Model selection

It should be noted that the three papers (Tolstikhin et al. 2021; Touvron et al. 2021; Trockman and Kolter 2022) mentioned in our main paper provide several variants for each model architecture. We choose Mixer-B/16 for MLP-Mixer, ResMLP-S24 for ResMLP, and ConvMixer-768/32 for ConvMixer. These three models are relatively similar in the number of parameters so that we can make reasonable comparisons.

Additional results

This section presents some additional results that are not mentioned in the main paper. These results either do not help much to derive the main conclusions or happen to be unsuccessful attempts. However, these results may still be meaningful for future investigation.

4-bit post-training quantization (PTQ) results

4-bit PTQ results are not mentioned in our main contents because 4-bit quantization settings usually require QAT to achieve desirable performance. However, some of these 4-bit PTQ results may bring insights to the MLP-based models. We include the 4bit PTQ results of baselines and the best quantization model of each MLP variant in Table 10. We should emphasize that we did not use percentile quantization to improve the accuracy here since we want to reflect each model's potential in 4-bit PTQ directly. However, we use asymmetric quantization in all 4-bit PTQ experiments to ease the activation sensitivity issue. Results show that MLP-Mixer does not suffer much under 4-bit quantization settings, while ResMLP and ConvMixer encounter severe accuracy degradation. These results are consistent with our analysis in the main paper that models with large activation ranges suffer more in quantization, and it also implies that the original MLP-Mixer has potential in ultra-low-bit quantization due to its uniform structure.

Table 10: PTQ results of 4-bit quantization.

Model	Precisioin	Method	Size(MB)	BOPS(G)	Top-1
MLP-Mixer	W32A32	Token-mixing	240	25825	76.64
		Multi-token-mixing	261	25825	78.35
Q-MLP-Mixer	W4A8	Token-mixing	30	807	75.82
		Multi-token-mixing	33	807	76.99
ResMLP	W32A32	Affine	120	12226	79.38
		LN	120	12226	79.59
Q-ResMLP	W4A8	Affine	15	382	60.67
		LN	15	382	61.12
ConvMixer	W32A32	ReLU	84	42762	80.16
		PACT	84	42762	80.22
Q-ConvMixer	W4A8	ReLU	11	1336	60.67
		PACT	11	1336	63.73

GELU vs. ReLU

We found that replacing GELU with ReLU does not help to improve the quantization performance or restrict the activation range. As shown in Table 11, GELU seems better for ResMLP, while ReLU works better for ConvMixer. However, the difference between GELU and ReLU is negligible, and small randomness during the training process may cause the difference between the two variants. The activation ranges of GELU and ReLU are also similar since the max absolute values of the two activations are close. The noticeable accuracy degradation in Table 11 implies that we need bounded activation functions (e.g., PACT) to deal with extremely large activation ranges, as discussed in the main text.

Table 11: GELU vs. ReLU

Method	Precisioin	Activation	Size(MB)	BOPS(G)	Top-1
ResMLP	W32A32	GELU	120	12226	79.38
		ReLU	120	12226	79.19
Q-ResMLP	W8A8	GELU	30	764	79.20
		ReLU	30	764	78.52
ConvMixer	W32A32	GELU	84	42762	79.73
		ReLU	84	42762	80.16
Q-ConvMixer	W4A8	GELU	11	2672	52.39
		ReLU	11	2672	57.81

Restricting activation range

ConvMixer has an extremely large activation range and faces severe performance degradation. In contrast to our methods proposed in the main contents, we would also like to discuss a few other methods that fail to restrict the activation range.

Firstly, exchanging the position of BNs and convolutions in each ConvMixer layer does not help to restrict the activation range. Besides, including a weight-decay term during training is also not helpful. Though the weight-decay option may slightly restrict the weight values, the activation range remains almost the same.

Limitation and future work

Though we have presented extensive tables and ablation studies in Section 4 of the main paper, some work remains to be done in the future. Firstly, since big model variants with different structures have relatively large differences in terms of parameter size and FLOPs, it is hard for us to make fair comparisons to all of them. Consequently, our experimental results mainly focus on a few selected (relatively small) MLP variants, and more results among larger MLPs, CNNs, and transformer models remain to be explored in future work. Secondly, we use models pretrained solely on ImageNet. In (Tolstikhin et al. 2021), the authors state that MLP-Mixer tends to overfit more than ViT, which implies that MLP-Mixer will potentially benefit more when pre-trained with larger datasets (for example, JFT-300M). Therefore, it is interesting to include some model variants pretrained on larger datasets to see if more data benefit MLP-based models more than transformers and CNNs in their quantization results. Lastly, our work only explores the merit of uniform quantization in order to maximize efficiency during inference. Aspects of mixed-precision quantization can be explored in future work.