EcoNAS: Finding Proxies for Economical Neural Architecture Search

Dongzhan Zhou\textsuperscript{1*} Xinchi Zhou\textsuperscript{1*} Wenwei Zhang\textsuperscript{2} Chen Change Loy\textsuperscript{2} Shuai Yi\textsuperscript{3} Xuesen Zhang\textsuperscript{3} Wanli Ouyang\textsuperscript{1}
\textsuperscript{1}The University of Sydney, SenseTime Computer Vision Research Group, Australia \textsuperscript{2}Nanyang Technological University \textsuperscript{3}SenseTime Research
\{d.zhou, xinchi.zhou1, wanli.ouyang\}@sydney.edu.au \{wenwei001, ccloy\}@ntu.edu.sg \{yishuai, zhangxuesen\}@sensetime.com

Abstract

Neural Architecture Search (NAS) achieves significant progress in many computer vision tasks. While many methods have been proposed to improve the efficiency of NAS, the search progress is still laborious because training and evaluating plausible architectures over large search space is time-consuming. Assessing network candidates under a proxy (i.e., computationally reduced setting) thus becomes inevitable. In this paper, we observe that most existing proxies exhibit different behaviors in maintaining the rank consistency among network candidates. In particular, some proxies can be more reliable – the rank of candidates does not differ much comparing their reduced setting performance and final performance. In this paper, we systematically investigate some widely adopted reduction factors and report our observations. Inspired by these observations, we present a reliable proxy and further formulate a hierarchical proxy strategy. The strategy spends more computations on candidate networks that are potentially more accurate, while discards unpromising ones in early stage with a fast proxy. This leads to an economical evolutionary-based NAS (EcoNAS), which achieves an impressive 400× search time reduction in comparison to the evolutionary-based state of the art [19] (8 vs. 3150 GPU days). Some new proxies led by our observations can also be applied to accelerate other NAS methods while still able to discover good candidate networks with performance matching those found by previous proxy strategies.

1. Introduction

Neural Architecture Search (NAS) has received wide attention and achieved significant progress in many computer vision tasks, such as image classification [9, 19, 26, 31], detection [6, 8, 24], and semantic segmentation [13]. Although recent NAS methods [14, 16, 18] improve the search efficiency from earlier works [30], the search progress is still time-consuming and requires vast computation overhead when searching in a large search space since all network candidates need to be trained and evaluated.

A widely adopted approach to alleviating this problem is by training and evaluating network candidates under proxies (i.e., computationally reduced settings [16, 18, 19, 31]). There are four common reduction factors, namely, the number of channels for CNNs (c), the resolution of input images (r), the number of training epochs (e), and the sample ratio (s) of the full training set. These reduction factors either reduce the computation costs of networks or reduce the training iterations to save the search time. While introducing proxies sounds appealing, a caveat is that different reduction factors exhibit different behaviors in keeping the rank consistency among network candidates. We observe that some architectures applied with certain reduction
factors win in the reduced setting but perform worse in the original setup on CIFAR-10 [12]. Some other reduction factors perform more consistently. Those consistent and efficient reduced settings accelerate existing NAS methods and achieve competitive results as shown in Fig. 1.

In this work, we investigate the behaviors of reduction factors \(c, r, s, e\). Two main observations were obtained: (1) with the same iteration numbers, using more training samples with fewer training epochs could be more effective than using more training epochs and fewer training samples in reducing the rank inconsistency; (2) reducing the resolution of input images is sometimes feasible while reducing the channels of networks is more reliable than reducing the resolution. The aforementioned observations motivate us to design reliable proxies that reduce channels and input resolutions while using all training samples. These new proxies apply well to many NAS methods, including evolutionary-based and gradient-based NAS methods [5, 16, 19]. In particular, we observe consistent search time reduction when these new proxies are applied, while the discovered networks are similarly competitive to those found by the original NAS methods.

We further formulate a novel hierarchical proxy strategy for evolutionary-based NAS. The goal of the hierarchical proxy strategy is to discard less promising candidates earlier with a faster proxy and evaluate more promising candidates with a more expensive proxy. The strategy is both effective and efficient: (1) this design saves substantial computation overhead by saving evaluation on less promising networks, and (2) assigning more resources to more promising networks help us to find good architectures more precisely. Thanks to the hierarchical proxy strategy, the proposed economical evolutionary-based NAS (EcoNAS) enjoys nearly 400× reduced search time (8 vs. 3150 GPU days) in comparison to the evolutionary-based state of the art [19] while maintaining comparable performance.

