6 Supplementary Materials

Visualization of decisions of DFS-ResNet74

This section visualizes more decision distributions of learned DFS-ResNet74 on CIFAR-10, showing that as computation percentage increases, the DFS framework prefer quantization options to layer skip options.

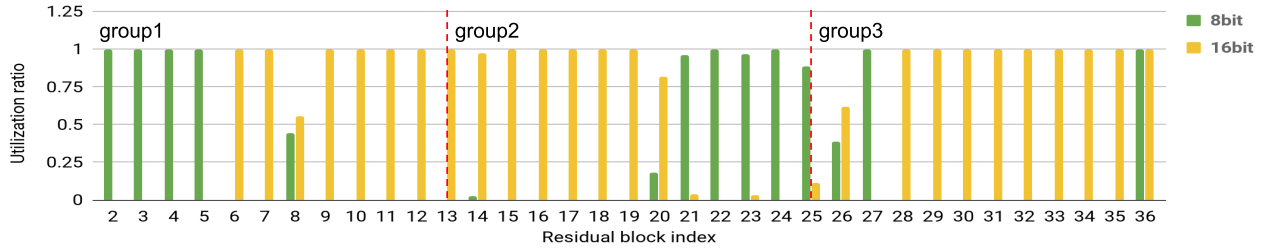

Figure 12: Visualization of layerwise decision distribution of DFS-ResNet74 on CIFAR10: computation percentage = 20%.

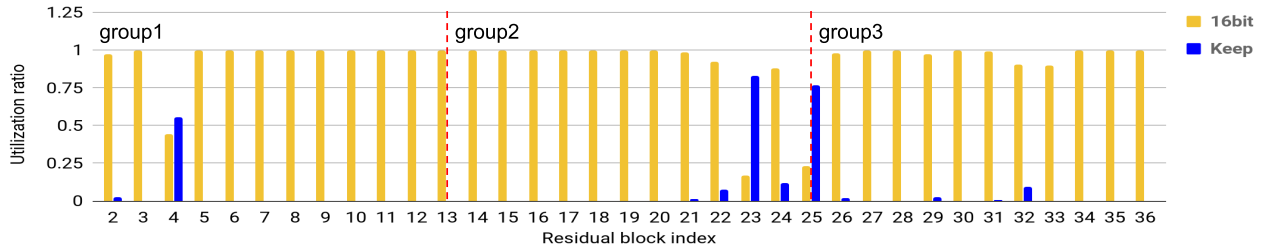

Figure 13: Visualization of layerwise decision distribution of DFS-ResNet74 on CIFAR10: computation percentage = 30%.

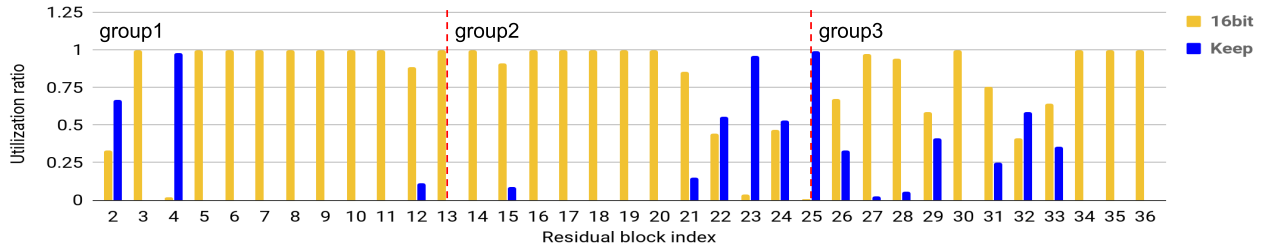

Figure 14: Visualization of layerwise decision distribution of DFS-ResNet74 on CIFAR10: computation percentage = 40%.

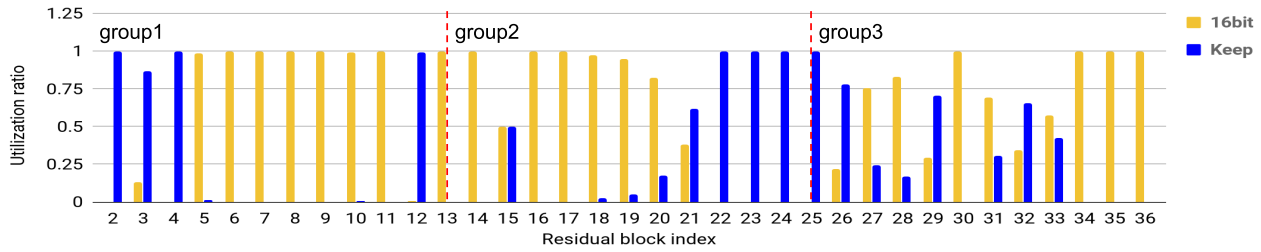

Figure 15: Visualization of layerwise decision distribution of DFS-ResNet74 on CIFAR10: computation percentage = 50%.

