NEURAL ARCHITECTURE SEARCH BY LEARNING ACTION SPACE FOR MONTE CARLO TREE SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural Architecture Search (NAS) has emerged as a promising technique for automatic neural network design. However, existing NAS approaches often utilize manually designed action space, which is not directly related to the performance metric to be optimized (e.g., accuracy). As a result, using manually designed action space to perform NAS often leads to sample-inefficient explorations of architectures and thus can be sub-optimal. In order to improve the sample efficiency, this paper proposes Latent Action Neural Architecture Search (LaNAS), which learns actions to recursively partition the search space into good or bad regions that contain networks with concentrated performance metrics, *i.e.*, low variance. During the search phase, as different architecture search action sequences lead to regions with different performance, the search efficiency can be significantly improved by biasing towards the good regions. On the largest NAS dataset NASBench-101, our experimental results demonstrated that LaNAS is $22\times$, $14.6\times$, $12.4\times$, $6.8\times$, $16.5\times$ more sample-efficient than Random Search, Regularized Evolution, Monte Carlo Tree Search, Neural Architecture Optimization, and Bayesian Optimization, respectively. When applied to the open domain, LaNAS achieves 98.0% accuracy on CIFAR-10 and 75.0% top1 accuracy on ImageNet in only 803 samples, outperforming SOTA AmoebaNet with $33\times$ fewer samples.

1 INTRODUCTION

During the past two years, there has been a growing interest in Neural Architecture Search (NAS) that aims to automate the laborious process of designing neural networks. Architectures found by NAS have achieved remarkable results in image classification (Zoph and Le (2016); Real et al. (2018)), object detection and segmentation (Ghiasi et al. (2019); Chen et al. (2019); Liu et al. (2019)), as well as other domains such as language tasks (Luong et al. (2018); So et al. (2019)).

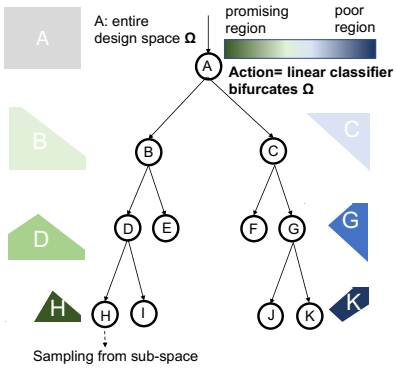
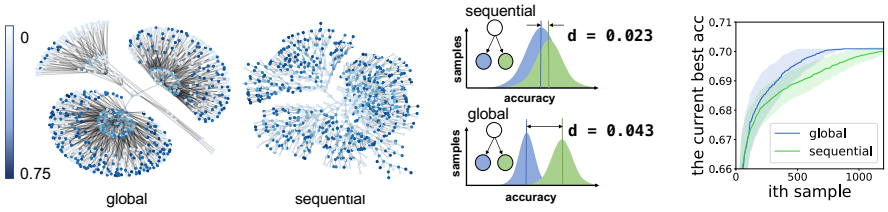


Figure 1: Latent Action Neural Architecture Search: Starting from the entire model space, at each search stage we learn an action (or a set of *linear constraints*) to separate good from bad models for providing distinctive rewards for better searching.

Starting from hand-designed discrete model space and action space, NAS utilizes search techniques to explore the search space and find the best performing architectures with respect to a single or multiple objectives (*e.g.*, accuracy, latency, or memory), and preferably with minimal search cost.

However, one common issue faced by the previous works on NAS is that the action space needs to be manually designed. The action space proposed by Zoph et al. (2018) involves sequential actions to construct a network, such as selecting two nodes, and choosing their operations. Other prior works including reinforcement learning based (Zoph and Le (2016); Baker et al. (2016)), evolution-based (Real et al. (2018; 2017)), and MCTS-based (Wang et al. (2019); Negrinho and Gordon (2017)) approaches, all use manually designed action spaces. As suggested in Sciuto et al. (2019); Li and Talwalkar (2019); Xie et al. (2019), action



(a) Visualization of MCTS search trees (b) Reward distribution at two child nodes (c) Search efficiency

Figure 2: **Illustration of motivation**: (a) visualizes the MCTS search trees using `sequential` and `global` action space. The node value (*i.e.* accuracy) is higher if the color is darker. (b) For a given node, the reward distributions for its children. d is the average distance over all nodes. `global` better separates the search space by network quality, and provides distinctive reward in recognizing a promising path. (c) As a result, `global` finds the best network much faster than `sequential`.

space design alone can be critical to network performance. Furthermore, it is often the case that manually designed action space is not related to the performance that needs to be optimized. In Sec 2, we demonstrate an example where subtly different action space can lead to significantly different search efficiency. Finally, unlike games that generally have a predefined action space (e.g., Atari, Chess and Go), in NAS, it is the final network that matters rather than the specific path of the search, which gives a large playground for action space learning.

Based on the above observations, we propose Latent Action Neural Architecture Search(LaNAS) that learns the action space to maximize search efficiency for performance metrics. Previous methods typically construct an architecture from an empty network by sequentially applying predefined actions, e.g. adding layers, setting the kernel or depth, resulting in a state space that provides undiscriminating rewards to inform the search (Fig. 2a `sequential`). LaNAS takes a dual approach and treats each action as a *linear constraint* that intersects with the search space Ω to bifurcate it into a good and bad region that contains networks with similar performance metric (Fig. 1). Once multiple actions are recursively applied to the entire search space, the search space is clearly separated (e.g. Fig. 2a `global`, Fig. 4) into regions that provide more informative rewards, thereby greatly improving the search efficiency. To achieve this goal, LaNAS iterates between *learning* and *searching* stage. In the learning stage, each action in LaNAS is learned to partition the model space into high-performance and low-performance regions, to achieve accurate performance prediction. In the searching stage, LaNAS applies MCTS on the learned action space to sample more model architectures. The learned actions provide an informed guide for MCTS to sample from performance-promising regions, while the adaptive exploration in MCTS collects more data to progressively refine the learned actions to zoom into more promising regions for further improving the sample efficiency. The iterative process is jump-started by first collecting a few random samples.