To summarize, our main contributions are as follows:
1. We conduct extensive experiments to study commonly applied reduction factors. While inconsistency of reduction factors is well-known in the community, our work presents the first attempt to analyze the behaviour systematically.
2. Observations from the experiments lead to some fast and reliable proxies that are applicable to a wide range of NAS methods, including evolutionary-based NAS [19] and gradient-based NAS methods [5, 16].
3. We present a hierarchical proxy strategy that leads to EcoNAS, reducing the search time requirement from thousands of GPU days to fewer than 10 GPU days.

2. Related Work

Neural Architecture Search. NAS aims at searching for good neural networks automatically in an elaborately designed search space. One common approach of NAS is to train and evaluate each network candidates on a proxy, i.e., computation reduced setting, and to search architectures using either evolutionary algorithms (EA) [2, 15, 19, 20], reinforcement learning (RL) [14, 18, 29, 30, 31], or gradient-based methods [16, 26]. One-Shot methods [3, 4, 9] usually train a supernet covering the search space once and then apply search algorithms to search the best path in this supernet as the searched architecture.

After the seminal work by Zoph and Le [30] that requires up to hundreds of GPU days to find a good architecture, many NAS methods [14, 16, 18] try to reduce the search cost through different approaches. The community first turns to searching for primary building cells rather than the entire network [29, 31]. Based on the cell search, some approaches try to use performance prediction based on learning curves [2] or surrogate models [14] to expedite the evaluation process. Parameter sharing between child models is also common for acceleration, e.g., in DARTS [16]. ENAS [18], and One-Shot methods [3, 4, 9]. Our EcoNAS is an evolutionary-based NAS method [2, 15, 19, 20]. Different from previous works that fix the proxies when searching for network architectures, we design a fast and consistent proxy to reduce the search time, and apply hierarchical proxy strategy to improve the search efficiency. It is noteworthy that another trend of NAS is to search for efficient architectures [5, 10, 22, 25] by adding constraints during search process such as latency [10, 22, 25] on specific platforms. This differs to our focus on reducing the search cost by more careful and systematic proxy design.

Accelerating neural network training. Parallel computing [1, 27, 28] could significantly accelerate the training process of deep neural networks. Although these methods save the training time to minutes, the reduction in training time comes with the cost of thousands of GPUs, and their computation overhead remains large. Many studies [7, 17, 23] evaluate the performance of different networks at reduced settings, e.g., using smaller images in Tiny ImageNet Challenge1, using the encodings of network layers [7], reducing the number of training iterations [17], and reducing samples [23]. These studies assume that a specific reduced setting is sufficiently consistent thus do not evaluate the influence of different reduced settings. In contrast, we extensively evaluate influences from different reduced settings.

3. Exploration Study

In this study, we investigate the behaviors of different reduced settings. Previous works have not studied such behaviours comprehensively and systematically. To facilitate our study, we construct a model zoo containing 50 networks

1https://tiny-imagenet.herokuapp.com
for comparing the rank consistency before and after applying reduced settings. Each network in the model zoo is a stack of cells with alternating normal cells and reduction cells [14, 18, 31], generated by random sampling. The cell can be regarded as a directed acyclic graph consisting of an ordered sequence of nodes. Each node is a feature map in convolution networks and each directed edge is associated with an operation to be searched or randomly sampled. The entire network structure and one example of the cell are shown in Fig. 2(a) and Fig. 2(c), respectively. The details are provided in the Appendix.

3.1. Reduction Factors

In our experiments, a reduced setting corresponds to a combination of four factors: (1) number of channels for CNNs (c), (2) resolution of input images (r), (3) training epochs (e), and (4) sample ratio (s). The degree of reduction of c, r, and s is represented by a subscript 0, 1, 2, 3, 4. Specific values for the subscript of each reduction factor, for CIFAR-10 dataset [12], are shown in Table 1. The subscript for e is an integer, e.g., e_30 indicates 30 training epochs. The value of c refers to the number of channels in the initial convolutional cell. As for the sample ratio s, we randomly select a subset from the training data according to the value of s and then train the network on this subset. For CIFAR-10, (c_0, r_0, s_0, e_0) corresponds to the original or conventional setting [16], which uses 36 initial channels and 32 x 32 input image. All training samples are used for training with 600 epochs. In comparison, (c_3, r_0, s_c, e_x) corresponds to a reduced setting requiring around 1/2^{a+b+c} FLOPs, i.e., 2^{x+b+c} speed-up, when compared with (c_0, r_0, s_0, e_x).