Visualization of bitallocation of HAQ-ResNet38 on CIFAR10

This section visualizes the layer-wise bitwidth allocation policy of HAQ-ResNet38 on CIFAR10. The horizontal axis represents specific bitwidth options, and the height of each layer's column indicates which option is used to quantize the layer's weights.

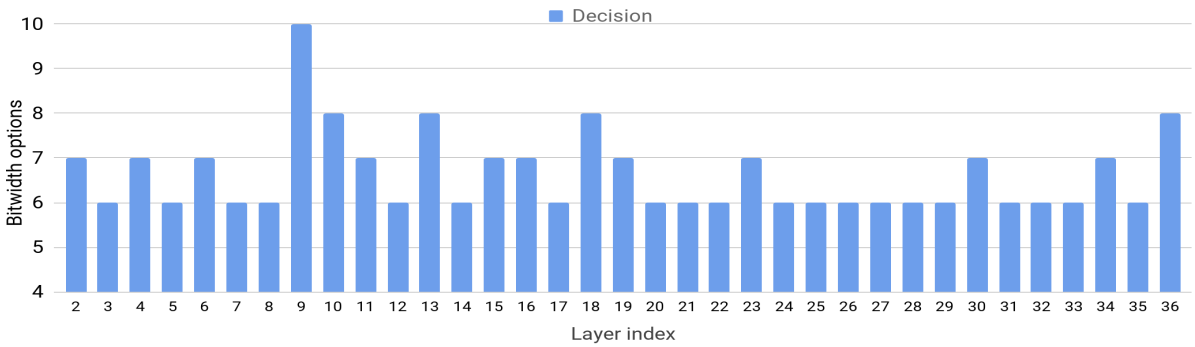

Figure 16: Visualization of layer-wise bitwidht allocation of HAQ method on CIFAR-10 when computation percentage = 20%

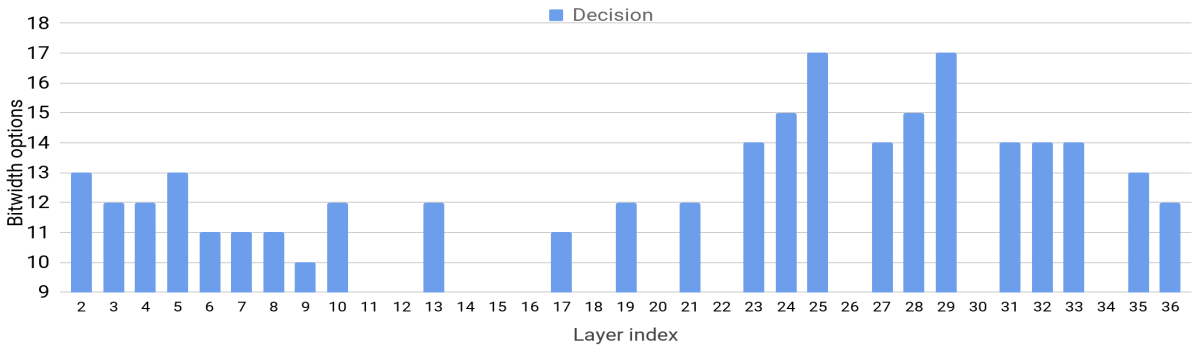

Figure 17: Visualization of layer-wise bitwidht allocation of HAQ method on CIFAR-10 when computation percentage = 30%

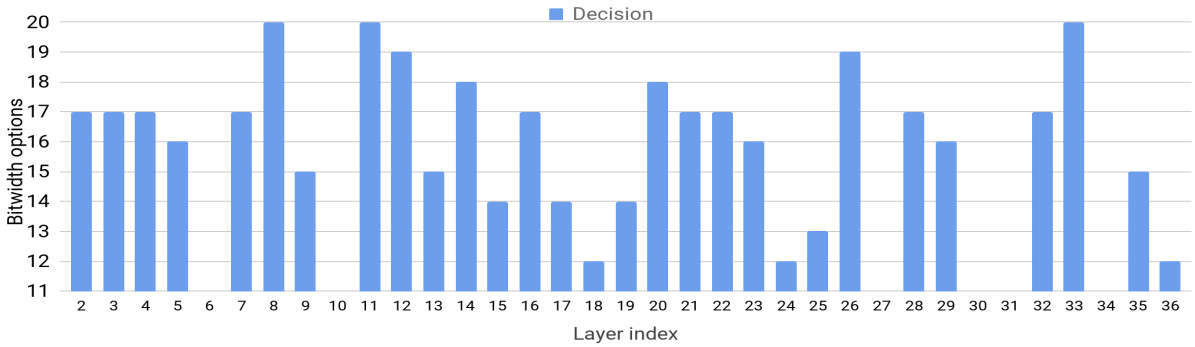

Figure 18: Visualization of layer-wise bitwidht allocation of HAQ method on CIFAR-10 when computation percentage = 50%