A. Appendix

A.1. Plots From Main CIFAR-10 Ablation Study

Due to space constraints, we ommitted plots of the layerwise weights data from the main paper since the BN-Folded weights were the ones being quantized. However, we can see that even prior to BN-Folding, the weights distributions of MobileNets are poorly behaved. BN-Folding further aggravates this issue. By analyzing the change in range/average precision of weights distributions before/after BN-Folding we can see how certain layers needed greater scaling to properly normalize their representations. This implies a greater misalignment in the learned distributions of that layer. Consequently, we observe even larger BN-Folding-induced distributional shifts in the ImageNet-trained MobileNets models (see Figure 8).

We also ommitted the layerwise activation plots as we felt that they followed very similar trends to the BN-Folded weights. Those plots can be seen below in Figures 1, 2.

Finally, the layerwise QMSE plots in the main paper had some outliers ommitted as those networks' QMSE spiked so high that we could not display the general behaviour of the other networks. The full layerwise QMSE plots with all networks included are in Figure 7. Incidentally, these outlier networks also suffered from the greatest quantization degradation.

A.2. ResNet-34 CIFAR-10 Experiment Details

In addition to the Regular-Conv and DWS-Conv networks, we also wanted to analyze the multiscale distributional dynamics of a more complex CNN block. Thus, we trained a ResNet-34 based architecture with some modifications for the CIFAR-10 image resolution. Those changes are as follows:

- We do not downsample with MaxPooling after the very first conv layer
- We do not downsample with stride = 2 at the first residual block with output channels 128
- We do not downsample with stride = 2 at the first residual block with output channels 512
- Thus, the input tensor shape to Global Average Pooling layer is $4 \times 4 \times 1024$ rather than $1 \times 1 \times 1024$

The rest of the training details are the same as those mentioned in the Experiments section of the main paper. Full results from all of the CIFAR-10 trained networks are in Table 1. Interestingly, the ResNet-34 quantization results are fairly similar to the Regular-Conv Networks. We had originally hypothesized that the skip connection may enable the convolution layers to learn more compact distributions and have less quantization degradation. However, as seen from the quantization results and the plots in Figures 3–6 this was not necessarily the case. It would appear that while ResNet-34 does indeed learn compact weight distributions (see weights/BN-Fold weights ranges in Figure 3), the introduction of a skip connection has led to more fluctuations in average precision and possibly offset any potential gains from the smaller dynamic ranges. However, further analysis is required to properly understand this. Overall, it is currently unclear how the introduction of skip connections might affect the quantization dynamics of the system. It would be interesting to see how a ResNet and Regular-ConvNet of equivalent layer-depths behave once trained and quantized. Additionally, comparing bottleneck residual block with inverted bottleneck residual (ie. MobileNet-V2) block should yield further insights on the interplay between reduced range, increased distributional mismatches, and the complex quantization dynamics these systems yield.

A.3. Layerwise Plots From ImageNet Analysis

Here, we've included all of the ommitted layerwise plots from the fine-grained, multiscale analysis of the ImageNet trained networks. As mentioned in the main paper, it was hard to compare the layerwise QMSE and Activations data. The differences in preprocessing for the two networks led to widely different scales/ranges. Worth noting is the overall trends in these distributions. QMSE increases and then decreases in VGG-19 while for MobileNet-V1 it stays generally high (compared to our CIFAR-10 networks which were preprocessed the same way. Eg. Normalized to the range [-1, 1]) and spikes near the end (see Figure 9). Furthermore, we still observe similar trends in fluctuating ranges/average precisions from the layerwise activations plots in Figure 10. Thus, further supporting our hypothesis that depthwise-separable convolutional networks tend towards learning mismatched distributions, regardless of training data.

Worth noting is that the diversity of a large-scale dataset like ImageNet also seems to have increased mismatch of distributions for VGG-19 (see in Figure 10 as well as ImageNet results in the main paper). Though the distributions of VGG-19 are better behaved in general, we see that regular-conv networks might also benefit from better-aligned distributional dynamics and the quantized accuracy drop could be related to the drop in precision plus spike in range we observe for both the weights and activations near the "middle" of VGG-19.