3.2. Evaluation Metric

We use Spearman Coefficient, denoted as $\rho_{sp}$, as the metric to evaluate the reliability of reduced settings. The same metric is also used in [14] but it is employed for evaluating the performance of accuracy predictors. In our work, we use the metric to measure the dependence of ranks between the original setting and the reduced settings when the reduced settings are used for ranking the models in the model zoo. Assuming the model zoo has K networks, the formulation of $\rho_{sp}$ is

$$\rho_{sp} = 1 - \frac{6 \sum_{i=1}^{K} d_i^2}{K(K^2 - 1)},$$

where $d_i$ is the difference for network i between its rank in the original setting and its rank in the reduced setting.

A higher Spearman Coefficient corresponds to a more reliable reduced setting. An architecture found in a reliable reduced setting is more likely to remain high rank in the original setting. For example, Fig. 3 shows two different reduced settings, i.e., one with a small $\rho_{sp}$ (Fig. 3(a)) and another one with a large $\rho_{sp}$ (Fig. 3(b)). When $\rho_{sp}$ is large, the dependence of ranks increases between the original setting and the reduced setting for the same model. This indicates that the average change of ranks when switching from the reduced setting to the original one is small; thus, the reduced setting is more consistent and reliable. When $\rho_{sp}$ is small, the rank obtained from the reduced setting is less reliable. The models with low ranks and high ranks at the original setting could rank highly and low at the reduced setting more frequently, respectively.

3.3. Experimental Observations

We try different combinations of reduction factors and investigate the rank consistency between the original and

---

Table 1. Specific values of reduction factors for CIFAR-10 [12]. For the values of $c$, $r$, and $s$, if $c = \frac{c_0 + c_1}{117.78}$, $r$ is an integer, the integer is taken directly ($c_2$, $r_2$). Otherwise, the nearest number divisible by 4 is selected.

| Reduction factor | 0 | 1 | 2 | 3 | 4 |
|------------------|---|---|---|---|---|
| $c$              | 36| 24| 18| 12| 8 |
| $r$              | 32| 24| 16| 12| 8 |
| $s$              | 1.0| 0.5| 0.25| 0.125| |
| $e$              | 30| 60| 90| 120| |
more samples in each epoch is a better choice than train-
the results show that training with less epochs but using
As shown in Fig. 4, Spearman Coefficient
and
feasible while reducing the channels of networks is more
2) Reducing the resolution of input images is sometimes
were used to evaluate monotonic increasing or
decreasing trend for the \( \rho_s \) corresponding to the change of \( r \) or \( c \).
reduced settings by Spearman Coefficient on the CIFAR-
10 dataset [12]. The experimental setting is described in the
Two useful observations are obtained as follows:
1) With the same iteration numbers, using more training
samples with fewer training epochs is more consistent than using more training epochs and fewer training
samples. We first analyze the influence of sample ratio \( s \) and training epoch \( e \). There are 25 combinations of \( c \) and \( r \) for each combination of \( s \) and \( e \) as indicated in Table 1.
As shown in Fig. 4, Spearman Coefficient \( \rho_{sp} \) increases when the number of epochs and sample ratio increases. Therefore, the rank consistency improves with more training epochs and/or more training samples. The increase of \( \rho_{sp} \) from 30 epochs to 60 epochs is the most obvious, after which the benefits brought about by more epochs become less apparent. When comparing the reduced setting pairs that have the same number of iterations such as \( c_x r_y s_0 e_2 \) and \( c_x r_y s_1 e_2 \), the results show that training with less epochs but using more samples in each epoch is a better choice than training with more epochs but fewer samples, especially when training for more iterations. Considering the trade-off between computation consumption and benefits, we find the combination \( s_0 e_60 \) to be a more optimal setting than other combinations.
2) Reducing the resolution of input images is sometimes
reliable than reducing the resolution. We further analyze the effect of reduction factors \( c \) and \( r \) with fixed \( s_0 e_60 \) as discovered in our previous experiments. The 3D-Bar chart in Fig. 5 shows the changes in \( \rho_{sp} \) along the dimensions \( c \) (X-axis) and \( r \) (Y-axis). The Z-axis represents \( \rho_{sp} \) for each \( c_x r_y \) setting, illustrated by the height of the bars. We use a measurement called entropy denoted by \( \rho_e \) to indicate the monotonicity of \( \rho_{sp} \) along a particular dimension of reduction factor. The details of this measurement are provided in the Appendix. The entropy ranges from -1 to 1. The absolute value of entropy close to 1 indicates that along this dimension, the objective set has an obvious monotonous increasing or decreasing trend, with only few small fluctuations. On the contrary, if the absolute value of entropy is relatively small, e.g., less than 0.5, then the monotonic increasing or decreasing trend along this dimension is less apparent.
As shown by the values of entropy close to the Y-axis in Fig. 5, the trend along the reduction factor \( c \) is obvious. For most of the fixed \( r \) settings, fewer channels lead to a better-behaved reduced setting, except for few points e.g., \( c_2 r_2 \). The variation of entropy close to the X-axis indicates that the trends along \( r \) for the fixed \( c \) are not apparent. But for most of the cases, a smaller \( r \) will lead to a smaller \( \rho_{sp} \), which indicates that reducing the resolution is not reliable.
Figure 6(a) shows the effects of \( c \) and \( e \), where the other two factors are fixed to \( r_0 s_0 \). The entropies close to the Y-axis indicate that the decrease in \( c \) will increase \( \rho_{sp} \) for the same number of epochs, which is consistent with the

