Supplementary Material for the Paper: Does Interference Exist When Training a Once-For-All Network?

Jordan Shipard[1], Arnold Wiliem[2], and Clinton Fookes[1]

[1]Signal Processing, Artificial Intelligence and Vision Technologies (SAIVT), Queensland University of Technology, Australia
[2]Sentient Vision Systems, Australia
{jordan.shipard@hdr., c.fookes@}qut.edu.au, arnoldw@sentientvision.com

1. RSS-Short Ablation Results

We conducted additional ablation experiments to study the effect of the number of subnets sampled and the selection scheme on the proposed RSS-Short method. As mentioned earlier, RSS trains for 590 epochs and RSS-Short trains for 180 epochs. Other than this, all RSS-Short experiments shared the same hyperparameters as their RSS counterparts. All experiments are conducted on the CIFAR100 dataset [3].

1.1. Effect of Number of Subnets Sampled

Fig. 1 shows the results from comparing per epoch to per batch sampling during training. These results follow the results presented in the paper with a decrease in population accuracy between per epoch and per batch methods. A further drop occurs when training two subnets per batch and combining their gradients.

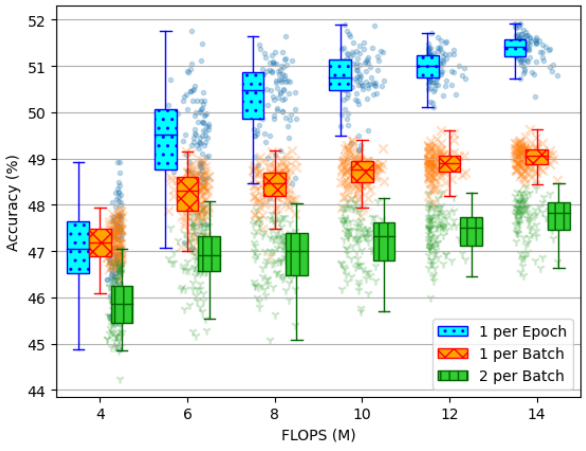


Figure 1. RSS-Short (proposed) results for sampling different numbers of subnets during training. The same pattern can be observed as in the main ablation results with per batch sampling and combined gradients performing worse than per epoch sampling.

1.2. Effect of Different Selection Schemes

Figs. 2 and 3 show the results from altering the subnet selection scheme during training. In Fig. 2, the selection schemes tested are **Smallest Only**, **Middle Only** and **Largest Only**, where the smallest, middle and largest subnets train for the entire 180 epochs. These results again follow the results presented in the paper. Each single subnet selection scheme shows a heavy bias towards the trained subnet.

Fig. 3 shows the two subnet selection scheme with different training orders. The two subnets trained are the largest (**max**) and smallest (**min**). **Max then Min** trains the largest subnet for the first 90 epochs and the smallest subnet for the remaining 90 epochs. **Min then Max** does the opposite, training the smallest then the largest. **Alternating** switches between the two subnets, training each for one epoch at a time. The results in Fig. 3 differ slightly from the RSS results presented in the paper. For **Max then Min**, min has

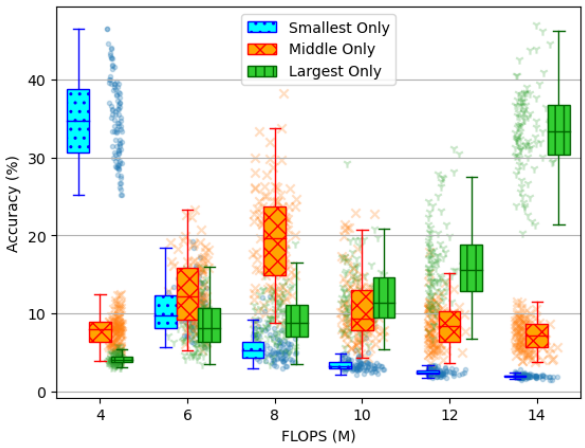


Figure 2. RSS-Short (proposed) population results from training a single subnet only. The results are the same as in the main ablation results.