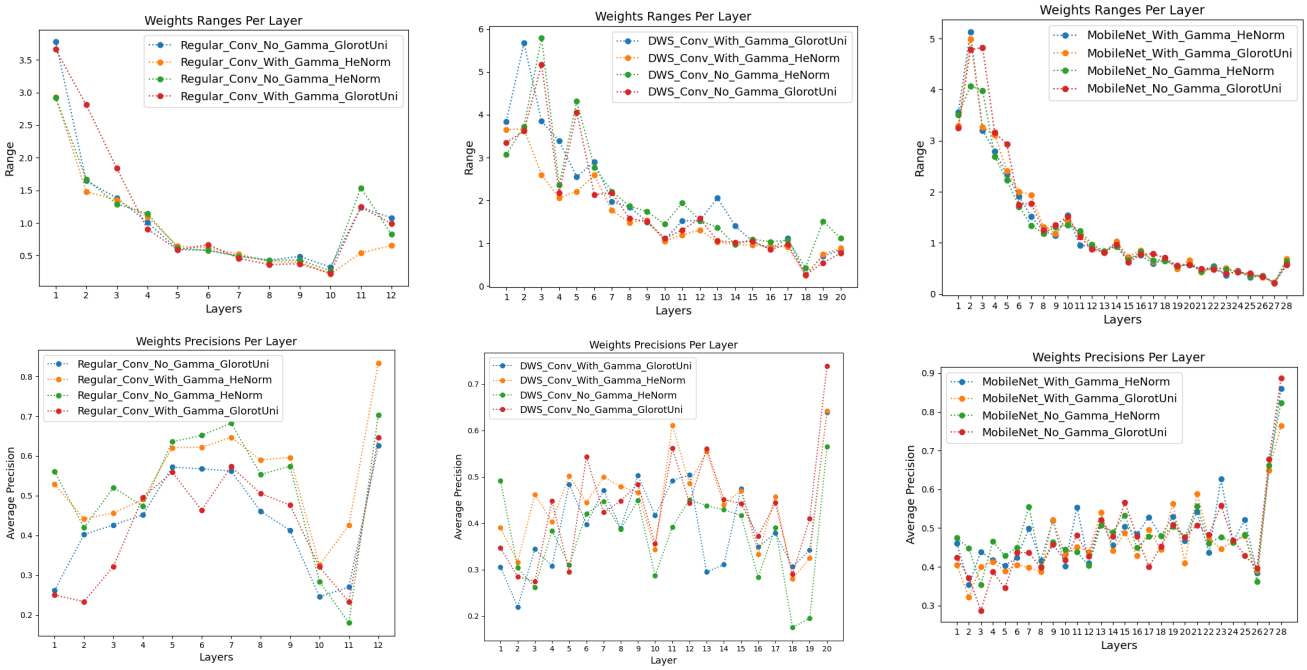

Figure 1. Layerwise weights range (top row) and average precision (bottom row) for all trained networks. Regular CNN (left), DWSCNN (center), and MobileNets-V1 (right).

