Supplementary Materials

A. Using the infoNCE Loss

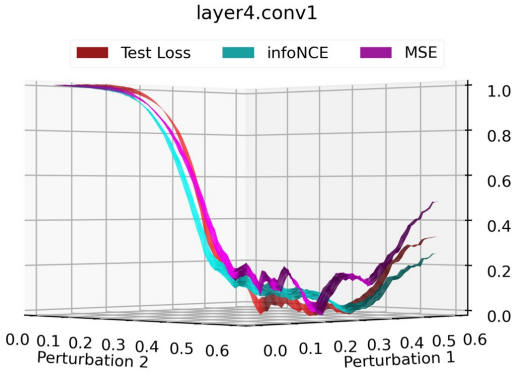
The infoNCE loss is an effective self-supervised learning technique to learn intermediary representations, but why apply it to quantization?

We find, both experimentally (in Fig. 5 of main paper) and qualitatively in Fig. 8, that the infoNCE loss provides better results than existing loss functions for global quantization. To perform global quantization, we try to minimize a loss between the quantized and full precision outputs given by:

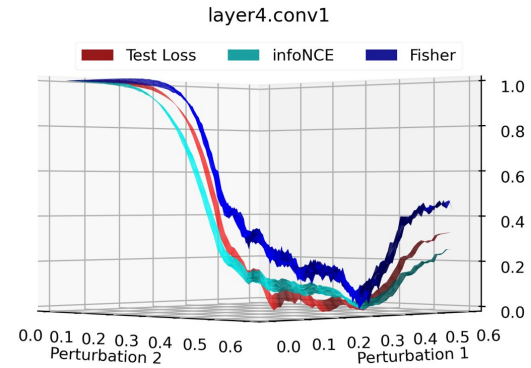
$$\arg\min_{\Delta} \mathcal{L}\big(x_Q(\Delta), \;\; x_{FP} \big) \tag{1}$$

where $x_Q(\Delta)$ is the quantized prediction parameterized by quantization scales Δ, and x_{FP} is the full precision prediction. It may seem reasonable to use the mean-squared error (MSE) or cosine similarity as a loss function in this setup. Unfortunately, PTQ methods only have access to a small calibration dataset, making it very easy for these loss functions to overfit to the few predictions available. The infoNCE loss combats this by using negative samples to encourage dissimilarity between x_Q and other predictions in the batch. We can see in Fig. 8a that the infoNCE loss provides a smoothing effect when compared to the MSE loss. The infoNCE loss has a flatter minima which aids in generalization to the unknown test distribution.

Additionally, Hessian-based loss functions allow for second order gradient information, however, they must be estimated using some form of approximation such as the Fisher loss used in BRECQ [6]. In Fig. 8b, we find the Fisher estimation to be noisy, and furthermore, does not accurately represent the underlying test loss landscape. The Fisher loss is an *empirical estimation*, and is a poor approximation when the training distribution does not match the test data distribution [5]. We find that the infoNCE loss performs much better since it does not rely on any gradient approximation, and more closely resembles the test loss. In Fig. 8b, we can see that the infoNCE and Fisher losses share a similar minimum, but the infoNCE provides a flatter neighborhood around the minimum which is more robust to data distribution shift [4, 3]. As discussed above, the infoNCE loss encourages diversity of representations by encouraging dissimilarity between predictions.



(a) Comparison of the test loss landscape with MSE and infoNCE loss landscapes.



(b) Comparison of the test loss landscape with Fisher and infoNCE loss landscapes.

Figure 8: Evaluating loss functions on ResNet-18. The infoNCE loss closely resembles the test loss (in red). In comparison, the MSE and Fisher loss are less smooth and do not accurately represent the test loss.

B. Ablation: Passes vs. Cycles

In Fig. 9, we ablate the number of passes, P, from 1 to 35. As we can see, a majority of the accuracy improvement occurs in the first 10 passes, so we choose

$P = 10$ for all experiments above. This allows for our method to run in less than one hour. However, we note that an additional accuracy boost may be enjoyed with more passes.