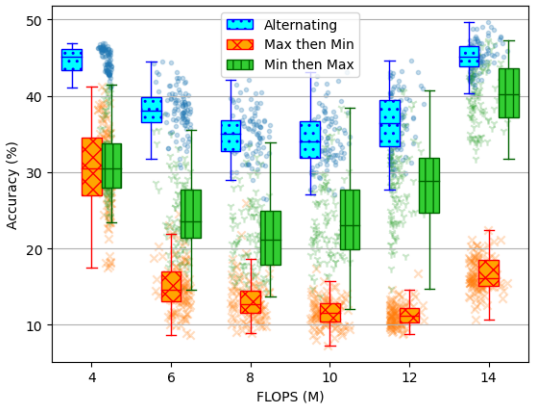


Figure 3. RSS-Short (proposed) population results from training the same largest and smallest subnets in different sequences. These results show a more significant bias towards the largest and smallest subnets than the main ablation results.

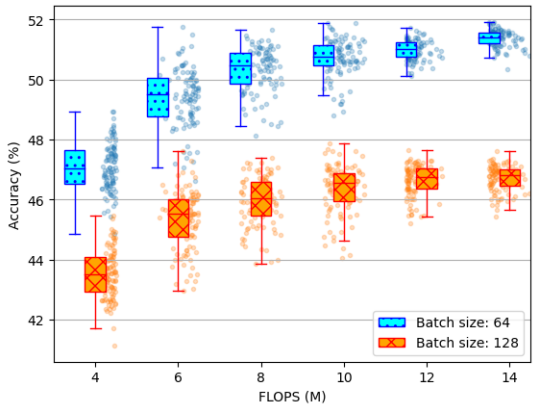
better performance than max; for **Min then Max**, max has better performance than min; and for **Alternating**, min and max have similar performance. The accuracy of subnets in the 8-10 MFLOP range trained via the **Alternating** selection scheme perform significantly worse than smaller and larger subnets. This result is not seen in the main results, suggesting that additional epochs are required to reduce the bias in the subnet selection scheme.

2. Hyperparameter Ablation Studies

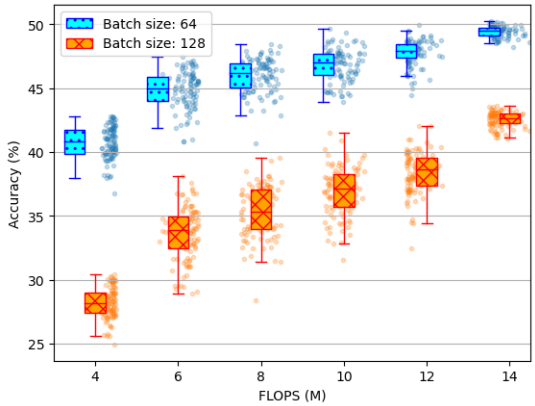
We alter various hyperparameters to ensure our findings are not limited to specific hyperparameter values. We adjust the batch size, dropout rate and learning rate, comparing the effect on RSS-Short and OFA. Fig. 4 shows that increasing the batch size from 64 to 128 results in a overall accuracy drop for both RSS-Short and OFA. Fig. 5 shows that increasing the dropout rate from 0.1 to 0.3 again results in a decrease in accuracy for both methods, with a more significant decrease for OFA. Lastly, Fig. 6 shows that increasing the learning rate from 0.01 to 0.02 reduces the accuracy of both methods. These results show that hyperparameter changes have the same effect for both methods. Confirming our results are not limited to specific hyperparameter values.

3. ProxylessNAS Base Architecture

We change the base architecture used during training, showing that results obtained are not unique to the MobileNetV3 [2] base architecture. Fig. 7 shows this as we switch to using a ProxylessNAS [1] base architecture and achieve similar results. The population settings remain the same as before; however, the resulting subnet population only ranges from 4 MFLOPs to 12 MFLOPs. Therefore, we only show subnets from sizes 4, 6, 8, 10 and 12 MFLOPs.



(a) Resulting RSS-Short (proposed) subnet populations.



(b) Resulting OFA subnet populations.

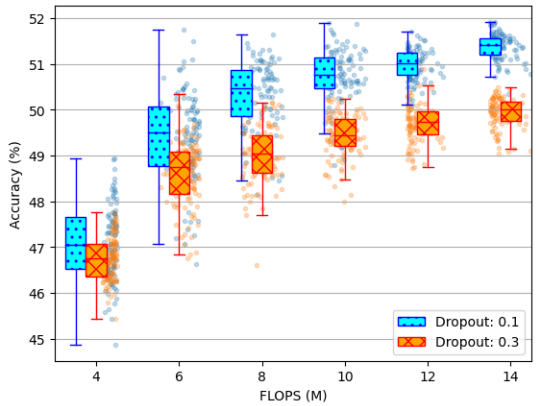
Figure 4. Resulting RSS-Short (proposed) and OFA subnet populations from increasing the batch size during training. These results show that the batch size has the same effect for both methods.

