Supplementary material for Bit-Mixer: Mixed-precision networks with runtime bit-width selection

Adrian Bulat
Samsung AI Cambridge
adrian@adrianbulat.com

Georgios Tzimiropoulos
Samsung AI Cambridge
Queen Mary University of London
g.tzimiropoulos@qmul.ac.uk

Configuration	avg. bits	Top-1 acc.
$[4, 4, 4, 4, 4, 4, 4, 4]$	4.0	69.04%
$[3, 4, 4, 4, 4, 4, 4, 4]$	3.875	69.02%
$[3, 4, 3, 4, 4, 4, 4, 4]$	3.75	69.04%
$[3, 4, 4, 3, 3, 4, 4, 4]$	3.625	69.02%
$[3, 4, 3, 3, 3, 4, 4, 4]$	3.5	69.01%
$[3, 4, 3, 3, 3, 3, 4, 4]$	3.375	69.20%
$[3, 3, 3, 3, 3, 3, 4, 4]$	3.25	68.94%
$[3, 3, 3, 3, 3, 3, 4, 3]$	3.125	68.73%
$[3, 3, 3, 3, 3, 3, 3, 3]$	3.0	68.47%
$[3, 2, 3, 3, 3, 3, 3, 3]$	2.875	68.22%
$[3, 2, 3, 3, 3, 2, 3, 3]$	2.75	67.80%
$[3, 2, 3, 2, 3, 3, 2, 3]$	2.625	67.22%
$[3, 3, 2, 2, 2, 3, 2, 3]$	2.5	66.41%
$[3, 3, 2, 2, 2, 2, 2, 3]$	2.375	66.06%
$[2, 2, 2, 2, 2, 2, 2, 4]$	2.25	65.76%
$[2, 2, 2, 2, 2, 2, 2, 3]$	2.125	65.61%
$[2, 2, 2, 2, 2, 2, 2, 2]$	2.0	64.48%

Table 1: Configurations of the models presented in Fig. 4, Section 4.2 in the main paper [1], and their corresponding accuracy. Each value on the leftmost column indicates the bit-width of each block inside a ResNet-18 model. Notice that generally the middle blocks are the least sensitive to quantization while the last ones the most. In essence, our method behaves as an implicit network search for the bit-width of each layer/block.

References

[1] Adrian Bulat and Georgios Tzimiropoulos. Bit-mixer: Mixed-precision networks with runtime bit-width selection. *ICCV*, 2021. 1