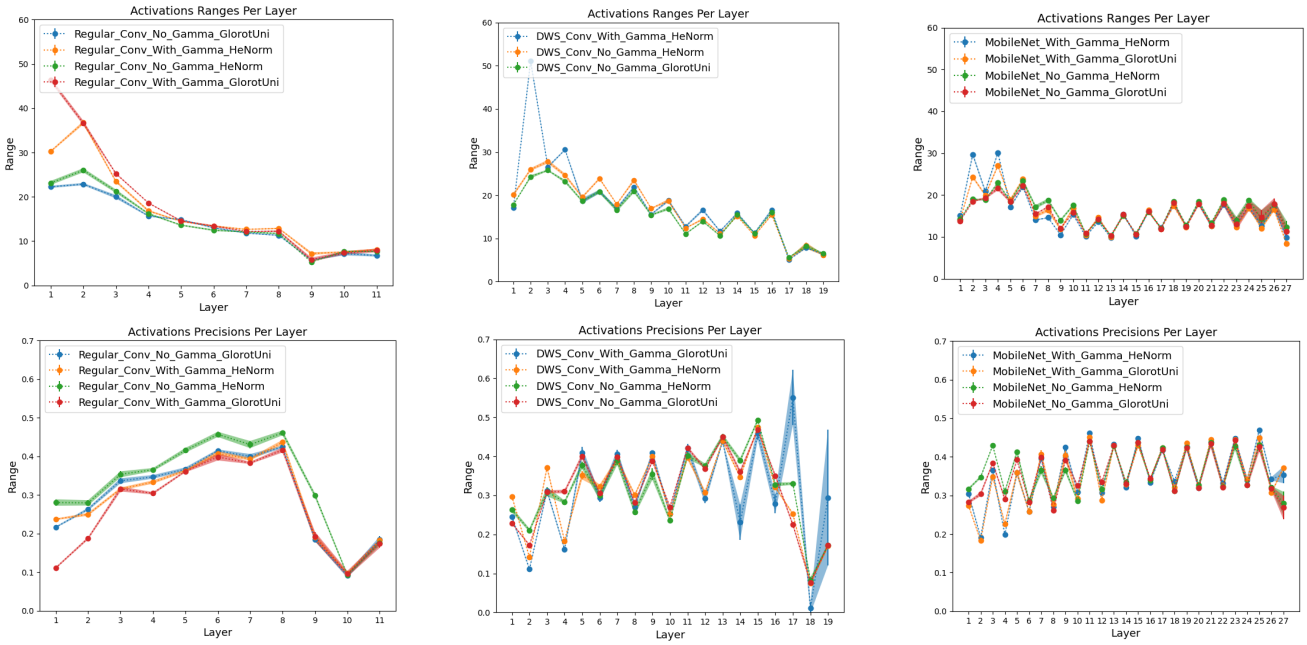

Figure 2. Layerwise activation range (top row) and average precision (bottom row) for all trained networks. Regular CNN (left), DWSCNN (center), and MobileNets-V1 (right). The solid line represents the average values across 5 quantization trials and the shaded region is the standard devation. We can see from the shaded regions that the choice of calibration dataset can lead to non-trivial variations in quantization parameters. **Note:** DWS-Conv-With-Gamma-GlorotUni has been ommitted from activation range plots due to its extreme outlier activation range of 200 at the first Dense layer (layer 18). This network also had the worst quantization behaviour out of all experiments.

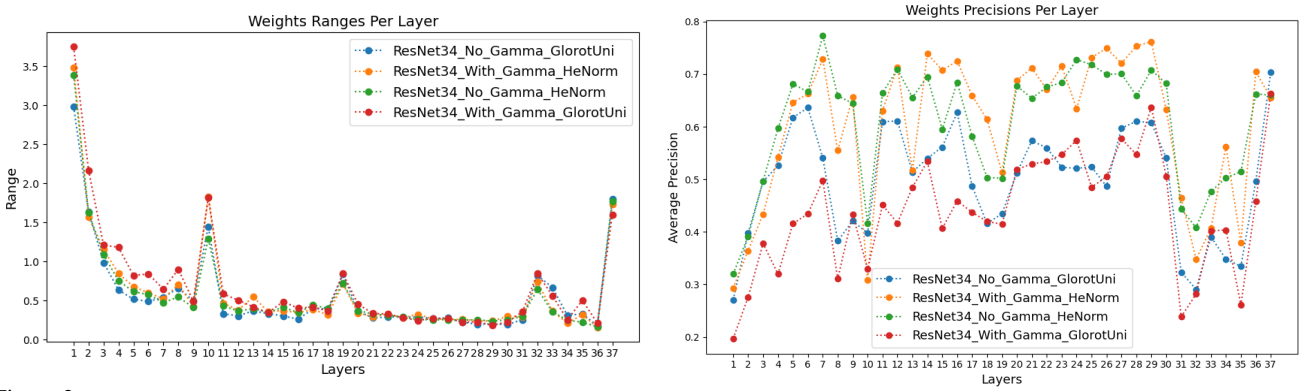

Figure 3. Layerwise weights range (left) and average precision (right) for ResNet-34 networks. Note, the spikes observed on the layerwise range plots at layers 10, 19, and 32 correspond to the 1×1 linear projection convolution.

