Automatic Neural Network Pruning that Efficiently Preserves the Model Accuracy
— SUPPLEMENTARY MATERIALS —

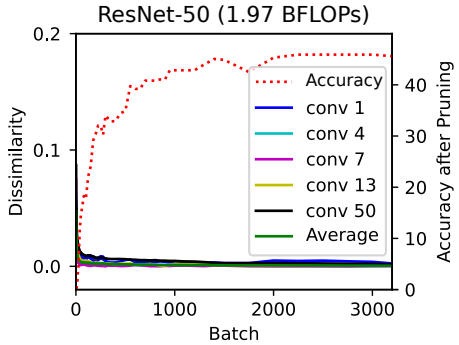
Thibault Castells and Seul-Ki Yeom*

Nota AI GmbH
Friedrichstrasse 200, 10117 Berlin, Germany
thibault@nota.ai, skyeom@nota.ai

A 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. Supplementary 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.



Supplementary 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.

B Pruning process time

Supplementary 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.

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

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

C Hardware specification

Supplementary Tab. 2 summarizes the hardware specifications used for the inference speed tests in our paper.

Supplementary Table 2: The specification of the hardware platforms deployed in our paper.

Platform	CPU	GPU	Memory
Jetson-Nano	Quad core Cortex -A57 @ 1.43GHz	128-core Maxwells^{TM} μA	4GB LPDDR4
Raspberry Pi 4B	Quad core Cortex -A72 @ 1.5GHz	No GPGPU	4GB LPDDR4
Raspberry Pi 3B+	Quad core Cortex -A53 @ 1.4GHz	No GPGPU	1GB LPDDR2
Raspberry Pi 2B	Quad core Cortex -A7 @ 900 MHz	No GPGPU	1GB SDRAM

D Inference speed detailed results

Supplementary Tab. 3 show the detailed results of the inference speed tests on CIFAR10 conducted in this work.

*Corresponding author.
Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Supplementary Table 3: Deployment test on different hardware platforms with our pruned model. Inference time for original model (ms) $\rightarrow$ inference time for pruned model (ms).

		Hardware (Processor)			
Model	FLOPs	Jetson-Nano (GPU)	Raspberry Pi 4B (CPU)	Raspberry Pi 3B+ (CPU)	Raspberry Pi 2B (CPU)
VGG-16	73.71M	61.63 → 13.33 (× **4.62**)	45.73 → 17.16 (× **2.66**)	79.98 → 35.17 (× **2.27**)	351.77 → 118.36 (× **2.97**)
VGG-16	108.61M	61.63 → 13.77 (× **4.48**)	45.73 → 19.95 (× **2.29**)	79.98 → 39.99 (× **2.00**)	351.77 → 143.95 (× **2.44**)
VGG-16	145.55M	61.63 → 19.24 (× **3.20**)	45.73 → 24.33 (× **1.88**)	79.98 → 50.27 (× **1.59**)	351.77 → 184.47 (× **1.91**)
ResNet-56	55.94M	16.47 → 13.71 (× **1.20**)	21.95 → 15.88 (× **1.38**)	60.42 → 39.78 (× **1.52**)	170.46 → 101.70 (× **1.68**)
ResNet-110	85.30M	28.10 → 26.36 (× **1.07**)	41.35 → 27.90 (× **1.48**)	112.57 → 72.71 (× **1.55**)	331.60 → 179.91 (× **1.84**)
GoogLeNet	0.45B	80.84 → 28.37 (× **2.85**)	146.68 → 57.25 (× **2.56**)	342.23 → 170.17 (× **2.01**)	1,197.65 → 400.89 (× **2.99**)
DenseNet-40	129.13M	35.25 → 33.46 (× **1.05**)	71.87 → 44.73 (× **1.61**)	171.86 → 102.75 (× **1.67**)	432.03 → 252.63 (× **1.71**)
DenseNet-40	168.26M	35.25 → 35.11 (× **1.00**)	71.87 → 53.08 (× **1.35**)	171.86 → 114.37 (× **1.50**)	432.03 → 302.49 (× **1.43**)