We show that LaNAS yields a tremendous acceleration on a diverse set of benchmark tasks, including publicly available NASBench-101 (420,000 NASNet models trained on CIFAR-10) (Ying et al. (2019)), our self-curated ConvNet-60K (60,000 plain VGG-style ConvNets trained on CIFAR-10), and LSTM-10K (10,000 LSTM cells trained on PTB). Our algorithm consistently finds the best performing architecture on all three tasks with at least an order of fewer samples (on average) than Random Search, Regularized Evolution, MCTS, Neural Architecture Optimization and Bayesian Optimization. In the open domain search scenario, our algorithm finds a network that achieves 98.0% accuracy on CIFAR-10 and 75.0% top1 accuracy (mobile setting) on ImageNet in only 803 samples, using 33$\times$ fewer samples and achieving higher accuracy than AmoebaNet (Real et al. (2018)). Moreover, we empirically demonstrate that the learned latent actions can transfer to a new search task to further boost efficiency. Finally, we provide empirical observations to illustrate the search dynamics and analyze the behavior of our approach. We also conduct various ablation studies in together with a partition analysis to provide guidance in determining search hyper-parameters and deploying LaNAS in practice.

2 A Motivating Example

To demonstrate the importance of action space in NAS, we start with a motivating example. Consider a simple scenario of designing a plain Convolutional Neural Network (CNN) for CIFAR-10 image

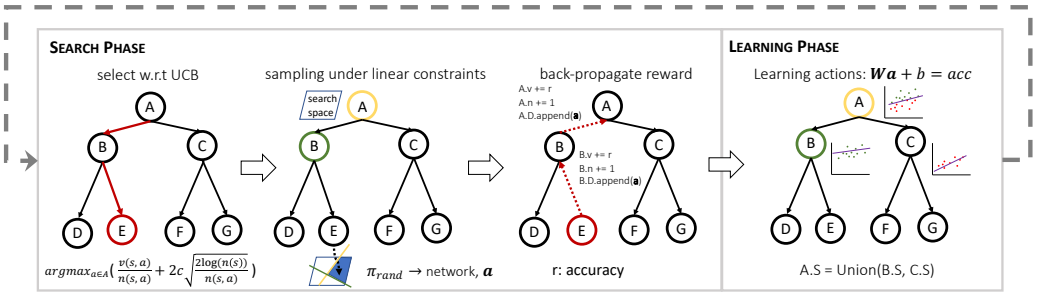


Figure 3: **An overview of LaNAS**: LaNAS is an iterative algorithm in which each iteration comprises a search phase and learning phase. The search phase uses MCTS to samples networks, while the learning phase learns a linear model between network hyper-parameters and their accuracies.

classification. The primitive operation is a Conv-ReLU layer. Free structural parameters that can vary include network depth $L = \{1, 2, 3, 4, 5\}$, number of filter channels $C = \{32, 64\}$ and kernel size $K = \{3 \times 3, 5 \times 5\}$. This configuration results in a search space of 1,364 networks. To perform the search, there are two natural choices of the action space: `sequential` and `global`. `sequential` comprises actions in the following order: adding a layer l, setting kernel size K_l, setting filter channel C_l. The actions are repeated L times. On the other hand, `global` uses the following actions instead: {Setting network depth L, setting kernel size $K_{1,\ldots,L}$, setting filter channel $C_{1,\ldots,L}$}. For these two action spaces, MCTS is employed to perform the search. Note that both action spaces can cover the entire search space but have very different search trajectories.

Fig. 2(a) visualizes the search for these two action spaces. Actions in `global` clearly separates desired and undesired network clusters, while actions in `sequential` lead to network clusters with a mixture of good or bad networks in terms of performance. As a result, the overall distribution of accuracy along the search path (Fig. 2(b)) shows concentration behavior for `global`, which is not the case for `sequential`. We also demonstrate the overall search performance in Fig.2(c). As shown in the figure, `global` finds desired networks much faster than `sequential`.

This observation suggests that changing the action space can lead to very different search behavior and thus potentially better sample efficiency. In other words, an early exploration on the network depth is critical. Increasing the depth is an optimization direction that can potentially lead to better model accuracy. One might come across a natural question from this motivating example. Is it possible to find a principle way to distinguish a good action space from a bad action space in NAS? Is it possible to *learn an action space* such that it can best fit the performance metric to be optimized?

3 Learning Latent Actions

In this section, we describe LaNAS, which comprises two phases: (1) learning phase, and (2) search phase. Fig. 3 presents a high level description of LaNAS, of which the corresponding algorithms are further described in Alg.1.

3.1 Learning Phase

In the learning phase at iteration t, we have a dataset $D_t = \{(\mathbf{a}_i, v_i)\}$ obtained from previous explorations. Each data point $(\mathbf{a}_i, v_i)$ in D_t has two components: $\mathbf{a}_i$ represents network attributes (e.g., depth, number of filters, kernel size, connectivity, etc) and v_i represents the performance metric estimated from training (or from pre-trained dataset such as NASBench-101). Our goal is to learn a good action space from D_t to guide future exploration as well as to find the model with the desired performance metric efficiently.

Starting from the entire search space Ω, the learnt actions recursively (and greedily) split it into smaller regions such that the estimation of performance metric becomes more accurate. This helps us prune away poor regions as soon as possible and increase the sample efficiency of NAS.

In particular, we model the recursive splitting process as a tree. The root node corresponds to the entire model space Ω, while each tree node j corresponds to a region Ω_j (Fig. 1). At each tree node j, we partition Ω_j into disjoint regions $\Omega_j = \cup_{k \in (j)} \Omega_k$, such that on each child region Ω_k, the estimation of performance metric $V(\Omega_k)$ is the most accurate (or equivalently, has lowest variance).

