Progressive Neural Architecture Search (Supplementary Material)

Chenxi Liu[1⋆], Barret Zoph[2], Maxim Neumann[2], Jonathon Shlens[2], Wei Hua[2], Li-Jia Li[2], Li Fei-Fei[2,3], Alan Yuille[1], Jonathan Huang[2], and Kevin Murphy[2]

[1] Johns Hopkins University
[2] Google AI
[3] Stanford University

A Search Efficiency of PNAS with RNN-ensemble

In Section 5.3 of the paper, we focused on the performance of MLP-ensemble as the surrogate model. Here we provide analysis of RNN-ensemble as well.

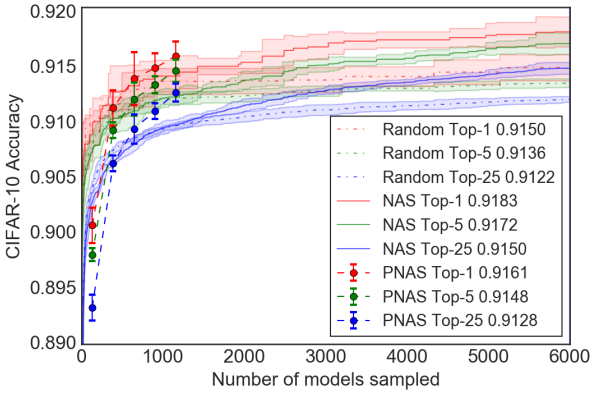

Fig. 1. Comparing the relative efficiency of PNAS (using RNN-ensemble) with NAS and random search under the same search space.

B	Top	Accuracy	# PNAS	# NAS	Speedup (# models)	Speedup (# examples)
5	1	0.9161	1160	2222	1.9	5.1
5	5	0.9148	1160	2489	2.1	5.4
5	25	0.9128	1160	2886	2.5	5.7

Table 1. Relative efficiency of PNAS (using RNN-ensemble predictor) and NAS under the same search space.

⋆ Work done while an intern at Google.

Again, each method is repeated 5 times to reduce the randomness in neural architecture search, and both performance mean and the variance are plotted in Figure 1. A more quantitative breakdown is given in Table 1. We see that PNAS with RNN-ensemble is about twice as efficient than NAS in terms of number of models trained and evaluated, and five times as efficient by the number of examples. Speedup measured by number of examples is greater than speedup in terms of number of models, because NAS has an additional reranking stage, that trains the top 250 models for 300 epochs each before picking the best one.

B Searching Cells with More Blocks

Using the MLP-ensemble predictor, we tried to continue the progressive search beyond cells with 5 blocks, all the way till $B = 10$. The result of this experiment is visualized in Figure 2, which extends Figure 4 of the main paper. As can be seen, PNAS is able to find good performing models over the much larger search spaces of $B > 5$. Note that the unconstrained search space size increases by about 4 orders of magnitude for every B level, reaching $\sim 10^{33}$ possible model configurations at $B = 10$. This is one of the main advantages of PNAS, to examine a highly focused search space of arbitrary size progressively. Notice that the NAS curve for comparison is still for $B = 5$, and if we search cells with more blocks using NAS, this curve is likely to go down, because of the growth in search space.

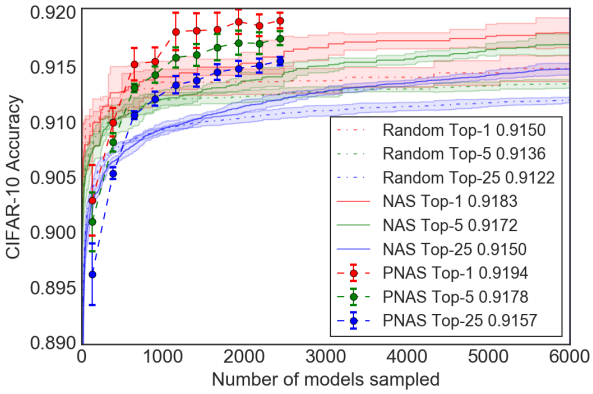

Fig. 2. Running PNAS (using MLP-ensemble) from cells with 1 block to cells with 10 blocks.

C Intermediate Level PNASNet Models

Our Progressive Neural Architecture Search algorithm explores cells from simple to complex by growing the number of blocks. We choose $B = 5$, and indeed the

best model found in the final level (PNASNet-5; visualized in the left plot of Figure 1) demonstrates state-of-the-art performance. In this section, however, we are interested in the best models found in smaller, intermediate levels, namely $b = 1, 2, 3, 4$. We call these models PNASNet-{1, 2, 3, 4}.

