a_{ii} b_{ii} + 2\left(\sum_{i,j} a_{ij} b_{ij} - \sum_i a_{ii} b_{ii}\right) \\
&= \sum_i a_{ii} \sum_k b_{kk} + 2 \sum_{i,j} a_{ij} b_{ij} \\
&= \text{Tr}(A)\text{Tr}(B) + 2\text{Tr}(AB)
\end{aligned} \tag{8}$$

A special case occurs when $A = w_1 {w_1}^\top$ and $B = w_2 {w_2}^\top$:

$$\mathbb{E}[({w_1}^\top z)^2 ({w_2}^\top z)^2] = ||w_1||_2^2 \cdot ||w_2||_2^2 + 2({w_1}^\top w_2)^2 \tag{9}$$

□

Although this theorem imposes strict constraints on the SSM inputs u_t (Gaussian distribution) and A_t (constant), it sufficiently reveals the following fact: outliers in the channel dimension P and state dimension N can be attributed to the variables $x_t \in \mathbb{R}^{(T,P)}$ and $B_t \in \mathbb{R}^{(T,N)}$, respectively. Figure 4(b) provides a visualization of this phenomenon.

A.3 More Ablation Studies

Visualization of DSQ. Figure 2 illustrates how DSQ improves performance. The presence of outliers causes MinMax quantization to waste a significant portion of available quantization slots, resulting in large rounding errors. Although introducing channel scales $S_{channel}$ helps make the quantization slots non-uniform, the mean norm remains sensitive to outliers, even unexpectedly amplifying them (as shown in the middle figure).

Trainable parameters in ESR. Table 2 demonstrates the effectiveness of our choice of trainable parameters in ESR: Fine-tuning selective parameters (B, C, and Δ), layer normalization, and convolution yields the best perplexity. In contrast, including x and z results in worse performance. We attribute this to the fact that fine-tuning all parameters can lead to overfitting and necessitates end-to-end training.

Norm	Δ,B,C,D	Conv-1D	X,Z	Wikitext2	C4
				25.73	29.94
✓				24.76	29.02
	✓			23.27	27.22
		✓		25.24	29.09
			✓	24.99	28.88
✓	✓			22.51	27.00
✓		✓		24.93	28.87
✓			✓	25.31	29.43
	✓	✓		22.68	26.91
	✓		✓	22.97	26.41
		✓	✓	25.66	28.89
✓	✓	✓		**21.92**	**25.99**
✓	✓		✓	23.63	27,43
✓		✓	✓	24.89	29.04
	✓	✓	✓	23.01	26.98
✓	✓	✓	✓	23.73	28.19

Table 2: The performance of W16A16H4 quantization for Mamba2-370M with different trainable parameters in the ESR.

REFERENCES

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. AWQ: activation-aware weight quantization for LLM compression and acceleration. *CoRR*, abs/2306.00978, 2023. doi: 10.48550/ARXIV.2306.00978. URL https://doi.org/10.48550/arXiv.2306.00978.

Guangxuan Xiao, Ji Lin, Mickaël Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38087–38099. PMLR, 2023. URL https://proceedings.mlr.press/v202/xiao23c.html.