At each node j, we learn a classifier that embodies an latent action to split the model space Ω_j. The linear classifier takes the portion of the dataset that falls into its own region $D_t \cap \Omega_j$, The performance of a region Ω_j, $V(\Omega_j)$, is estimated by $1/N \sum_{i \in D_t \cap \Omega_j} v_i$. To minimize the variance of $V(\Omega_k)$ for all child nodes, we learn a linear regressor f_j that minimizes $\sum_{i \in D_t \cap \Omega_j} (f_j(\mathbf{a}_i) - v_i)^2$. Once learned, the parameters of f_j and $\hat{V}(\Omega_j)$ form a linear constraint that bifurcates Ω_j into a good region ($> \hat{V}(\Omega_j)$) and a bad region ($\leq \hat{V}(\Omega_j)$) for sampling. For convenience, the left child always represents the good region. The partition threshold $\hat{V}(\Omega_j)$, combined with f_j, forms two latent actions at node j, going left ($f_j(\mathbf{a}_i) > \hat{V}(\Omega_j)$) and going right ($f_j(\mathbf{a}_i) \leq \hat{V}(\Omega_j)$).

Note that we need to initialize each node classifier properly with a few random samples to establish initial boundary in the search space. An ablation study on the number of samples for initialization is provided in Fig. 6c and Fig. 6d.

3.2 Search Phase

Once actions are learned, the search phase follows. The search uses the learned actions to sample more architectures $\mathbf{a}_i$ as well as their performance v_i, and store $(\mathbf{a}_i, v_i)$ in dataset D_t to refine the action space in the next learning phase. Note that in the search phase, the tree structure and the parameters of those classifiers are fixed and static. Instead, the search phase learns to decide which region Ω_j on tree leaves to sample $\mathbf{a}_i$, with a proper balance between the exploration of less known Ω_j and exploitation of promising Ω_j.

Following the construction of the classifier at each node, a trivial go-left greedy based search strategy can be used to exclusively exploit the most promising Ω_k defined by the current action space. However, this is not a good strategy since it only *exploits* the current action space, which is learned from the current samples and may not be optimal. There can be good model regions that are hidden in the right (or bad) leaves that need to be explored.

In order to overcome this issue, we use Monte Carlo Tree Search (MCTS) as the search method, which has the characteristics of adaptive exploration, and has shown superior efficiency in various tasks. MCTS keeps track of visiting statistics at each node to achieve the balance between exploitation of existing good region and exploration of new region. In lieu of MCTS, our search phase also has *select*, *sampling* and *backpropagate* stages. LaNAS skips the *expansion* stage in regular MCTS since the connectivity of our search tree is static. When the action space is updated, previous sampled networks and their performance metrics in D_t are reused and redirected to (maybe different) nodes in initializing visitation counts n(s) and node values v(s). When there is no learned action space, we random sample the model space to get jump started.

Algorithm 1: LaNAS search procedure.

```
Data: specifications of the search domain
Function get_ucb (X̄_next, n_next, n_curt )
    c = 0.5
    return X̄_next / n_next + 2c sqrt(2log(n_curt) / n_next) if n_next ≠ 0
      else +∞
begin
    while acc < target do
        for n ∈ Tree.N do
            n.g.train()
        for i = 1 → #selects do
            leaf, path = ucb_select(root)
            constraints =
              get_constraints(path)
            network =
              sampling(constraints)
            acc = network.train()
            back − propagate(network, acc)
```

Algorithm 2: Subroutines in Alg. 1.

```
Function get_constraints (s_path)
    constraints = []
    for node ∈ s_path do
        W, b = node.g.params(), X̄ = node.X̄
        if node on left then
            constraints.add(Wa + b ≥ X̄) else
            constraints.add(Wa + b < X̄)
    return constraints
Function ucb_select (c = root)
    path = []
    while c not leaf do
        path.add(c)
        l_ucb = get_ucb(c.left.X̄, c.left.n, c.n),
          r_ucb =
          get_ucb(c.right.X̄, c.right.n, c.n)
        if l_ucb > r_ucb then c = c.left else
          c = c.right
    return path, c
```

3.3 LANAS PROCEDURES

Learning phase: (Alg.1 line $6 \rightarrow 7$) the learning phase is to train f_j at every node j. With more samples, f becomes more accurate, and $V(\Omega_j)$ becomes closer to $\hat{V}(\Omega_j)$. However, as MCTS biases towards selecting good samples, LaNAS will zoom into the promising hyper-space to further improve the sample efficiency. We provided an analysis of such search dynamics in Fig. 4.

Search phase: (Alg.1 line $8 \rightarrow 13$) the search phase consists of 3 major steps.

1. ***select w.r.t UCB***: the calculation of UCB follows the standard definition in Auer et al. (2002). We present the equation of π_{UCB} in Fig. 3. The input to π_{UCB} is the number of visits of current node $n(s)$ and next nodes $n(s,a)$, and the next node value $v(s,a)$, where s stands for a node, and a stands for an action. Therefore, the next node is deterministic given s and a. At node j, the number of visits $n(s)$ is determined by the number of samples in $D_t \cap \Omega_j$, while $v(s)$ is determined by $\hat{V}(\Omega_j)$. The selecting policy π_{ucb} takes the node with the largest UCB score. Starting from $root$, we follow π_{ucb} to traverse down to a leaf (Alg. 2 line 7-13).
2. ***sampling from a leaf***: Fig. 3 highlights the sequence of actions impose several linear constraints on the search space Ω, bounding a polytope for the leaf. Alg. 2 line 1-6 explains how to obtain constraints from an action sequence. There are various techniques to perform uniform sampling in a polytope with linear constraints, e.g. George and McCulloch (1993); Neal et al. (2003). Here we use a variant of Markov Chain Monte Carlo (MCMC) sampler (PyMC) to get uniformly distributed samples.
3. ***back-propagate reward***: after training the sampled network, LaNAS back-propagates the reward, i.e. accuracy, to update the node statistics $n(s)$ and $v(s)$. It also back-propagates the sampled network so that every parent node j keeps the network in $D_t \cap \Omega_j$ for training.

3.4 PARTITION ANALYSIS

The sample efficiency, i.e. the number of samples to find the global optimal, is closely related to the partition quality of each tree nodes. Here we seek an upper bound for the number of samples in the leftmost leaf (the most promising region) to characterize the sample efficiency.

Assumption 1 *Given a search domain Ω containing finite samples N, there exists a probabilistic density f such that $P(a < v < b) = \int_a^b f(v)dv$, where v is the performance of a network* $\mathbf{a}$.