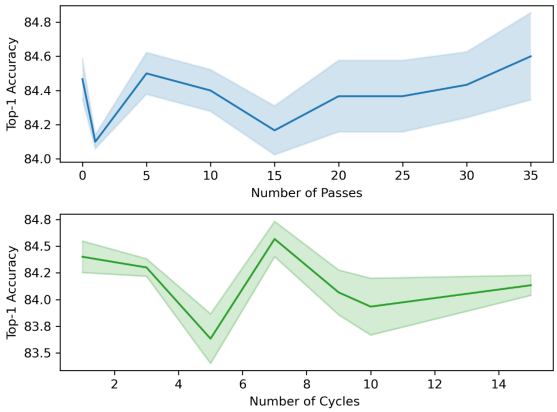


Figure 9: Ablation on number of cycles and passes.

We also ablate the number of cycles, C, to determine how many mutations should occur per block. We use $C = 3$ even though we see $C = 7$ is optimal in our ablation study. In practice, we find that the choice of C is random seed and model dependent. We find that for some runs, the best choice is simply 1 cycle, but in others it is 3, 5 or 7. Ultimately, we choose $C = 3$ for consistency across experiments.

C. Ablation on Calibration Set Size

As the calibration dataset increases, we'd expect better performance for our PTQ method. However, Fig. 10 suggests that a 512 images yields the highest performance, whereas 2,000 and 5,000 images makes performance worse than FQ-ViT (which uses 1,000 calibration images). This is likely an artifact of the way we implement contrastive loss.

When we apply contrastive loss on a batch of images, the contrastive loss minimizes the distance to the corresponding full precision prediction, but maximizes the dissimilarity across all other images, regardless of the whether of not the other images are in the same class. Ideally, we want to avoid maximizing the dissimilarity within a class, so a smaller calibration dataset will minimize the likelihood of two images belonging to the same class.

We use 1,000 images in this paper as in prior work, however, accuracy may be improved by using only 512 images. Alternatively, a labelled calibration set may allow the contrastive loss to ignore other images belonging to the same class.

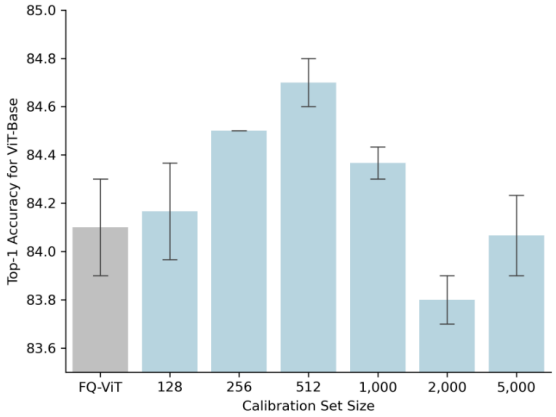


Figure 10: Ablation on Evol-Q's calibration dataset for sizes 128 to 5,000.

D. On Variation across Random Seeds

In Fig. 11, we show the performance of Evol-Q compared to the baseline method, FQ-ViT. Across twelve random seeds, ten runs improve performance over FQ-ViT, and three result in top-1 accuracy that is superior to the full precision model.

The random seed dictates which images are chosen for the calibration dataset, and we attribute the poor accuracy in seeds 4 and 5 to the poor choice of calibration set. This is a limitation of PTQ methods which rely on a calibration dataset, and so we employ a contrastive loss to combat overfitting (we can only minimize it's effect and not eliminate it).