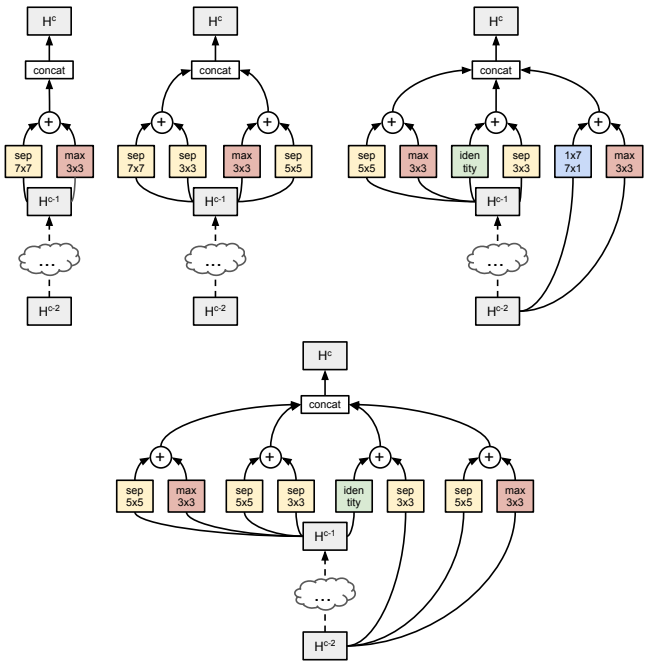

Fig. 3. Cell structures used in PNASNet-{1, 2, 3, 4}.

Model	B	N	F	Error	Params	M_1	E_1	M_2	E_2	Cost
PNASNet-4	4	4	44	3.50±0.10	3.0M	904	0.9M	0	0	0.8B
PNASNet-3	3	6	32	3.70±0.12	1.8M	648	0.9M	0	0	0.6B
PNASNet-2	2	6	32	3.73±0.09	1.7M	392	0.9M	0	0	0.4B
PNASNet-1	1	6	44	4.01±0.11	1.6M	136	0.9M	0	0	0.2B

Table 2. Image classification performance on CIFAR test set. "Error" is the top-1 misclassification rate on the CIFAR-10 test set. (Error rates have the form $\mu \pm \sigma$, where μ is the average over multiple trials and σ is the standard deviation. In PNAS we use 15 trials.) "Params" is the number of model parameters. "Cost" is the total number of examples processed through SGD ($M_1 E_1 + M_2 E_2$) before the architecture search terminates.

We visualize their cell structures in Figure 3, and report their performances on CIFAR-10 in Table 2. We see that the test set error rate decreases as we progress from $b = 1$ to $b = 5$, and the performances of these PNASNets with smaller number of blocks are still competitive.

D Transferring from CIFAR-10 to ImageNet

Figure 4 shows that the accuracy on CIFAR-10 (even for models which are only trained for 20 epochs) is strongly correlated with the accuracy on ImageNet, which proves that searching for models using CIFAR-10 accuracy as a fast proxy for ImageNet accuracy is a reasonable thing to do.

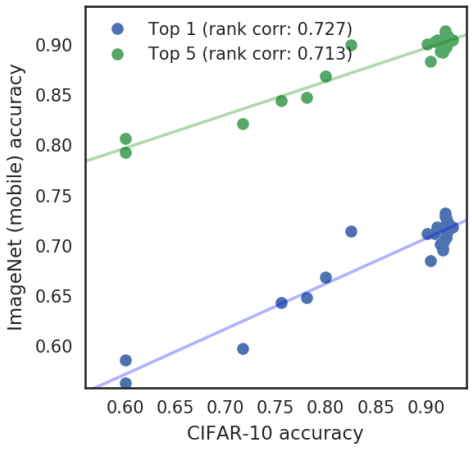

Fig. 4. Relationship between performance on CIFAR-10 and ImageNet for different neural network architectures. The high rank correlation of 0.727 (top-1) suggests that the best architecture searched on CIFAR-10 is general and transferable to other datasets. (Note, however, that rank correlation for the higher-value points (with CIFAR score above 0.89) is a bit lower: 0.505 for top-1, and 0.460 for top-5.)