With this assumption, we can count the number of networks in the accuracy range of $[a,b]$ by $N * P(a \leq v \leq b)$. Since $v \in [0,1]$ and $\sigma_(v) < \infty$, the following holds (Mallows (1991))

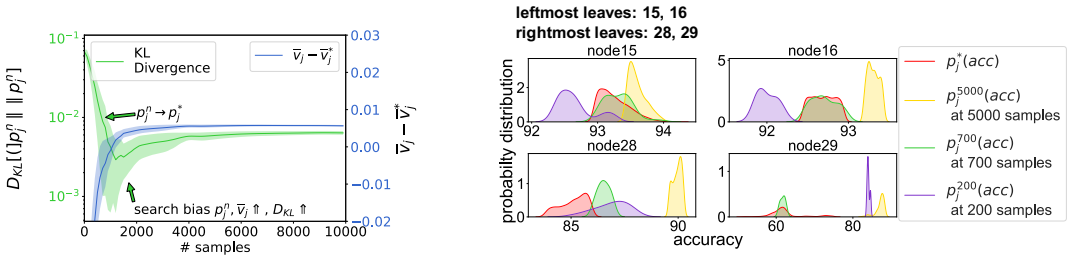
$$|E(\overline{v} - M_v)| < \sigma_v \tag{1}$$

$\overline{v}$ is the mean performance in Ω, and M_v is the median performance. Note $v \in [0,1]$, and let's denote $\epsilon = |\hat{v} - \overline{v}|$. Therefore, the maximal distance from $\hat{v}$ to M_v is $\epsilon + \sigma_v$; and the number of networks falling between $\hat{v}$ and M_v is $N * max(\int_{\hat{v}-\epsilon-\sigma_v}^{M_v} f(v)dv, \int_{M_v}^{\hat{v}+\epsilon+\sigma_v} f(v)dv)$, denoted as δ. Therefore, the root partitions Ω into two sets that have $\leq \frac{N}{2} + \delta$ architectures.

Theorem 1 *Given a search tree of height = h, the sub-domain represented by the leftmost leaf contains at most $2 * \delta_{max}(1 - \frac{1}{2^h}) + \frac{N}{2^h}$ architectures, and δ_{max} is the largest partition error from the node on the leftmost path.*

Proof: In the worst scenario, the left child is always assigned with the large partition; and let's recursively apply this all the way down to the leftmost leaf h times, resulting in $\delta^h + \frac{\delta^{h-1}}{2} + \frac{\delta^{h-2}}{2^2} + ... + \frac{N}{2^h} \leq 2 * \delta_{max}(1 - \frac{1}{2^h}) + \frac{N}{2^h}$. δ is related to ϵ and σ_v; note $\delta \downarrow$ with more samples as $\epsilon \downarrow$, and σ_v becomes more accurate.

The analysis indicates that LaNAS is approximating the global optimum at the speed of $N/2^h$, suggesting 1) the performance improvement will remain near plateau as $h \uparrow$ (verified by Fig 6a), while the computational costs ($2^h - 1$ nodes) exponentially increase; 2) the performance improvement w.r.t random search (cost $\sim N/2$) is more obvious on a large search space (verified by Fig.5 (a)$\rightarrow$(c)).



(a) Search dynamics: KL-divergence and mean distance

(b) Search dynamics: sample distribution vs. dataset distribution

Figure 4: **Evaluations of search dynamics**: (a) sample distribution p_j approximates dataset distribution p_j^* when the number of samples $n \in [200, 700]$. The search algorithm then zooms into the promising sub-domain, as shown by the growth of $\bar{v}_j$ when $n \in [700, 5000]$. (b) KL-divergence of p_j and p_j^* dips and bounces back. $\bar{v} - \bar{v}^*$ continues to grow, showing the average metric $\bar{v}$ over different nodes becomes higher when the search progresses.

4 Experiment

We performed extensive experiments on both offline collected benchmark datasets (e.g., NAS-Bench Ying et al. (2019)) and open search domain to validate the effectiveness of LaNAS.

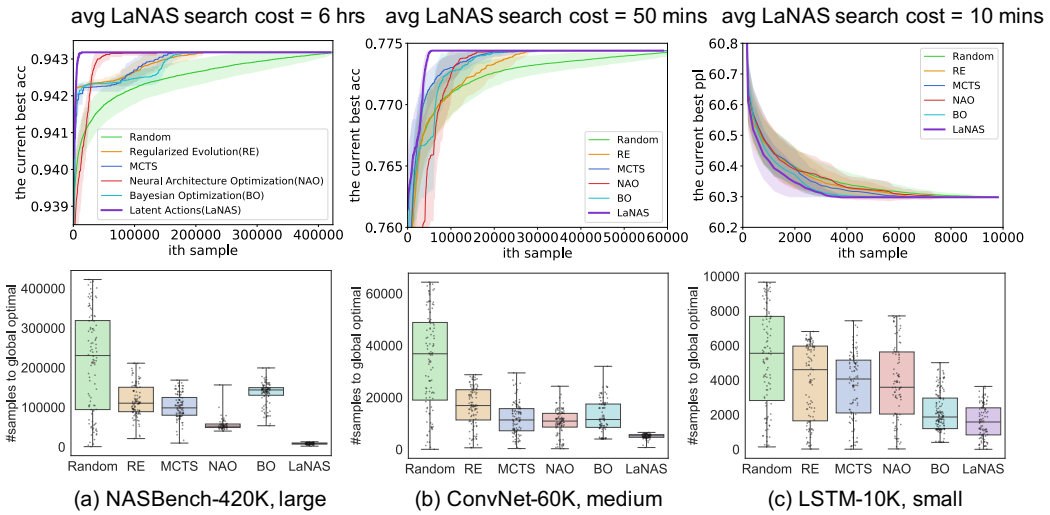
4.1 Analysis of Search Algorithm

We analyze LaNAS by looking into the dynamics of sample distributions on tree leaves during the search. We used NASBench-101 in experiments to form a constrained search space that contains 420K models. NASBench-101 provides us with the true distribution of model accuracy, given any subset of model specifications, or equivalently a collection of actions (or constraints). By construction, left nodes contain regions of the good metric while right nodes contain regions of the poor metric. Therefore, at each node j, we can construct *reference* distribution $p_j^*(v)$ by training toward the entire dataset to partition the search space into small regions with concentrated performances on leaves. We compare $p_j^*(v)$ with the estimated distribution $p_j^n(v)$, where n is the number of accumulated samples in $D_t \cap \Omega_j$ at the node j. Since the *reference* distribution $p_j^*(v)$ is static, comparing $p_j^n(v)$ to $p_j^*(v)$ enables us to see the variations of Ω_j on tree leaves w.r.t the growing samples. To provide a quantitative view, we also calculated the KL-divergence $D_{KL}[p_j^n \| \| p_j^*]$, and their mean value $\bar{v}_j^* = \mathbb{E}_{p_j^*}[v]$ and $\bar{v}_j = \mathbb{E}_{p_j^n}[v]$.