Figure 4. The influence of factors \( s \) and \( e \). The Y-axis is the average \( \rho_{sp} \) for 25 different reduced settings that only differ in \( c \) and \( r \) but have the same \( s \) and \( e \).
reduced settings on CIFAR-10 [12]. The X-axis represents the ratio of FLOPs at the original setting to that at the reduced setting. The Y-axis provides the corresponding $\rho_{sp}$ (Eq. (1)). The reduced settings are getting more efficient along X-axis and getting more consistent in ranking different network candidates along Y-axis. The blue points show all the reduced settings, the orange points are good ones, and the green point is the recommended one.

Figure 8. The rank consistency and acceleration ratios of different reduced settings on CIFAR-10 [12]. The X-axis represents the ratio of FLOPs at the original setting to that at the reduced setting. The Y-axis provides the corresponding $\rho_{sp}$ (Eq. (1)). The reduced settings are getting more efficient along X-axis and getting more consistent in ranking different network candidates along Y-axis. The blue points show all the reduced settings, the orange points are good ones, and the green point is the recommended one.

The results in Fig. 5 for the same resolution. However, lower resolution leads to smaller $\rho_{sp}$, which is also consistent with the results in Fig. 5. Both Fig. 6(a) and Fig. 6(b) show that more epochs improves $\rho_{sp}$, which also validates the results in Fig. 4.

The results in Fig. 5 and Fig. 6(a) indicate that decreasing the number of channels will lead to an increase of $\rho_{sp}$, which means that more reliable setting can be achieved by fewer channels with less computation and memory. Fig. 7 also illustrates this phenomenon, where smaller accuracy difference appears with fewer channels between training and testing dataset. This phenomenon may be caused by overfitting when the amount of parameters for the same architecture is reduced. We hypothesize that overfitting has an adverse effect on the rank consistency.

4. Economical Evolutionary-Based NAS

Based on the investigations in the previous section, we propose Economical evolutionary-based NAS (EcoNAS), an accelerated version of the evolutionary-based state of the art [19]. Existing evolutionary-based NAS methods [2, 15, 19, 20] suffer from large computation cost due to time wasted on unpromising networks. Taking heed of prior observations (Sec. 3.3), we designed a more efficient and consistent proxy (Sec. 4.1) that reduces the search cost by a large margin. Then we propose a hierarchical proxy strategy that trains networks with different proxies based on their respective accuracy. This further improves the search efficiency by focusing more on good architectures and using the most accurate ones for evolution. The search strategy and algorithm pipeline are described in Sec. 4.2.

4.1. Efficient Proxy

To replace the original proxies in previous NAS methods [19, 31] with a more efficient one, we further analyze the acceleration ratio for FLOPs and $\rho_{sp}$ of the 200 reduced settings and provide empirical results on the good reduced settings. We divide the acceleration ratios into different groups and list the reduced settings that perform best in each group. We also compare two reduced settings in previous works [19, 31] where the stack number $N$ is set to 2. In NASNet [31] the reduced setting is $c_{0}r_{0}\bar{n}_{0}\bar{c}_{20}$, and in AmoebaNet [19] the setting is $c_{1}r_{1}\bar{n}_{1}\bar{c}_{25}$. These two reduced settings exhibit less consistency as shown in Fig. 8.