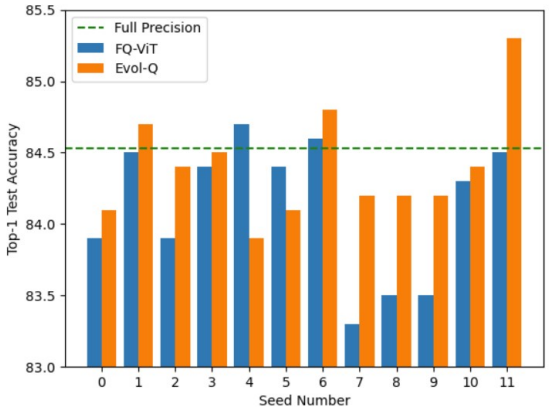


Figure 11: Comparing performance across 12 random seeds for 8W8A ViT-Base. 10/12 runs improve over the initial FQ-ViT quantization.

E. Impact on Attention Maps

We find that Evol-Q preserves the spatial integrity of the full precision feature maps even as quantization

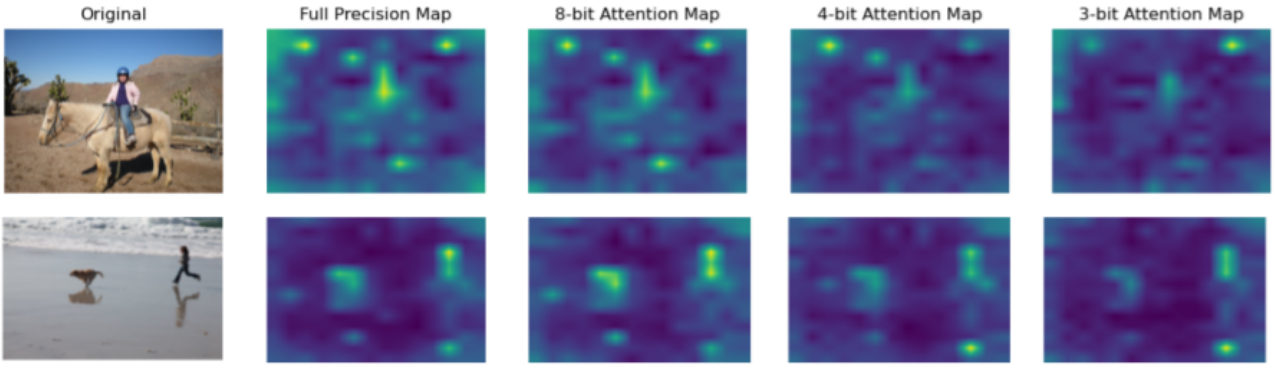


Figure 12: Attention maps for different quantization levels. Evol-Q's quantized models preserve the spacial locality of the full precision feature map. As the quantization level becomes more extreme, the attention map becomes subject to decreased resolution.

forces discretization of the attention mechanism. In Fig. 12, as quantization becomes more severe from 8-bit to 3-bit, the resolution of the feature map degrades, as is expected when only a finite number of values can be expressed in the quantized scheme. This attention map visualization is averaged over all blocks, and serves as qualitative inspection of how the quantized network's attention mechanism is performing. All in all, Fig. 12 provides confidence that Evol-Q's quantized attention maps learn reasonable representations of the original full precision network.

F. Layer-wise Weight Distributions

The weight distributions for ViT-Base's projection layers are shown in Fig. 14. To recap, the projection layer is the final linear layer of each attention block[1].

The beauty of Evol-Q is in its global optimization strategy – learning quantization scales with respect to a global objective allows Evol-Q to choose scales for the intermediary layers which improve quantization for other layers. FQ-ViT may approximate the full precision weight distribution well, however, a matching layer-wise distribution may not translate to overall performance gain. As explained in the main paper, a small perturbation in quantization scale can reap a huge accuracy gain. We can see that Evol-Q's layer-wise distributions are not very different than FQ-ViT, yet Evol-Q has a 0.15% accuracy improvement over FQ-ViT for ViT-Base. In summary, we find that Evol-Q's slight adjustment in quantization scale can greatly improve accuracy.