In our experiments, we use a complete *binary* tree with the height of 5. We label nodes 0-14 as internal nodes, and nodes 15-29 as leaves. By definition, $\bar{v}_{15}^* > \bar{v}_{16}^* ... > \bar{v}_{29}^*$ reflected by $p_{15,16,28,29}^*$ in Fig. 4b. At the beginning of the search ($n = 200$ for random initialization), $p_{15,16}^{200}$ are expected to be smaller than $p_{15,16}^*$, and $p_{28,29}^{200}$ are expected to be larger than $p_{15,16}^*$; because the tree still learns to partition at that time. With more samples ($n = 700$), p_j starts to approximate p_j^*, manifested by the increasing similarity between $p_{15,16,28,29}^{700}$ and $p_{15,16,28,29}^*$, and the decreasing D_{KL} in Fig. 4a. This is because MCTS explores the under-explored regions. As the search continues ($n \to 5000$), LaNAS explores deeper into promising regions and p_j^n is biased toward the region with good performance, deviated from p_j^*. As a result, D_{KL} bounces back in Fig. 4a. The mean accuracy of $p_{15}^{700,5000} > p_{16}^{700,5000} > p_{28}^{700,5000} > p_{29}^{700,5000}$ in Fig. 4(b) indicates that LaNAS successfully minimizes the variance of rewards on a search path making architectures with similar metrics concentrated in a region, and LaNAS correctly ranks the regions on tree leaves. These search dynamics show how our model adapts to different stages during the course of the search, and validate its effectiveness for the purpose of searching.

4.2 Performance on NAS Datasets

Evaluating on NAS Datasets: We use NASBench-101 (Ying et al. (2019)) as one benchmark that contains over 420K NASNet CNN models. For each network, it records the architecture information and the associated accuracy for fast retrieval by NAS algorithm, avoiding time-consuming model retraining. In addition, we construct two more datasets for benchmarking, ConvNet-60K (60K plain ConvNet models, VGG-style, no residual connections, trained on CIFAR-10) and LSTM-10K (10K



(a) NASBench-420K, large (b) ConvNet-60K, medium (c) LSTM-10K, small

Figure 5: **Evaluations of sample-efficiency** on NASBench, ConvNet-60K and LSTM-10K. Each search algorithm is repeated 100 times on each datasets with different random seeds. The top row shows the time-course of the search algorithms (current best accuracy with interquartile range), while the bottom row illustrates the number of samples to reach the global optimum.

LSTM cells trained on PTB) to further validate the effectiveness of LaNAS. These 3 datasets cover a diverse of tasks and sizes of search space ranging from small (10K), medium (60K) to large (420K). In each task, we perform an independent search for 100 times with different random seeds. The mean performance, along with the 25th and 75th percentile, is shown in Fig.5. The structure of search tree is consistent across all 3 tasks, using a tree with height = 5, and #selects = 100. We randomly pick 400 networks to initialize LaNAS, which is counted into the total sample complexity.

Baselines: we compare LaNAS with 5 NAS methods that cover a diverse types of search algorithms, ranging from Random Search, Bayesian Optimization, gradient based NAS, evolutionary algorithms, to Monte Carlo Tree Search (MCTS). `Random Search` randomly picks a sample from a dataset of size n, and it can find the global optimum in expected $n/2$ samples. It is fully parallelizable and agnostic to both datasets and tasks. `Regularized Evolution` (Real et al. (2019)) is a type of evolutionary algorithm that keeps a pool of top performing architectures for random mutations, and is applied in AmoebaNet (Real et al. (2018)) that achieves SoTA performance for image recognition. Our implementation of `Regularized Evolution` is from NASBench-101[1]. AlphaX (Wang et al. (2018)) adds a surrogate model in `MCTS` to boost the efficiency by predicting the architecture performance, and `MCTS` has succeeded in Go playing (Silver et al. (2016)) that render similar challenges to NAS in exploring a vast state space. We used AlphaX [2] as our `MCTS` implementation, and we set #prediction=1 for a rollout. `Neural Architecture Optimization(NAO)` (Luo et al. (2018)) is a gradient based NAS method that finds the architecture on a performance predictor in a continuous embedding space with gradient descent. We adopted NAO from its public release[3]. To modify `NAO` for NAS datasets, the new sample is selected by predicting the networks in the remaining dataset, which is equivalent of solving a global optimization on the predictor. `Bayesian Optimization(BO)` is a competitive method for the black box optimization, and it is applied in AutoML (Hutter et al. (2011)) and NAS (Zhou et al. (2019)). We implemented `BO` that used Gaussian Process as the surrogate model and expected improvement as the acquisition function based on here[4]. All the baseline methods propose 100 architectures at every search iteration, i.e. #select = 100, to be consistent with LaNAS.

Analysis of Results: Fig. 5 demonstrates that LaNAS consistently outperforms the baselines by significant margins on three separate tasks. Particularly, on NASBench, LaNAS is on average using **22x**, **14.6x**, **12.4x**, **6.8x**, **16.5x** fewer samples than `Random Search`, `Regularized Evolution`,

[1]https://github.com/google-research/nasbench/blob/master/NASBench.ipynb
[2]https://github.com/linnanwang/AlphaX-NASBench101
[3]https://github.com/renqianluo/NAO
[4]http://krasserm.github.io/2018/03/21/bayesian-optimization/

Table 1: Results on CIFAR-10, c/o is cutout.

