A. Details of ImageNet Experiments

When porting our CIFAR-10 models to work on ImageNet, we made the following changes:

- The input image size was 224×224 rather than 32×32.
- We used eight cells instead of six. As for CIFAR-10, we halved the image height/width and doubled the number of filters after the second, fourth, and sixth cells.
- Instead of using a 3x3 convolution for the model stem, we used a 7x7 convolution followed by a 3x3 max pooling layer. Both had stride 2.

After making these changes, we re-ran our calibration experiments. Correlations between a trained one-shot model and stand-alone model architectures trained on ImageNet for 6 epochs are shown in Figure 8. Even more than for CIFAR-10, we see a strong correlation between one-shot accuracies and ImageNet accuracies after a shortened training period.

Experiments on this dataset were reported in Table 2. When we compared our models to previously published mobile results such as Howard et al. (2017) and Zoph et al. (2017), we found that although our models had higher accuracies for the same number of parameters, they also had more MAC (Multiply-Add) operations. In order to make our results more directly comparable, we made the following changes to our search space:

- Instead of using a 1x1 convolution followed by a stride-2 max-pooling layer to reduce the image height/width and double the number of filters, we used a strided 1x1 convolution.
- We changed the model stem to use a single stride-2 depthwise separable convolution with a 7x7 kernel.
- Following Zoph et al. (2017), we decreased the number of MACs in the first two cells of the model. In our initial ImageNet model, the first two cells had height/width 56 and F filters. In our updated model, the first cell had 112×112 input images and $F/4$ filters, while the second had 56×56 input images and $F/2$ filters.

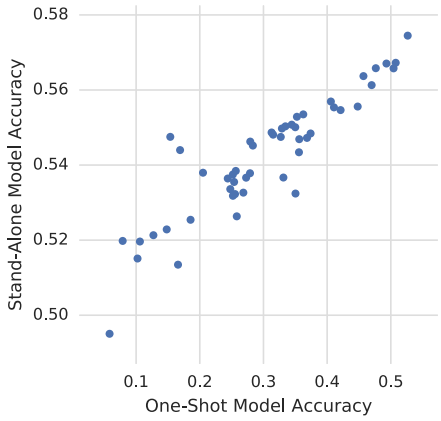

Figure 8. Correlation between one-shot and stand-alone model accuracies on ImageNet.

After making these changes, we re-evaluated the top cells found by our previous ImageNet search. The results are shown in Table 3. The general trends are the same as for our previous experiment. The architectures with the highest one-shot accuracies have higher stand-alone accuracies than random architectures, but also more parameters. The "small" models show significantly better trade-offs between the two.

Method	Param $\times 10^6$	Accuracy
One-Shot Optimized for FLOPS Top ($F = 16$)	3.1 ± 0.4	66.2 ± 1.0
One-Shot Optimized for FLOPS Top ($F = 24$)	6.8 ± 0.9	70.7 ± 0.6
One-Shot Optimized for FLOPS Top ($F = 32$)	11.9 ± 1.5	72.6 ± 0.4
One-Shot Optimized for FLOPS Small ($F = 16$)	1.4 ± 0.4	63.4 ± 0.8
One-Shot Optimized for FLOPS Small ($F = 24$)	2.9 ± 0.8	68.9 ± 0.5
One-Shot Optimized for FLOPS Small ($F = 32$)	5.0 ± 1.4	71.2 ± 0.5
Random Optimized for FLOPS ($F = 16$)	2.0 ± 0.5	63.3 ± 1.6
Random Optimized for FLOPS($F = 24$)	4.4 ± 1.0	68.5 ± 1.3
Random Optimized for FLOPS($F = 32$)	7.6 ± 1.9	70.9 ± 1.0

Table 3. Evaluation of Architecture Search Results on ImageNet after tuning to reduce MACs.