Please refer to the last page for Fig. 14.

[1] W^O in Pytorch's `torch.nn.MultiheadAttention()`

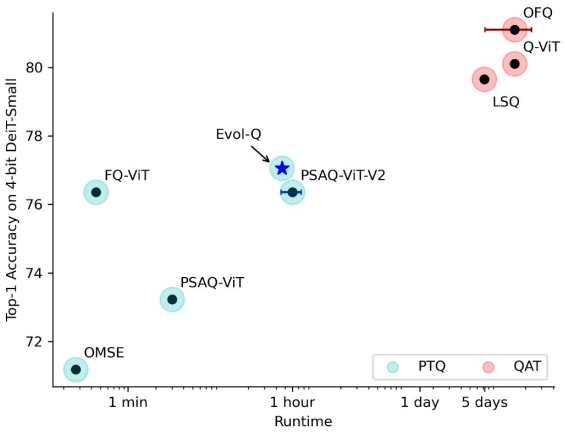


Figure 13: Runtime *vs.* Accuracy for 4-bit DeiT-Small using existing vision transformer techniques. We compare PTQ methods (blue) and QAT methods (red) on the same plot and show that Evol-Q is on the Pareto front. We estimate runtime for PSAQ-ViT-V2 [7] and OFQ [8], and indicate uncertainty using error bars.

G. Pareto Front for 4-bit DeiT-Small

Since most methods report 4-bit weights for DeiT-Small, we compare these methods in terms of both runtime & accuracy. In Fig. 13 we illustrate tradeoff between runtime and accuracy for PTQ and QAT methods. In comparison to 8-bit ViT-Base (Fig. 7 in the main paper), this figure includes QAT results which are unavailable in the 8-bit setting. We estimate runtime for PSAQ-ViT-V2 [7] and OFQ [8], since they do not open-source their code, nor report runtime. Evol-Q is on the Pareto curve (note x-axis is log scale), and has the best accuracy of all PTQ methods. Still, there is a

3-bit weights, 8-bit activations (3W8A)				
Method	DeiT-T	DeiT-S	DeiT-B	ViT-B
FQ-ViT	35.79	60.58	72.11	55.33
+ OMSE	52.03	65.27	75.00	62.83
+ Bias Corr	56.17	68.53	77.57	73.27
Evol-Q (ours)	**58.93**	**69.93**	**78.40**	**75.00**

(a) 3-bit weights, 8-bit activations

4-bit weights, 8-bit activations (4W8A)				
Method	DeiT-T	DeiT-S	DeiT-B	ViT-B
FQ-ViT	66.91	76.93	79.99	78.73
+ OMSE	66.03	77.17	80.30	78.90
+ Bias Corr	67.27	78.03	80.43	79.37
Evol-Q (ours)	**68.47**	**78.30**	**81.07**	**80.37**

(b) 4-bit weights, 8-bit activations

8-bit weights, 8-bit activations (8W8A)				
Method	DeiT-T	DeiT-S	DeiT-B	ViT-B
FQ-ViT	71.61	79.17	81.20	83.30
+ OMSE	72.17	80.30	82.17	82.47
+ Bias Corr	72.33	79.87	82.07	82.43
Evol-Q (ours)	**72.37**	**80.33**	**82.47**	**84.40**

(c) 8-bit weights, 8-bit activations

Table 1: We add OMSE quantization and Bias Correction (Bias Corr) on top of FQ-ViT. Finally, we apply Evol-Q on top of all three methods to achieve state-of-the-art PTQ quantization. We show results for 3W8A, 4W8A, 8W8A in Tab. 1a, Tab. 1b, and Tab. 1c respectively.

performance gap ($\sim 2.5-3\%$) when compared to QAT methods, illustrating that there is room to improve for PTQ methods.