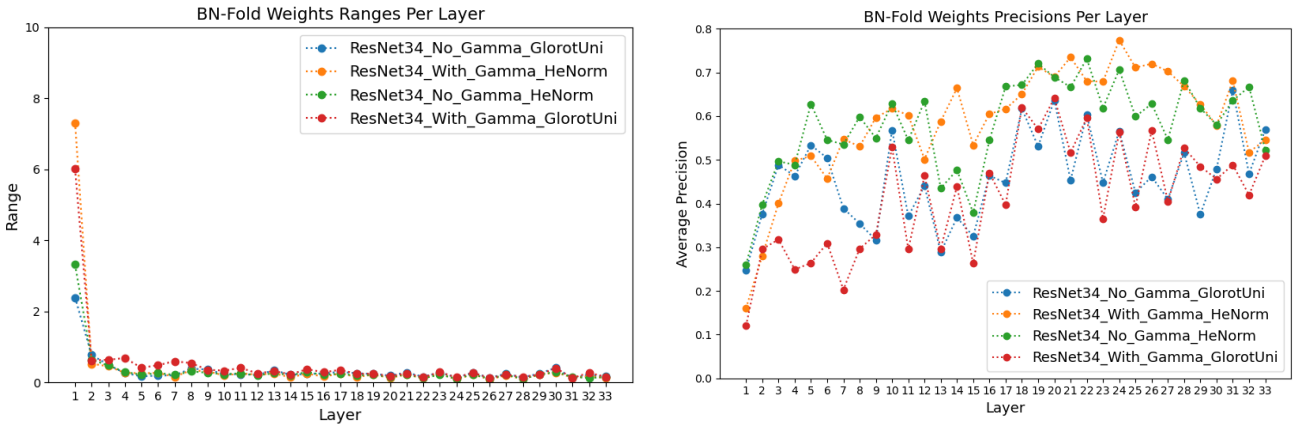

Figure 4. Layerwise BN-folded weights dynamic range (left) and average precision (right) for ResNet-34 networks.

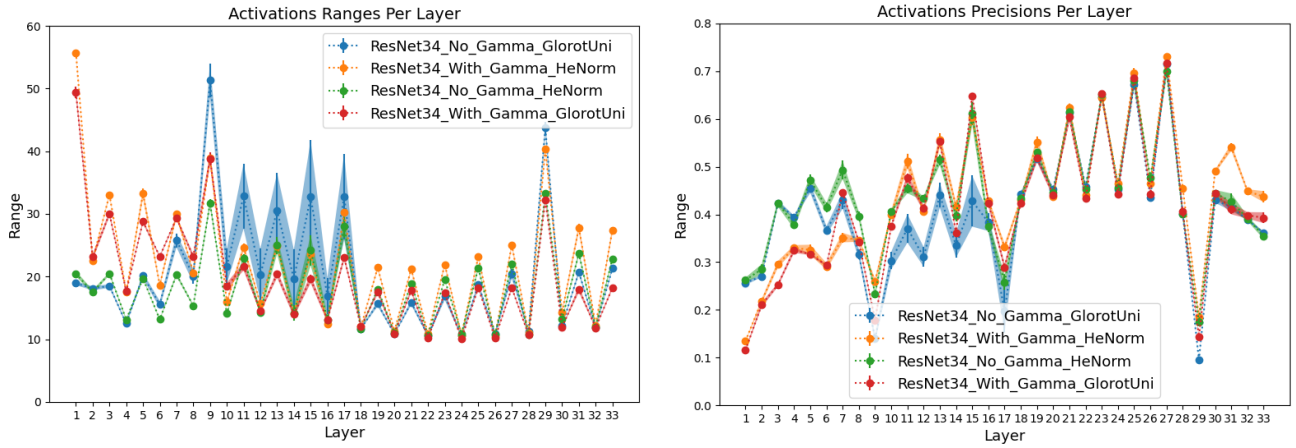

Figure 5. Layerwise activation range (left) and average precision (right) for ResNet-34.

