To address this shortcoming, we propose a method for training binary neural networks with a mixture of bits, yielding effectively fractional bitwidths. We demonstrate that our method is not only effective in allowing finer tuning of the speed to accuracy trade-off, but also has inherent representational advantages.

Due to heavy approximation error, binary models typically have reduced accuracy. However, the accuracy loss can be reduced by using multiple ($n$) bits.

Residual error binarization has each additional bit best reduce remaining approximation error. Excellent for fitting arbitrary distributions, unlike linear binarization.

In residual error binarization, each bit represents a mean of the remaining error. Thus, values close to that mean do not require additional steps, leading to Middle-Out’s excellent performance.

We introduce an algorithm for distributing bitwidths to each individual value of a tensor, which can be considered a high dimension hyperparameter. However, we demonstrate that the Middle-Out sorting algorithm allows bitwidths to be efficiently learned jointly with the parameters of the model, yielding high accuracy inference with fewer bits.

We perform a sweep of training runs with varying weight bitwidths using AlexNet.

Accuracy of HGNs scales super-linearly relative to bitwidth.

1.4 average bits outperforms a 2-bit state of the art model.

All mixtures of bitwidths with the same average have the same final accuracy.

We compare a selection of our results to other state-of-the-art models and confirm that Heterogeneous Binarization transfers to MobileNet.

Fractional bitwidths always compare favorably to higher integer bitwidths.

HBNN provides fine-grained control of the accuracy to speed-up tradeoff.

Transfers well heavily hand-optimized models such as MobileNet, not just AlexNet.

HBNN is especially well suited to FPGA and ASIC systems.

Control of bitwidth allows better use of FPGA space.

Chips can be made to hit a designated accuracy without wasted power or gates.