Model	Filters	Params	Top1 err	M	GPU days
NASNet-A+c/o	32	3.3 M	2.65	20000	2000
AmoebaNet-B+c/o	36	2.8 M	$2.55_{\pm 0.05}$	27000	3150
PNASNet-5	48	3.2 M	$3.41_{\pm 0.09}$	1160	225
NAO	36	10.6 M	3.18	1000	200
ENAS+c/o	36	4.6 M	2.89	-	0.45
DARTS+c/o	36	3.3 M	$2.76_{\pm 0.09}$	-	1.5
BayesNAS+c/o	32	3.4 M	$2.81_{\pm 0.04}$	-	0.2
ASNG-NAS+c/o	32	3.9 M	$2.83_{\pm 0.14}$	-	0.11
LaNet	32	3.2 M	$\mathbf{3.13}_{\pm 0.03}$	803	150
LaNet+c/o	32	3.2 M	$\mathbf{2.53}_{\pm 0.05}$	803	150
NAO+c/o	128	128.0 M	2.11	1000	200
AmoebaNet-B+c/o	128	34.9 M	$2.13_{\pm 0.04}$	27000	3150
LaNet+c/o	128	38.7 M	$\mathbf{1.99}_{\pm 0.02}$	803	150

M: number of samples selected.

Table 2: Results on ImageNet (mobile setting)

Model	FLOPs	Params	top1 / top5 err
NASNet-A (Zoph et al. (2018))	564M	5.3 M	26.0 / 8.4
NASNet-B (Zoph et al. (2018))	488M	5.3 M	27.2 / 8.7
NASNet-C (Zoph et al. (2018))	558M	4.9 M	27.5 / 9.0
AmoebaNet-A (Real et al. (2018))	555M	5.1 M	25.5 / 8.0
AmoebaNet-B (Real et al. (2018))	555M	5.3 M	26.0 / 8.5
AmoebaNet-C (Real et al. (2018))	570M	6.4 M	**24.3 / 7.6**
PNASNet-5 (Liu et al. (2018a))	588M	5.1 M	25.8 / 8.1
DARTS (Liu et al. (2018b))	574M	4.7 M	26.7 / 8.7
FBNet-C (Wu et al. (2018))	375M	5.5 M	25.1 / -
RandWire-WS (Xie et al. (2019))	583M	5.6 M	25.3 / 7.8
BayesNAS (Zhou et al. (2019))	-	3.9 M	26.5 / 8.9
LaNet	570M	5.1 M	**25.0 / 7.7**

`MCTS`, `NAO` and `Bayesian Optimization` to find the global optimum. On LSTM, LaNAS still performs the best despite that the dataset is small. The decreasing speedup w.r.t `RS` from a large to small dataset are consistent with our analysis in Sec.3.4.

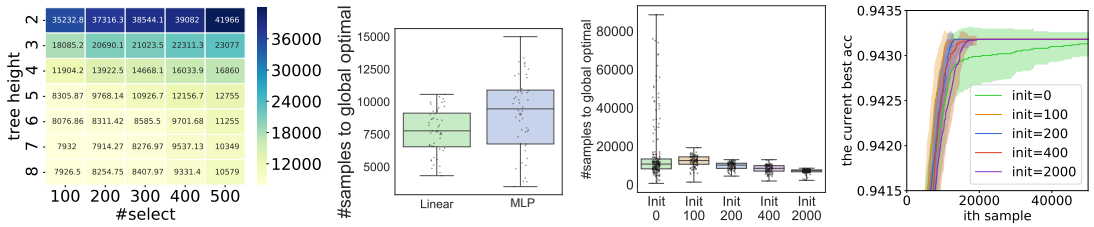
In LaNAS, the action space is learned so that in each node, the representing search space can be partitioned into a separate good/bad region using a linear classifier based on the network attributes (e.g., depth/width of the network). With the learned action space, the resulting search space clearly separates good/bad regions, and the search is biased towards the regions that contain better architectures, therefore becoming more efficient. Besides, with more samples, Fig. 4 shows LaNAS zooms in the promising regions pinpointing more promising regions to further improve the efficiency. In contrast, the action spaces of `MCTS` are manually pre-defined, e.g. adding layers, or setting kernels, resulting in a state space that provides undiscriminating rewards, even predicted from the surrogate model, for searching, e.g. Fig. 2(a), thereby less sample efficient. `Random Search` gives poor performance, in particular in a large search space like NASBench. `Regularized Evolution` utilizes a static exploration strategy that maintains a pool of top 500 architectures for random mutations, which is not guided by previous search experience. The surrogate model in `NAO` maps network attributes to accuracy that needs sufficient samples before generalizing well; while our model aims as an easier target that approximates the mean accuracy of a search space to progressively maps the attributes of a network into good/bad regions. Though `Bayesian Optimization(BO)` shows slightly worse performance than LaNAS on small LSTM-10K, scaling `BO` to high dimensional and hundreds of thousands of samples remain open (Wang et al. (2017)), as `BO` incurs O(n^3) cost and n is the number of samples. For example, `BO` took 8 hours to fit 30000 samples on a CPU, however, during the search, `BO` needs fit on the collected samples (up to 420K) for hundreds of thousands of search iterations, which is too expensive to benchmark. In contrast, on the same CPU, LaNAS only took 6 hours to complete the search on NASBench-101. To mitigate the scalability issues, on NASBench-420K and ConvNet-60K, we apply `BO` on collected samples using a similar strategy to SGD that splits the entire samples into batches, then iteratively fits `BO` at a batch size of 2000 for 2 epochs. So, the result is worse than using full data points.

4.3 Performance on Open Domain Search

To be consistent with existing works, we also evaluated LaNAS on the NASNet open search domain.