Network Architecture	FP32 Acc (%)	QUINT8 Acc (%)	QMSE	QCE	QKL-Div	Percent Acc Decrease
MobileNet No Gamma HeNorm	78.63	65.92 +/- 5.76	0.0266 +/- 0.011	1.08 +/- 0.45	0.76 +/- 0.43	16.16 +/- 7.32
MobileNet No Gamma GlorotUni	78.73	71.88 +/- 1.79	0.0186 +/- 0.0029	0.77 +/- 0.065	0.45 +/- 0.065	8.70 +/- 2.28
MobileNet With Gamma HeNorm	78.65	66.11 +/- 4.51	0.0278 +/- 0.0076	1.17 +/- 0.40	0.84 +/- 0.37	15.95 +/- 5.74
MobileNet With Gamma GlorotUni	78.46	65.87 +/- 2.36	0.0273 +/- 0.0046	1.05 +/- 0.13	0.72 +/- 0.13	16.05 +/- 3.01
DWS Conv No Gamma HeNorm	79.48	65.89 +/- 8.29	0.0236 +/- 0.012	1.03 +/- 0.33	0.60 +/- 0.33	17.10 +/- 10.43
DWS Conv No Gamma GlorotUni	78.63	68.93 +/- 2.90	0.0187 +/- 0.0036	0.88 +/- 0.11	0.48 +/- 0.11	12.34 +/- 3.69
DWS Conv With Gamma HeNorm	80.16	70.72 +/- 2.91	0.0187 +/- 0.0040	0.84 +/- 0.095	0.45 +/- 0.095	11.78 +/- 3.63
DWS Conv With Gamma GlorotUni	78.17	45.43 +/- 20.53	0.0592 +/- 0.037	2.51 +/- 1.34	2.04 +/- 1.34	41.88 +/- 26.26
Regular Conv No Gamma HeNorm	87.27	78.29 +/- 2.67	0.0170 +/- 0.0045	0.66 +/- 0.13	0.46 +/- 0.13	10.29 +/- 3.06
Regular Conv No Gamma GlorotUni	86.63	83.32 +/- 0.75	0.00764 +/- 0.0016	0.37 +/- 0.044	0.18 +/- 0.044	3.83 +/- 0.87
Regular Conv With Gamma HeNorm	88.01	80.26 +/- 0.77	0.0152 +/- 0.0013	0.55 +/- 0.022	0.38 +/- 0.022	8.80 +/- 0.88
Regular Conv With Gamma GlorotUni	86.58	81.15 +/- 2.17	0.0113 +/- 0.0044	0.46 +/- 0.11	0.26 +/- 0.11	6.27 +/- 2.51
ResNet34 No Gamma HeNorm	84.93	80.91 +/- 0.73	0.0123 +/- 0.0012	0.45 +/- 0.039	0.30 +/- 0.038	4.73 +/- 0.86
ResNet34 No Gamma GlorotUni	84.34	77.04 +/- 2.42	0.0181 +/- 0.0048	0.65 +/- 0.15	0.48 +/- 0.15	8.66 +/- 2.87
ResNet34 With Gamma HeNorm	85.35	83.79 +/- 0.65	0.00820 +/- 0.0021	0.35 +/- 0.056	0.20 +/- 0.056	1.83 +/- 0.76
ResNet34 With Gamma GlorotUni	85.67	78.95 +/- 1.88	0.0166 +/- 0.0031	0.60 +/- 0.12	0.44 +/- 0.12	7.84 +/-2.20

Table 1. Detailed quantization results for each network trained on CIFAR-10. Quantization results are reported as mean and standard deviation across five different quantization trials. The output QMSE, QCE and QKL-Div of depthwise-separable convolution based networks are noticeably higher than regular CNNs in most cases