H. Adding Bias Correction and OMSE

OMSE quantization [2] and Bias Correction [1] are statistical techniques we can use to improve quantization performance. We apply them on the original FQ-ViT model, and then use Evol-Q to achieve state-of-the-art PTQ performance. In Tab. 1 (last page), we can see the benefits of applying OMSE and Bias Correction techniques and how adding Evol-Q on top of these can boost performance even more.

In this paper, we have shown how Evol-Q can boost performance in a variety of scenarios and does not require a cherry-picked setting. We show that Evol-Q works on top of BRECQ for CNNs, FQ-ViT for ViTs, and even works in this setting, where we boost FQ-ViT's accuracy by adding Bias Correction and OMSE.

In summary, we are confident that Evol-Q's novel optimization method in conjunction with evaluating small scale perturbations is orthogonal to other quantization methods and can be used in a variety of scenarios to improve accuracy.

References

[1] Ron Banner, Yury Nahshan, and Daniel Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. *Advances in Neural Information Processing Systems*, 32, 2019. 4

[2] Yoni Choukroun, Eli Kravchik, Fan Yang, and Pavel Kisilev. Low-bit quantization of neural networks for efficient inference. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3009–3018. IEEE, 2019. 4

[3] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018. 1

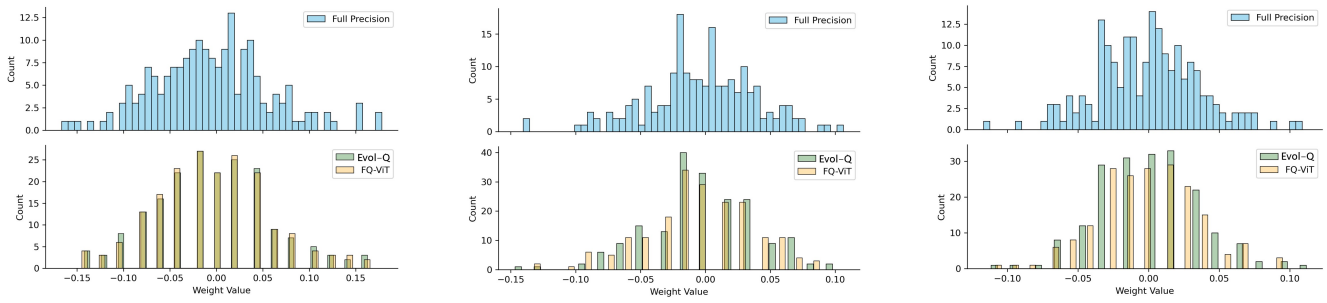
[4] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016. 1

[5] Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical fisher approximation for natural gradient descent. *Advances in neural information processing systems*, 32, 2019. 1

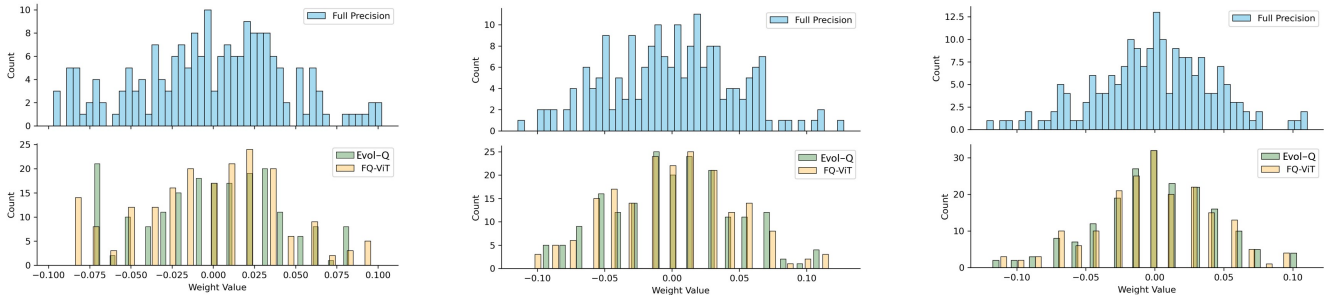
[6] Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. Brecq: Pushing the limit of post-training quantization by block reconstruction. *arXiv preprint arXiv:2102.05426*, 2021. 1