Figure 8 suggests that a proxy that substantially reduces the computation overhead does not necessarily have a poor rank consistency. There are many good reduced settings that possess large acceleration ratio and high consistency at the same time, just as the orange points in Fig. 8. The orange points in Fig. 8 show that many recommended reduced settings include $c_{4}$, $c_{90}$, $c_{120}$. The observation verifies our previous conclusion about the benefits of more epochs and fewer channels. We adopt the reduced setting $c_{4}r_{4}\bar{n}_{4}\bar{c}_{60}$ in EcoNAS because this reduced setting exhibits relatively higher consistency and acceleration ratios as shown by the green point in Fig. 8. These reduced settings’ exact values are shown in Table 1. Under this reduced setting, we enlarge the batch size from 96 to 384 to further compress the searching time.

4.2. Hierarchical Proxy

Although we found an efficient and consistent proxy, we observe that training and evaluating each architecture with the same proxy still hurts the search efficiency because unpromising networks count a lot and waste most of the search time. An intuitive approach to this problem is to reduce the training epochs further based on our newly designed consistent proxy. This strategy could find and discard inaccurate architectures faster, however, the best model found from such a proxy with less training epochs might not be as good because the consistency will degrade given fewer training epochs. Therefore, we propose hierarchical proxy strategy as a trade-off to precisely rank good architectures and save time from assessing unpromising networks.

Population set. We divide the population set into three subsets $P_{1}$, $P_{2}$, $P_{3}$, which contains networks with low, middle, and high accuracy and has small, middle, and high probability of being chosen for mutation, respectively. For networks in the sets from $P_{1}$ to $P_{3}$, the networks are trained with faster but less consistent proxies (based on observations in Sec. 3.3). We design these three proxies sharing the same $c$, $r$, $s$ but training networks with $3E$, $2E$, and $E$ epochs for $P_{3}$, $P_{2}$, and $P_{1}$ (e.g., 90, 60, and 30 if $E = 30$), respectively. We apply this design because it exploits the weight sharing strategy and it is simple. For example, when several newly mutated networks remain after training $E$ epochs, we only need to train them with $E$ more epochs to rank them more precisely rather than to train with a new proxy from scratch. More diverse proxies for the hierarchical proxy...
strategy could also be tried and we leave it to future research.

During each evolution cycle, we use proxy with the lowest computation cost to quickly remove the less accurate networks from newly mutated networks. Then the remaining networks, which are more accurate, will be evaluated with slower but more consistent proxy and assigned into the subsets $P_2$, $P_3$ hierarchically. When choosing networks for mutation from the same subsets, the networks with a higher rank are more likely to be chosen to increase the probability of producing new accurate architectures. Such a design saves the computation overhead by only training those networks with low accuracies using less epochs. Furthermore, allocating resources to more promising architectures allows the good ones to be found with higher precision according to the observation in Section 3.3. Assigning larger probabilities to these networks for mutation also potentially helps to produce more promising architectures in fewer evolution cycles.

Algorithm pipeline. The pipeline of EcoNAS is shown in Algorithm 1. The initial architectures are randomly generated. All the searched architectures and their accuracies will be stored in the history set. As explained before, in EcoNAS, there are three population sets $P_1$, $P_2$, and $P_3$, also noted as $P_E$, $P_{2E}$, $P_{3E}$ because they store architectures that have been trained for $E$, $2E$, $3E$ epochs, respectively. For each evolution cycle, the following three steps are conducted:

1. A batch of randomly sampled architectures from the population sets $P_E$, $P_{2E}$, and $P_{3E}$ will be mutated (Algorithm 1, line 12-17). Architectures with higher performance are more likely to be chosen. We follow the mutation rules in AmoebaNet [19] except that we remove the ‘identity’ mutation, because in EcoNAS the amount of networks at each cycle is relatively fewer. The mutated networks are trained from scratch for $E$ epochs and then added to set $P_E$.

2. Choose the top architectures from set $P_E$ and $P_{2E}$ to $P_{2E}$ and $P_{3E}$, respectively (Algorithm 1, line 18-25). These architectures are loaded from checkpoints and trained for $E$ more epochs, then those top architectures are more precisely ranked and moved to the corresponding subsets.

3. Remove dead architectures from set $P_E$, $P_{2E}$ and $P_{3E}$ to force the algorithm to focus on newer architectures [19].

Protocol Details. For EcoNAS, the number of nodes (Fig. 2) within one architecture is four and we set the stack number $N = 6$ to build the networks. We found in our experiments that $N = 2$, a setting used in previous works [19, 31] is detrimental to the rank consistency (Fig. 8). Our searching experiments are conducted on CIFAR-10 [12] and we split 5k images from the training set as validation set. The search space is similar to that in [16]. Initially, $P_1$ has 50 randomly generated architectures, while $P_2$ and $P_3$ are empty. At each cycle, 16 architectures sampled from $P_k$, $k \in [E, 2E, 3E]$ will be mutated. We set the number of cycles $C$ to 100 and training epoch $E$ to 20. All the experiments and search cost calculation are conducted using NVIDIA GTX 1080Ti GPUs.