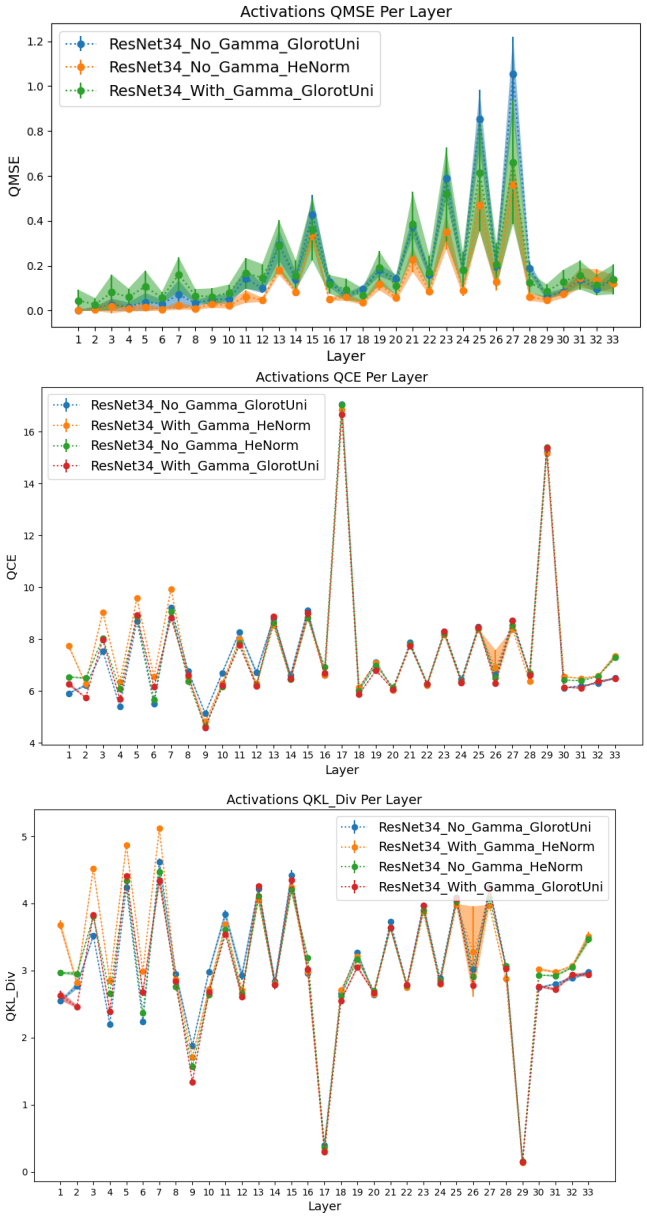

Figure 6. Layerwise QMSE (top) QCE (center) and QKL-Div (bottom) for ResNet-34. The solid line represents the average values across 5 quantization trials and the shaded region is the standard devation. Note, one outlier removed due to large QMSE scale.

