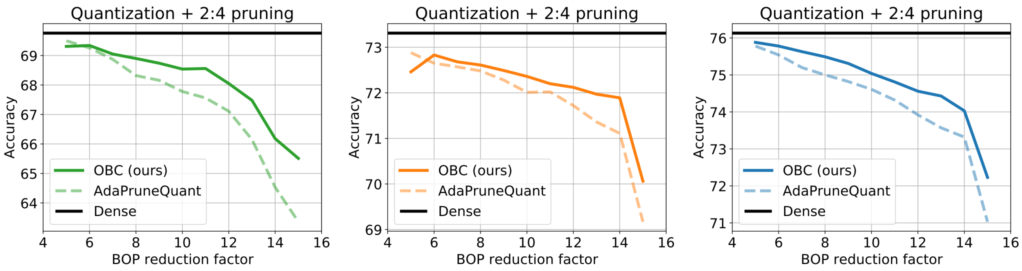
A.10 Compound Compression Comparisons

In the main paper, we focused on independent comparisons for quantization and pruning since existing methods are generally only designed for a single compression approach. In this section, we additionally provide compound comparisons for our GPU and CPU scenarios which combine sparsity and quantization. In particular, we construct a strong baseline by substituting OBC in our mixed setup with the best independent layer-wise pruning and quantization methods, AdaPrune and AdaQuant, respectively. We now provide detailed comparisons for all experiments of Figure 2 from the main text, in Figures 4, 5 and 6.

In summary, it appears that, as expected, the accuracy improvements for the individual compression types shown by the experiments in Section 6 also transfer to the combined setting. More concretely, for the reduction target ranges highlighted in the main paper, that is $12 - 14\times$ for ResNet models and $7 - 8\times$ for others, there is a consistent $0.5 - 1.5$ point gap between OBC and the AdaPruneQuant baseline. For lower BOP reduction / inference time speedup targets, the gap is typically smaller, which is expected as only the less sensitive layers have to compressed more than to the generally very easy 8-bit level. In contrast, the gaps are largest for the highest targets that also require high compression of sensitive layers as this is where the effects of OBC's more accurate layer-wise compression become particularly noticeable.

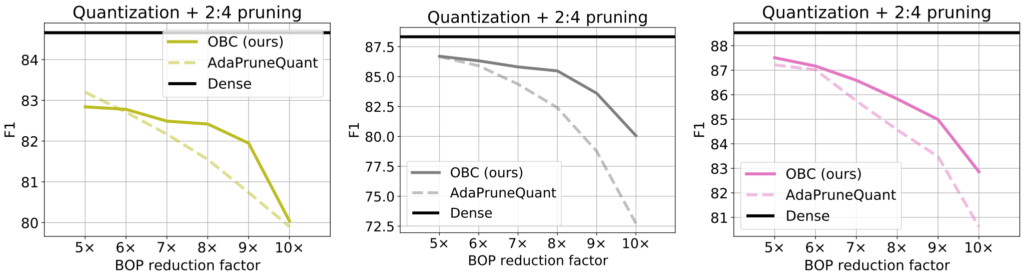


(a) ResNet18.

(b) ResNet34.

(c) ResNet50.

Figure 4: Mixed quantization and 2:4 pruning for various BOP reduction targets on ResNet models.

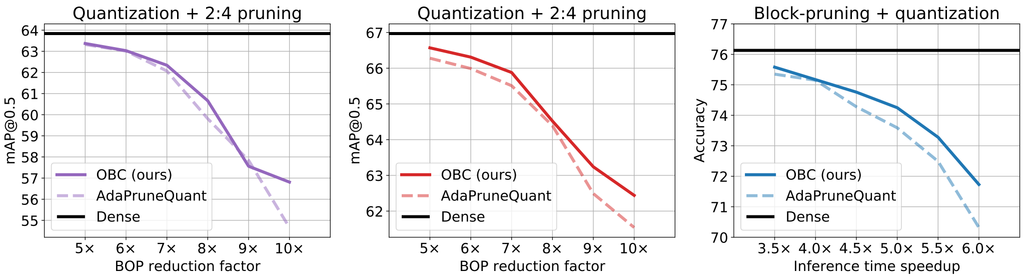


(a) BERT3.

(b) BERT6.

(c) BERT.

Figure 5: Mixed quantization and 2:4 pruning for various BOP reduction targets on BERT models.



(a) YOLOv5m.

(b) YOLOv5l.

(c) ResNet50 – CPU.

Figure 6: (a) & (b): Mixed quantization and 2:4 pruning for various BOP reduction targets on YOLO models. (c) Block sparsity & quantization for real-time CPU inference speedup targets on ResNet50.