5.1. Overall Results on CIFAR-10 and ImageNet

The searching process yields several candidate convolutional cells and these cells are evaluated on two datasets, CIFAR-10 [12] and ImageNet [21] for image classification tasks. Unlike previous works [19, 29], which select dozens or hundreds of cells to evaluate, we only pick up the top-5 cells from history to evaluate the performance of our search algorithm. Retraining top-5 cells saves the retraining overhead and is enough in our experiments; the reason is analyzed in Section 5.3.

Results on CIFAR-10. For the task of CIFAR-10, we set $N = 6$ and the initial channel $c = 36$. The networks are trained from scratch for 600 epochs with batch size 96. Other hyper-parameters follow previous works [16, 26]. The experimental results on CIFAR-10 are shown in Table 2. The best model found by our algorithm achieves a test
Table 2. CIFAR-10 test errors of EcoNAS and the state-of-the-art networks. The ‘Error’ refers to top-1 error rate and ‘Cost’ refers to the number of GPU days.

| Network               | Error (%) | Params. (M) | Cost |
|-----------------------|-----------|-------------|------|
| DenseNet-B[11]        | 3.46      | 25.6        | -    |
| NASNet-A [31]         | 2.65      | 33.2        | 1800 |
| Amoeba-A [19]         | 3.34±0.06 | 3.2         | 3150 |
| Amoeba-B [19]         | 2.55±0.05 | 2.8         | 3150(TPU) |
| Hierarchical Evo [15] | 3.75±0.12 | 15.7        | 300  |
| PNAS[14]              | 3.41±0.09 | 3.2         | 225  |
| ENAS[18]              | 2.89      | 4.6         | 0.5  |
| DARTS(1st order) [16] | 2.94      | 2.9         | 1.5  |
| DARTS(2nd order) [16] | 2.83±0.06 | 3.4         | 4    |
| SNAS[26]              | 2.85±0.02 | 2.8         | 1.5  |
| EcoNAS                | 2.62±0.02 | 2.9         | 8    |

Table 3. ImageNet Test errors of EcoNAS and the state-of-the-art networks. The ‘Error’ refers to top-1 error rate and ‘Cost’ refers to the number of GPU days.

| Architecture        | Error (%) | Params. (M) | Cost |
|---------------------|-----------|-------------|------|
| NASNet-A [31]       | 26.0      | 5.3         | 1800 |
| NASNet-B [31]       | 27.2      | 5.3         | 1800 |
| NASNet-C [31]       | 27.5      | 4.9         | 1800 |
| AmoebaNet-A [19]    | 25.5      | 5.1         | 3150 |
| AmoebaNet-B [19]    | 26.0      | 5.3         | 3150 |
| AmoebaNet-C [19]    | 24.3      | 6.4         | 3150 |
| DARTS [16]          | 26.9      | 4.9         | 4    |
| SNAS[26]            | 27.3      | 4.3         | 1.5  |
| EcoNAS              | 25.2      | 4.3         | 8    |

efficiency of 2.60%, a rate that is on par with state-of-the-art evolution-based methods and is much lower than most Reinforcement Learning (RL)-based methods. Importantly, the model uses about 400× less computation resources. Under the same magnitude of computational complexity, our result is superior to gradient-based methods [16, 26] and weight sharing method [18].

Results on ImageNet. To evaluate the transferability of the cells searched by EcoNAS, we transfer the architectures to ImageNet, where we only change the stack number N to 14 and enlarge the initial channel c to 48. As shown in Table 3, the best architecture found by EcoNAS on CIFAR-10 generalizes well to ImageNet. EcoNAS achieves top-1 error that outperforms the previous works that consume the same magnitude of GPUs [16, 26]. EcoNAS also surpasses models with similar amount of parameters (i.e., having fewer than 5M parameters) found by reinforcement learning [31] and evolution algorithms [19], which require about 200× more GPU resources.

5.2. Ablation Study of EcoNAS

Reduced setting for NAS using evolutionary algorithms. We evaluate the reduced setting on the conventional evolutionary algorithms in Table 4. AmoebaNet [19], NAS-
for networks as shown in Fig. 10. The difference of average accuracy for networks in those sets increase gradually, which deviate much during evolution cycles, but the average accuracy for networks in those sets increase gradually, which

sign newly mutated networks into three strategies.