[7] Zhikai Li, Mengjuan Chen, Junrui Xiao, and Qingyi Gu. Psaq-vit v2: Towards accurate and general data-free quantization for vision transformers. *arXiv preprint arXiv:2209.05687*, 2022. 3

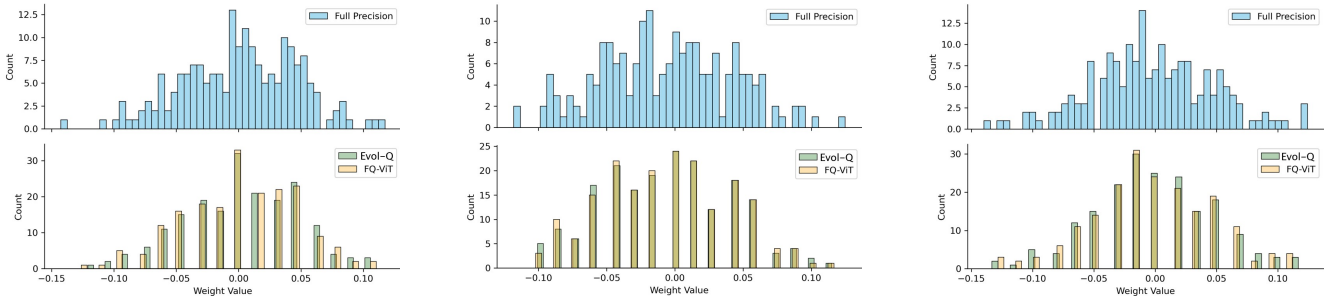
[8] Shih-Yang Liu, Zechun Liu, and Kwang-Ting Cheng. Oscillation-free quantization for low-bit vision transformers. *arXiv preprint arXiv:2302.02210*, 2023. 3



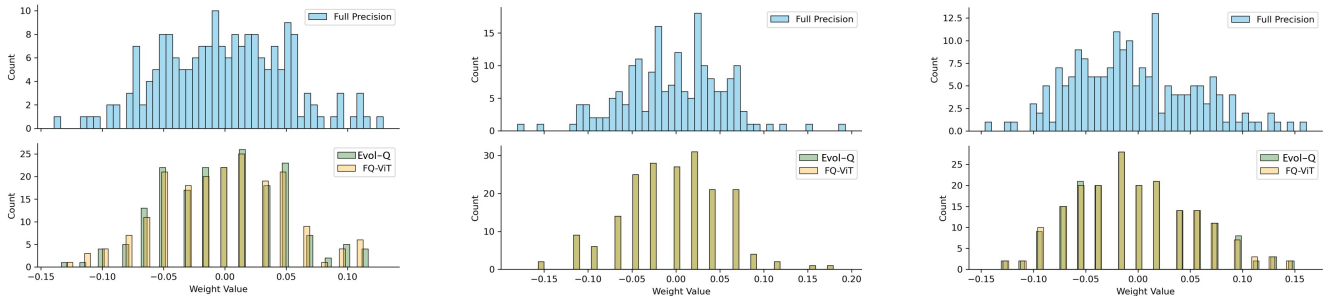
(a) Block #0 (b) Block #1 (c) Block #2



(d) Block #3 (e) Block #4 (f) Block #5



(g) Block #6 (h) Block #7 (i) Block #8



(j) Block #9 (k) Block #10 (l) Block #11

Figure 14: Weight distributions for the projection layers of all attention blocks for ViT-Base. The 12 blocks are numbered from 0-11, with block #1 being the same as reported in Fig. 4 of the main paper. Evol-Q (green) has a 0.15% Top-1 accuracy improvement over FQ-ViT (yellow).