Search space: Our search space is consistent with the widely adopted NASNet (Zoph et al. (2018)). The operations are 3x3, 5x5 and 7x7 max pool, 3x3, 5x5, 7x7 depth-separable conv, 1x1 convolution and identity. The number of nodes within a cell is 5, and we constructed the network by stacking a cell 6 times. Existing works in NAS tends to have very different sizes of search space according to Table.3 in Radosavovic et al. (2019). Though we follow NASBench-101 and PNAS to share the normal and reduction cell in constructing the network, our search space contains 4.36×10^{13} architectures, which is decent large for evaluations.

Experiment setup: During the search phase, we set filters to 32 and pre-trained an architecture for 120 epochs. Then we selected top-20 architectures collected from the search, and re-trained them



(a) tree height vs. #select (b) choice of classifier (c) #samples for init (d) #samples for init

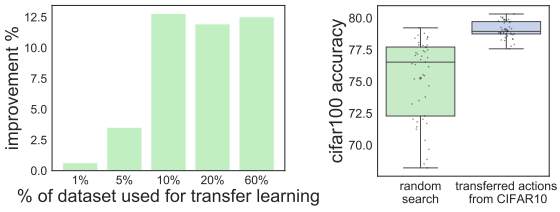
Figure 6: **Ablation study**: (a) the effect of different tree heights and #select in MCTS. Number in each entry is #samples to reach global optimal. (b) the choice of predictor for splitting search space. (c-d) the effect of #samples for initialization toward the search performance.

for 600 epochs to acquire the final accuracy in Table. 1. We reused the training logic from DARTS and their training settings. The height of search tree is 8; we used 200 random samples for the initialization, and #select = 50.

Table. 1 compares our results in the context of searching NASNet style architecture on CIFAR-10. We found the best performing architecture (LaNet) at the 803th sample. LaNet demonstrates an average accuracy of 97.47% (#filters = 32, #params = 3.22M) and 98.01% (#filters = 128, #params = 38.7M), which is better than all existing NAS-based results on CIFAR-10. It is worth noting that we achieved this accuracy with 33x fewer samples than AmoebaNet. Though one-shot learning methods, e.g. Liu et al. (2018b); Zhou et al. (2019); Akimoto et al. (2019), and their weight sharing variants, e.g. Pham et al. (2018), exhibit weaker performance, they render much lower search costs for greatly reducing network evaluation costs with techniques such as transfer learning or approximating final weights with only 1 training iteration. Fig. 5 indicates the search model is cheap to run, but evaluating networks during the search costs the majority of GPU days. Both NAO and ENAS show that weight sharing has greatly sped up network evaluations; our approach can also benefit from weight sharing to reduce the end-to-end search time. On the other hand, one shot learning methods show a small difference to LaNAS in accuracy, e.g. -0.8%, at much less cost. However, as suggested in Fig. 5, almost 99.9% search costs, i.e. samples, is at improving the last 1% accuracy. Therefore, one-shot learning is fast to yield a reasonable result, while LaNAS aims at giving the best accuracy with the fewest samples.

4.4 Transfer Learning

Transfer LaNet to ImageNet: Transferring the best performing architecture (found through searching) from CIFAR10 to ImageNet has already been a standard technique. Following the mobile setting (Zoph et al. (2018)), Table. 2 shows that LaNet found on CIFAR-10, when transferred to ImageNet mobile setting (FLOPs are constrained under 600M), achieves competitive performance.



(a) intra-task transfer (b) inter-task transfer

Figure 7: **Latent actions transfer**: learned latent actions can generalize within the same task or across different tasks, to further boost search efficiency.

Intra-task latent action transfer: We learn actions from a subset (1%, 5%, 10%, 20% 60%) of NASBench and test their transferability to the remaining dataset, as shown in Fig. 7a. Interestingly, the improvement remain steady after 10%. Consistent with Fig. 4, it is enough to use 10% of the samples to learn the action space.

Inter-task latent action transfer: We compare 100 architectures selected by LaNAS from sec.4.3 (on CIFAR10) with 100 random trials. Networks are trained for 100 epochs on CIFAR-100 and their performances are compared. Fig. 7b indicates that inter-task action transfer is also beneficial.

4.5 ABLATION STUDIES

The effect of tree height and #selects: Fig. 6a relates tree height (h) and the number of selects (#selects) to the search performance. In Fig. 6a, each entry represents #samples to achieve optimality on NASBench, averaged over 100 runs. A deeper tree leads to better performance, since the model space is partitioned by more leaves. Similarly, small #select results in more frequent updates of action space allowing the tree to make up-to-date decisions, and thus leads to improvement. On the other hand, the number of classifiers increases exponentially as the tree goes deeper, and a small #selects incurs frequent learning phase. Therefore, both can significantly increase the computation cost.

Choice of classifiers: Fig.6b shows that using a linear classifier performs better than an multi-layer perceptron (MLP) classifier. This indicates that adding complexity to decision boundary of actions may not help with the performance. Conversely, performance can get degraded due to potentially higher difficulties in optimization.

#samples for initialization:We need to initialize each node classifier properly with a few samples to establish the initial boundary in the search space. As shown in Fig.6c6d, cold start is necessary (init > 0 is better than init = 0). Also, small init=100-400 converges to top 5% performance much faster than init=2000 (Fig. 6c), while init=2000 gets the best performance faster (Fig. 6d).

5 FUTURE WORK

Recent work on shared model (Luo et al. (2018); Pham et al. (2018)) improves the training efficiency by reusing trained components from the similar previously explored architectures, e.g., weight sharing (Liu et al. (2018b); Luo et al. (2018); Pham et al. (2018)). Our work focuses on sample-efficiency and is complementary to the above techniques.

To encourage reproducibility in NAS research, various of architecture search baselines have been discussed in Sciuto et al. (2019); Li and Talwalkar (2019). We will also open source the proposed framework, together with the three NAS benchmark datasets used in our experiments.

REFERENCES

Y. Akimoto, S. Shirakawa, N. Yoshinari, K. Uchida, S. Saito, and K. Nishida. Adaptive stochastic natural gradient method for one-shot neural architecture search. *arXiv preprint arXiv:1905.08537*, 2019.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

B. Baker, O. Gupta, N. Naik, and R. Raskar. Designing neural network architectures using reinforcement learning. *arXiv preprint arXiv:1611.02167*, 2016.