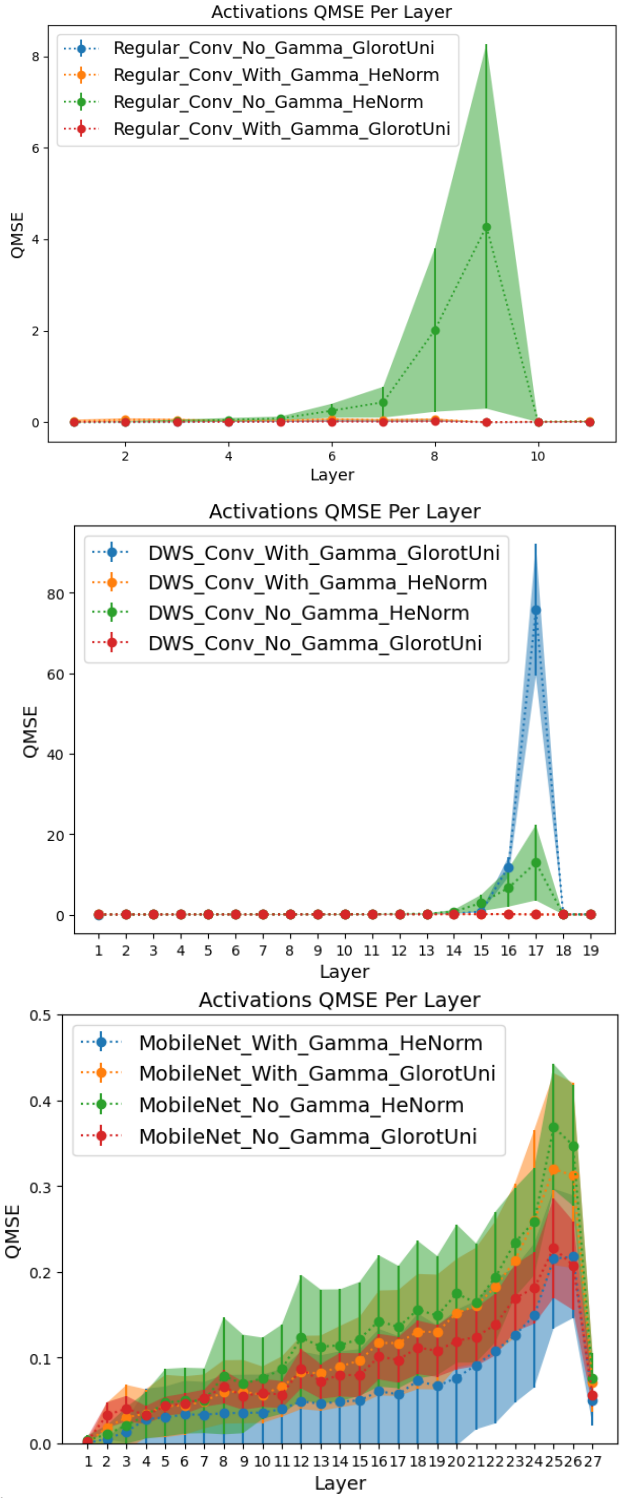

Figure 7. Layerwise QMSE for all trained networks. Regular-ConvNets (top), DWS-ConvNets (center), and MobileNets-V1 (bottom). The significant spike for DWS_Conv_With_Gamma_GlorotUni explain the major degradation in quantized performance. **Note** the difference in y-axis scales for Regular-ConvNets and DWS-ConvNets. The solid line represents the average values across 5 quantization trials and the shaded region is the standard deviation.

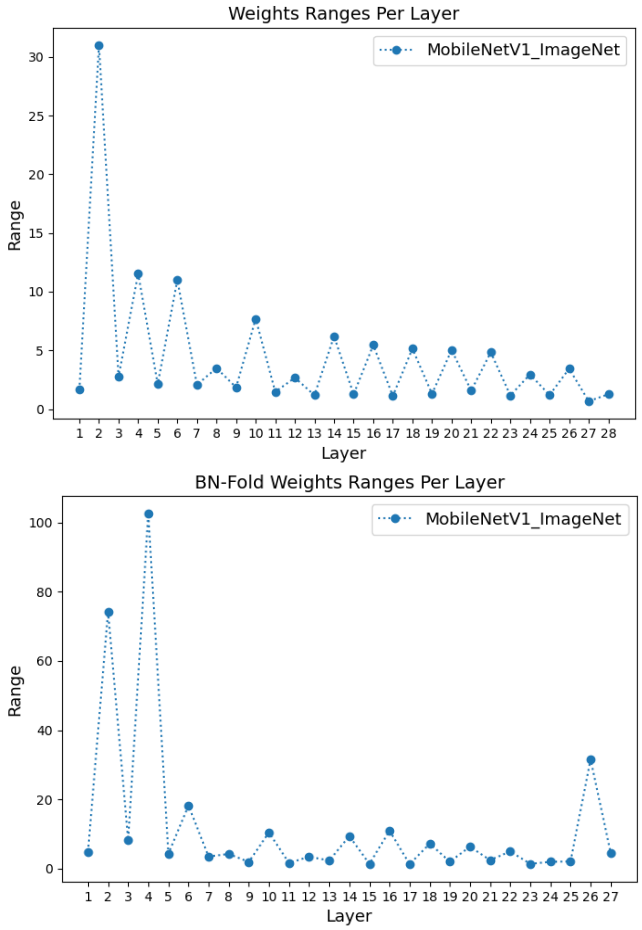

Figure 8. Layerwise weights range (top) and BN-Folded weights range (bottom). Observing the BN-Folding induced distributional shift can also yield interesting insights on the scaling required at each layer. Note the very different y-axis scales.

