Supplementary Material: Automatic Neural Network Pruning that Efficiently Preserves the Model Accuracy

Anonymous submission

1 Reduced training data on ILSVRC2012

As discussed in the Pruning Strategy sub-section, AutoBot only requires a small subset of the dataset to converge optimally. This section demonstrates the same phenomena with a supplementary experiment on ILSVRC2012. Fig. 1 shows the accuracy after pruning (before finetuning) and the dissimilarity between filters ranking between two parameters updates when pruning ResNet-50 on ILSVRC2012. Dissimilarity is computed using the normalised Kendall tau distance, which is a common tool to measure dissimilarity between rankings.

In this experiment, we can observe that the accuracy after pruning can converge at around 3000 batches (15.0% of the training dataset), while the dissimilarity between filters is also stable throughout the network.

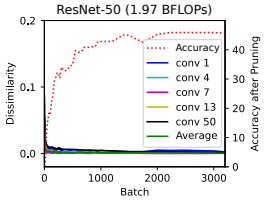

Figure 1: Evolution of accuracy after pruning (before finetuning) and of dissimilarity between filters ranking (normalised Kendall tau distance) when increasing the number of batches, on ILSVRC2012.

2 Pruning process time

Tab. 1 shows the time it takes for the whole pruning process with AutoBot, including the model and data loading, the bottlenecks training, the computation of the optimal threshold, and the actual pruning of the filters. On CIFAR-10, regardless of the architectures' heaviness, AutoBot can provide the optimal pruning ratio layer-by-layer under the target FLOPs within a minute. Even on ILSVRC2012, which requires larger computational times in general, AutoBot can efficiently prune ResNet-50 in around 10 minutes.

To summarize, this table highlights that the model compression is relatively fast thanks to the fast convergence of the bottlenecks.

Dataset	CIFAR-10					ILSVRC2012
Model	VGG-16	ResNet-56	ResNet-110	GoogleNet	DenseNet-40	ResNet-50
GPU hours	0.005	0.009	0.011	0.011	0.013	0.182

Table 1: Pruning process time on NVIDIA RTX 2080 Ti, in GPU hours