Y. Chen, T. Yang, X. Zhang, G. Meng, C. Pan, and J. Sun. Detnas: Neural architecture search on object detection. *arXiv preprint arXiv:1903.10979*, 2019.

E. I. George and R. E. McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

G. Ghiasi, T.-Y. Lin, R. Pang, and Q. V. Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. *arXiv preprint arXiv:1904.07392*, 2019.

F. Hutter, H. H. Hoos, and K. Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *International conference on learning and intelligent optimization*, pages 507–523. Springer, 2011.

L. Li and A. Talwalkar. Random search and reproducibility for neural architecture search. *arXiv preprint arXiv:1902.07638*, 2019.

C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L.-J. Li, L. Fei-Fei, A. Yuille, J. Huang, and K. Murphy. Progressive neural architecture search. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 19–34, 2018a.

C. Liu, L.-C. Chen, F. Schroff, H. Adam, W. Hua, A. Yuille, and L. Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. *arXiv preprint arXiv:1901.02985*, 2019.

H. Liu, K. Simonyan, and Y. Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018b.

R. Luo, F. Tian, T. Qin, E. Chen, and T.-Y. Liu. Neural architecture optimization. In *Advances in neural information processing systems*, pages 7816–7827, 2018.

M.-T. Luong, D. Dohan, A. W. Yu, Q. V. Le, B. Zoph, and V. Vasudevan. Exploring neural architecture search for language tasks. *arXiv preprint arXiv:1901.02985*, 2018.

C. Mallows. Letters to the editor. *The American Statistician*, 45(3):256–262, 1991. doi: 10.1080/00031305.1991.10475815.

R. M. Neal et al. Slice sampling. *The annals of statistics*, 31(3):705–767, 2003.

R. Negrinho and G. Gordon. Deeparchitect: Automatically designing and training deep architectures. *arXiv preprint arXiv:1704.08792*, 2017.

H. Pham, M. Guan, B. Zoph, Q. Le, and J. Dean. Efficient neural architecture search via parameter sharing. In *International Conference on Machine Learning*, pages 4092–4101, 2018.

I. Radosavovic, J. Johnson, S. Xie, W.-Y. Lo, and P. Dollár. On network design spaces for visual recognition. *arXiv preprint arXiv:1905.13214*, 2019.

E. Real, S. Moore, A. Selle, S. Saxena, Y. L. Suematsu, J. Tan, Q. V. Le, and A. Kurakin. Large-scale evolution of image classifiers. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2902–2911. JMLR. org, 2017.

E. Real, A. Aggarwal, Y. Huang, and Q. V. Le. Regularized evolution for image classifier architecture search. *arXiv preprint arXiv:1802.01548*, 2018.

E. Real, A. Aggarwal, Y. Huang, and Q. V. Le. Regularized evolution for image classifier architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4780–4789, 2019.

C. Sciuto, K. Yu, M. Jaggi, C. Musat, and M. Salzmann. Evaluating the search phase of neural architecture search. *arXiv preprint arXiv:1902.08142*, 2019.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

D. R. So, C. Liang, and Q. V. Le. The evolved transformer. *arXiv preprint arXiv:1901.11117*, 2019.

L. Wang, Y. Zhao, Y. Jinnai, and R. Fonseca. Alphax: exploring neural architectures with deep neural networks and monte carlo tree search. *arXiv preprint arXiv:1805.07440*, 2018.

L. Wang, Y. Zhao, Y. Jinnai, Y. Tian, and R. Fonseca. Alphax: exploring neural architectures with deep neural networks and monte carlo tree search. *arXiv preprint arXiv:1903.11059*, 2019.

Z. Wang, C. Gehring, P. Kohli, and S. Jegelka. Batched large-scale bayesian optimization in high-dimensional spaces. *arXiv preprint arXiv:1706.01445*, 2017.

B. Wu, X. Dai, P. Zhang, Y. Wang, F. Sun, Y. Wu, Y. Tian, P. Vajda, Y. Jia, and K. Keutzer. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. *arXiv preprint arXiv:1812.03443*, 2018.

S. Xie, A. Kirillov, R. Girshick, and K. He. Exploring randomly wired neural networks for image recognition. *arXiv preprint arXiv:1904.01569*, 2019.

C. Ying, A. Klein, E. Real, E. Christiansen, K. Murphy, and F. Hutter. Nas-bench-101: Towards reproducible neural architecture search. *arXiv preprint arXiv:1902.09635*, 2019.

H. Zhou, M. Yang, J. Wang, and W. Pan. Bayesnas: A bayesian approach for neural architecture search. *arXiv preprint arXiv:1905.04919*, 2019.

B. Zoph and Q. V. Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.

B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.

A Appendix

A.1 NAS Datasets and Experiment Setup:

NAS dataset enables directly querying a model's performance, e.g. accuracy. This allows for truly evaluating a search algorithm by repeating hundreds of independent searches without involving the actual training. NASBench-101 Ying et al. (2019) is the only publicly available NAS dataset that contains over 420K DAG style networks for image recognition. However, a search algorithm might overfit NASBench-101, losing the generality. This motivates us to collect another 2 NAS datasets, one is for image recognition using sequential CNN and the other is for language modeling using LSTM.

Collecting ConvNet-60K dataset: following a similar set in collecting 1,364 networks in sec.2, free structural parameters that can vary are: network depth $D = \{1, 2, 3, 4, 5, 6, 7, 8\}$, number of filter channels $C = \{32, 64, 96\}$ and kernel size $K = \{3 \times 3, 5 \times 5\}$. We train every possible architecture 100 epochs, and collect their final test accuracy in the dataset.

Collecting LSTM-10K dataset: following a similar LSTM cell definition in Pham et al. (2018), we represent a LSTM cell with a connectivity matrix and a node list of operators. The choice of operators is limited to either a fully connected layer followed by RELU or an identity layer, while the connectivity is limited to 6 nodes. We randomly sampled 10K architectures from this constrained search domain, and training them following the exact same setup in Liu et al. (2018b). Then we keep the test perplexity as the performance metric for NAS.

A.2 Open Domain Search and Experiment Setup:

The best convolutional cell found by LaNAS is visualized below. We constructed LaNet by stacking this cell 6 times.