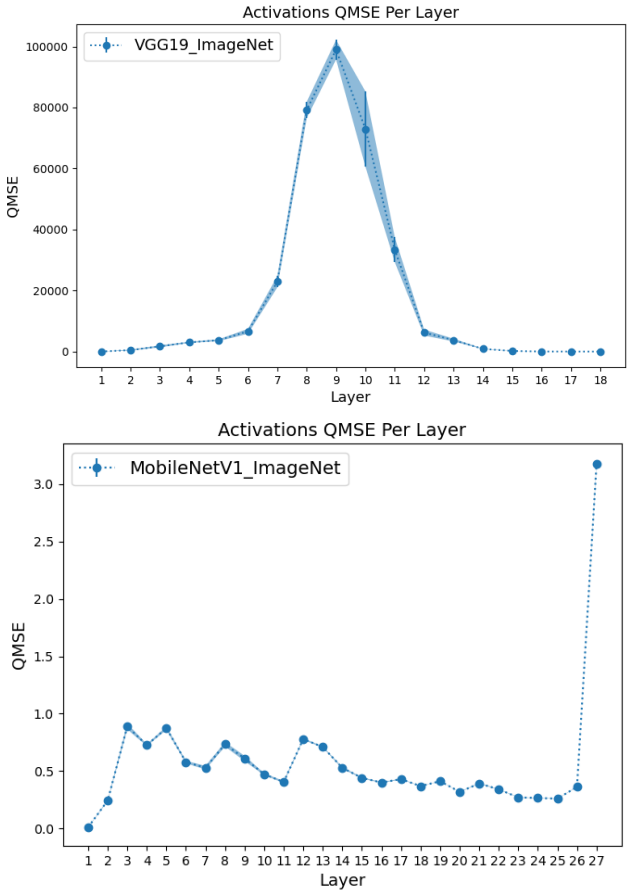

Figure 9. Layerwise QMSE for VGG-19 (top) and MobileNet-V1 (bottom) trained on ImageNet. Note the difference in y-axis scales.

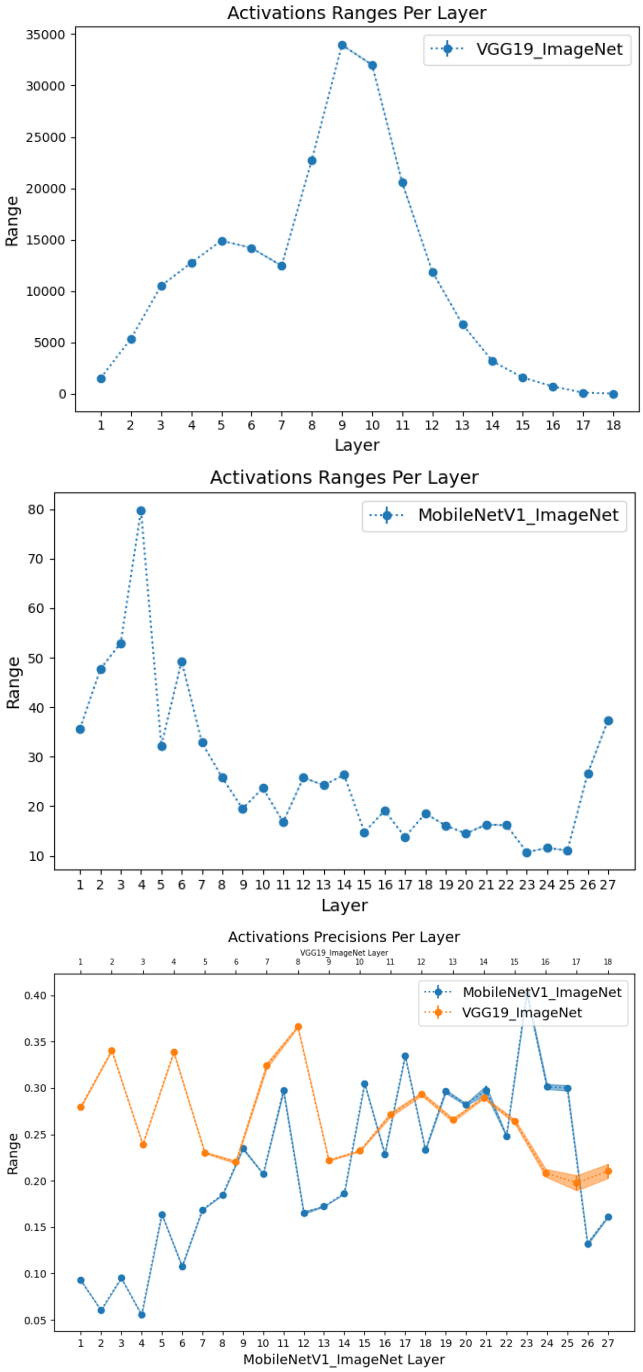

Figure 10. Layerwise activation ranges for VGG-19 (top) and MobileNet-V1 (center) trained on ImageNet. Layerwise activation precisions for VGG-19 vs. MobileNet-V1 trained on ImageNet (bottom).