References

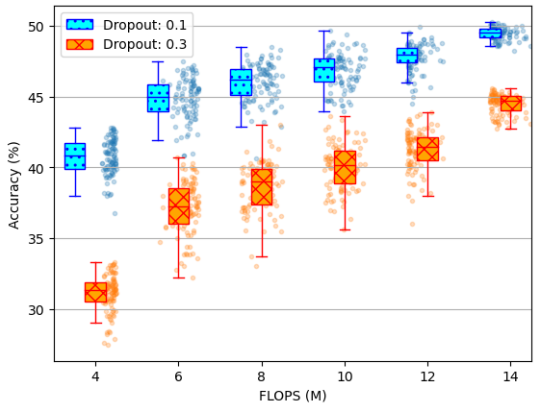
[1] Han Cai, Ligeng Zhu, and Song Han. ProxylessNAS: Direct neural architecture search on target task and hardware. In *International Conference on Learning Representations*, 2019. 2, 3

[2] Andrew G. Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324, 2019. 2, 3

[3] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. 1, 3

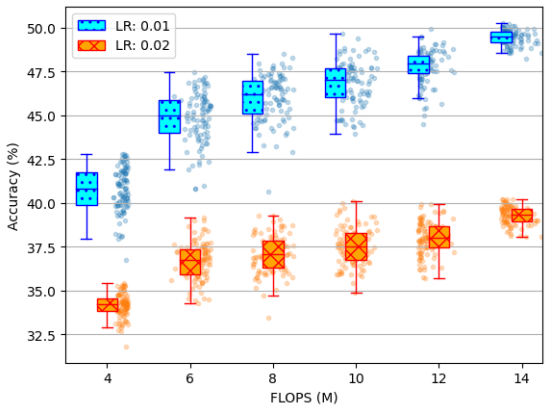


(a) Resulting RSS-Short (proposed) subnet populations.

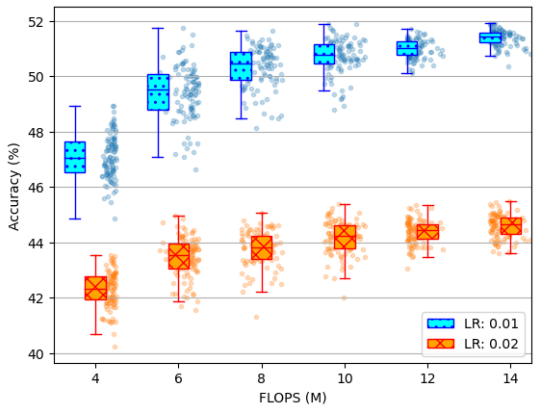


(b) Resulting OFA subnet populations.

Figure 5. Resulting RSS-Short (proposed) and OFA subnet populations from increasing the drop-out rate. These results show that both methods suffer from an increased dropout rate with OFA having a more significant accuracy drop.



(a) Resulting RSS-Short (proposed) subnet populations.



(b) Resulting OFA subnet populations.

Figure 6. Resulting RSS-Short (proposed) and OFA subnet populations from increasing the learning rate. These results show that the learning rate decreases the accuracy of both methods.

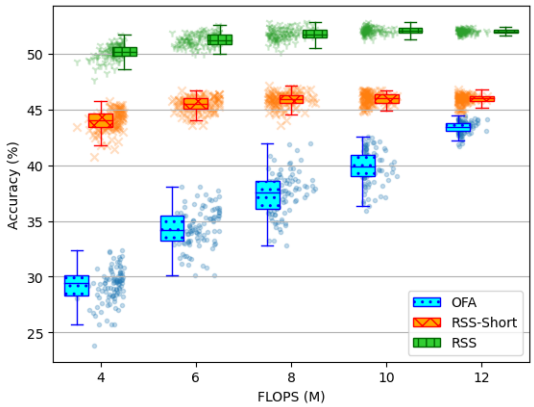


Figure 7. Resulting OFA, RSS-Short and RSS subnet populations from training with a ProxylessNAS [1] base architecture on CIFAR100 [3]. These results are consistent with the main results despite a lower overall accuracy for each method than with the MobileNetV3 [2] base architecture.