Diversity of structures by hierarchical training strategy. We further evaluate the ability of the reduced settings in retaining good models. If a top-10 model at the original setting is a top-15 or top-20 model at the reduced setting, it will be considered as a ‘good model retained’. We divide the reduced settings into different groups according to their corresponding $\rho_{sp}$ and calculate the average number of good models retained for reduced settings in the same group. As $\rho_{sp}$ increases, the number of good models retained also increases (Fig. 9), suggesting that reduced settings with better consistency can keep more good models in the top positions.

Figure 9 also indicates that reliable proxies can reduce not only the search cost but also the retraining overhead. Previous works need to select the best model after retraining hundreds of top models searched under proxies (e.g., 100 [29] for BlockQNN and 250 for NASNet [31]), AmoebaNet [19] also needs to retrain 20 models. One of the reasons for this phenomenon is that previous works adopt less reliable proxies as shown in Fig. 8. Since the ranks of networks under the proxies are less consistent with their actual ranks under the original setting, they need to select more networks for retraining to find the optimal one. According to the conclusion obtained from Fig. 9, a more consistent reduced setting retains more top models, thus allows the search algorithm to retrain fewer networks to obtain competitive accuracy. With a more reliable proxy, EcoNAS only retains top-5 networks and remarkably saves computation overhead for retraining models, which is usually overlooked in most NAS literature.

Diversity of structures by hierarchical training strategy. EcoNAS adopts hierarchical training strategy to assign newly mutated networks into three population sets and sample networks for mutation from these sets with different probabilities. Under this scheme, both the good and bad models enjoy chances to mutate so that the architectures in the next evolution cycle will not be trapped in a few local optimum structures. We evaluate the average accuracy of networks in the three population sets during evolution cycles as shown in Fig. 10. The difference of average accuracy for networks in $P_E$, $P_{2E}$, and $P_{3E}$ are apparent and do not deviate much during evolution cycles, but the average accuracy for networks in those sets increase gradually, which verifies the diversity of structures in those population sets.

The diversity of structures allows the search algorithms to find accurate architectures with fewer search costs by potentially helping to explore the search space more comprehensively. Since hierarchical training strategy provides the diversity of structures in the population sets, EcoNAS obtains similar competitive architectures after evaluating 1k models. As a comparison, BlockQNN [29] evaluates 11k models, AmoebaNet [19] evaluates 20k models, and NASNet [31] evaluates 45k models.

5.3. Analysis

Ability of reliable proxies to retain top models. We further evaluate the ability of the reduced settings in retaining good models. If a top-10 model at the original setting is a top-15 or top-20 model at the reduced setting, it will be considered as a ‘good model retained’. We divide the reduced settings into different groups according to their corresponding $\rho_{sp}$ and calculate the average number of good models retained for reduced settings in the same group. As $\rho_{sp}$ increases, the number of good models retained also increases (Fig. 9), suggesting that reduced settings with better consistency can keep more good models in the top positions.

Figure 9 also indicates that reliable proxies can reduce not only the search cost but also the retraining overhead. Previous works need to select the best model after retraining hundreds of top models searched under proxies (e.g., 100 [29] for BlockQNN and 250 for NASNet [31]), AmoebaNet [19] also needs to retrain 20 models. One of the reasons for this phenomenon is that previous works adopt less reliable proxies as shown in Fig. 8. Since the ranks of networks under the proxies are less consistent with their actual ranks under the original setting, they need to select more networks for retraining to find the optimal one. According to the conclusion obtained from Fig. 9, a more consistent reduced setting retains more top models, thus allows the search algorithm to retrain fewer networks to obtain competitive accuracy. With a more reliable proxy, EcoNAS only retains top-5 networks and remarkably saves computation overhead for retraining models, which is usually overlooked in most NAS literature.

Diversity of structures by hierarchical training strategy. EcoNAS adopts hierarchical training strategy to assign newly mutated networks into three population sets and sample networks for mutation from these sets with different probabilities. Under this scheme, both the good and bad models enjoy chances to mutate so that the architectures in the next evolution cycle will not be trapped in a few local optimum structures. We evaluate the average accuracy of networks in the three population sets during evolution cycles as shown in Fig. 10. The difference of average accuracy for networks in $P_E$, $P_{2E}$, and $P_{3E}$ are apparent and do not deviate much during evolution cycles, but the average accuracy for networks in those sets increase gradually, which

Last but not least, we find some proxies led by our observations are also applicable to other NAS methods [5, 16]. These proxies further reduce the search time while the discovered models achieve comparable or better performance. We wish our work could inspire the community to further explore more practical proxies and search algorithms to improve the efficiency of NAS.

Acknowledgement This work was supported by the Australian Research Council Grant DP200103223.
References

[1] Takuya Akiba, Shuji Suzuki, and Keisuke Fukuda. Extremely large minibatch SGD: Training ResNet-50 on ImageNet in 15 minutes. CoRR, abs/1711.04325, 2017.

[2] Bowen Baker, Otirik Gupta, Ramesh Raskar, and Nikhil Naik. Accelerating neural architecture search using performance prediction. CoRR, abs/1705.10823, 2017.

[3] Gabriel Bender, Pieter-Jan Kindermans, Barret Zoph, Vijay Vasudevan, and Quoc V. Le. Understanding and simplifying one-shot architecture search. In ICLR, 2018.

[4] Andrew Brock, Theodore Lim, James M. Ritchie, and Nick Weston. SMASH: One-shot model architecture search through hypernetworks. CoRR, abs/1708.05344, 2017.

[5] Han Cai, Ligeng Zhu, and Song Han. Proxylessnas: Direct neural architecture search on target task and hardware. In ICLR. 2019.

[6] Yukang Chen, Tong Rui Yang, Xiangyu Zhang, Gaofeng Meng, Chunhong Pan, and Jian Sun. DetNAS: Backbone search for object detection. In NeurIPS, 2019.

[7] Boyang Deng, Junjie Yan, and Dahua Lin. Peephole: Predicting network performance before training. CoRR, abs/1712.03351, 2017.

[8] Golnaz Ghiasi, Tsung-Yi Lin, Ruoming Pang, and Quoc V. Le. NAS-FPN: learning scalable feature pyramid architecture for object detection. CoRR, abs/1904.07392, 2019.

[9] Zichao Guo, Xiangyu Zhang, Haoyuan Mu, Wen Heng, Zechun Liu, Yichen Wei, and Jian Sun. Single path one-shot neural architecture search with uniform sampling. CoRR, abs/1904.00420, 2019.

[10] Andrew 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. CoRR, abs/1905.02244, 2019.

[11] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In CVPR, 2017.

[12] Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009. Technical report.

[13] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan L. Yuille, and Li Fei-Fei. Auto-DeepLab: Hierarchical neural architecture search for semantic image segmentation. In CVPR, 2019.

[14] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In ECCV, 2018.

[15] Hanxiao Liu, Karen Simonyan, Oriol Vinyals, Chrisantha Fernando, and Koray Kavukcuoglu. Hierarchical representations for efficient architecture search. CoRR, abs/1711.00436, 2017.

[16] Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: Differentiable architecture search. In ICLR, 2019.

[17] Dmytro Mishkin, Nikolay Sereyevskiy, and Jiri Matasa. Systematic evaluation of cnn advances on the imagenet. CVIU, 2017.

[18] Hieu Pham, Melody Y. Guan, Barret Zoph, Quoc V. Le, and Jeff Dean. Efficient neural architecture search via parameter sharing. In ICML, 2018.

[19] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V. Le. Regularized evolution for image classifier architecture search. In AAAI, 2019.

[20] Esteban Real, Sherry Moore, Andrew Selle, Saurabh Saxena, Yutaka Leon Suematsu, Jie Tan, Quoc V. Le, and Alexey Karakin. Large-scale evolution of image classifiers. In ICLR, 2017.

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. IJCV, 2015.

[22] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, and Quoc V. Le. MnasNet: Platform-aware neural architecture search for mobile. CoRR, abs/1807.11626, 2018.

[23] Kailas Vodrahalli, Ke Li, and Jitendra Malik. Are all training examples created equal? an empirical study. CoRR, abs/1811.12569, 2018.

[24] Ning Wang, Yang Gao, Hao Chen, Peng Wang, Zhi Tian, and Chunhua Shen. NAS-FCOS: Fast neural architecture search for object detection. CoRR, abs/1906.04423, 2019.

[25] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanhan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. FBNet: Hardware-aware efficient convnet design via differentiable neural architecture search. In CVPR, 2019.

[26] Sirui Xie, Huhui Zheng, Chunxiao Liu, and Liang Lin. SNAS: Stochastic neural architecture search. In ICLR, 2019.

[27] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. CoRR, abs/1708.03888v3, 2017.

[28] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In ICLR, 2018.

[29] Zhao Zhong, Junjie Yan, Wei Wu, Jing Shao, and Cheng-Lin Liu. Practical block-wise neural network architecture generation. In CVPR, 2018.

[30] Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. In ICLR, 2017.

[31] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning transferable architectures for scalable image recognition. In CVPR, 2018.