GreedyNASv2: Greedier Search with a Greedy Path Filter

Tao Huang\textsuperscript{1,2}, Shan You\textsuperscript{1,3}*, Fei Wang\textsuperscript{4}
Chen Qian\textsuperscript{1}, Changshui Zhang\textsuperscript{3}, Xiaogang Wang\textsuperscript{1,5}, Chang Xu\textsuperscript{2}
\textsuperscript{1}SenseTime Research \textsuperscript{2}School of Computer Science, Faculty of Engineering, The University of Sydney
\textsuperscript{3}Department of Automation, Tsinghua University, Institute for Artificial Intelligence, Tsinghua University (THUAI), Beijing National Research Center for Information Science and Technology (BNRist)
\textsuperscript{4}University of Science and Technology of China \textsuperscript{5}The Chinese University of Hong Kong

Abstract

Training a good supernet in one-shot NAS methods is difficult since the search space is usually considerably huge (e.g., $13^{21}$). In order to enhance the supernet’s evaluation ability, one greedy strategy is to sample good paths, and let the supernet lean towards the good ones and ease its evaluation burden as a result. However, in practice the search can be still quite inefficient since the identification of good paths is not accurate enough and sampled paths still scatter around the whole search space. In this paper, we leverage an explicit path filter to capture the characteristics of paths and directly filter those weak ones, so that the search can be thus implemented on the shrunk space more greedily and efficiently. Concretely, based on the fact that good paths are much less than the weak ones in the space, we argue that the label of “weak paths” will be more confident and reliable than that of “good paths” in multi-path sampling. In this way, we cast the training of path filter in the positive and unlabeled (PU) learning paradigm, and also encourage a path embedding as better path/operation representation to enhance the identification capacity of the learned filter. By dint of this embedding, we can further shrink the search space by aggregating similar operations with similar embeddings, and the search can be more efficient and accurate. Extensive experiments validate the effectiveness of the proposed method GreedyNASv2. For example, our obtained GreedyNASv2-L achieves 81.1% Top-1 accuracy on ImageNet dataset, significantly outperforming the ResNet-50 strong baselines.

1. Introduction

Neural architecture search (NAS) aims to boost the performance of deep learning by seeking an optimal architecture in the given space, and it has achieved significant improvements in the sight of applications, such as image classification [11, 26, 28, 32, 34] and object detection [2, 10]. One-shot NAS [11, 15, 23–25, 33, 34] stands out from the literature of NAS for the sake of its decent searching efficiency. Instead of exhaustively training each possible architecture, one-shot NAS fulfills the searching in an only one-shot trial, where a supernet is leveraged to embody all candidate architectures (i.e., paths). Each path can be parameterized by the corresponding weights within the supernet, and thus gets trained, evaluated, and ranked. Typical uniform sampling (SPOS) [11] is usually adopted to train the supernet because of the feasible single-path memory consumption and being friendly to large-scale datasets.

The architecture search space in NAS could be considerably huge (e.g., $13^{21}$). Equally treating different paths and uniformly sampling them from the supernet could lead to inappropriate training of the supernet, as the weak paths would disturb the highly-shared weights. Various sampling strategies have thus been proposed to address this issue, such as fair sampling [3] and Monte-Carlo tree search [23]. We are particularly interested in the strategy of multi-path sampling with rejection by GreedyNAS [34], which identifies good paths from weak paths and then only greed-
ily updates those good ones; it is easy to implement and more suitable for large search spaces among these methods. Working on the whole search space, GreedyNAS has to safely allocate only a medium level of partition of good paths (e.g., only 5 out of 10) to ensure a high probability of sampled paths being good. But the search will become infeasible and limited if the search space grows larger with more operation choices. Besides, GreedyNAS needs to maintain a candidate pool to recycle paths, which limits the number of stored paths, and many elite paths could be missed.

In this paper, we propose GreedyNASv2 to power the multi-path sampler with explicit search space shrinkage for one-shot NAS, which targets on a greedier search space with an tiny (e.g., only 1%) proportion of paths treated as “good paths”. Since good paths are usually much less than weak paths, the probability of picking out a good path by a multi-path sampler could be smaller than that of sampling weak paths. If weak paths can be captured with confidence, we can easily screen them out from the searching space and execute a greedier search on the shrunk space. By doing so, the supernet only needs to focus on evaluating those not too bad paths (potentially-good paths), which benefits the overall searching performance and efficiency simultaneously.

The key is then to learn a path filter to identify those weak paths from the entire architecture search space. Though it is hard to find a good path, we can have high confidence about weak paths in the multi-path sampling. These identified weak paths with confidence can be regarded as positive examples to be thrown away. As a precaution, the remaining paths in the search space are taken as unlabeled examples, as they may contain both weak paths (positive examples to be thrown away) and good paths (negative examples not to be thrown away). The learning of this path filter can thus be formulated as the Positive-Unlabeled (PU) learning problem [9,17]. Once the path filter has been well trained, a given new path can be efficiently predicted to specify whether it is weak or not. A path embedding is also learned with the path filter to encode the path as a better path representation. Since the path embedding is learned in the weak/good sense, if two operations have similar embeddings, it means that both operations have similar or even the same impact on discriminating paths, and they can be thus merged. This enables a greedy shrinkage of operations, which is expected to work together with the path shrinkage to boost the searching performance and efficiency further.

We conduct extensive experiments on ImageNet dataset to validate the effectiveness of our proposed GreedyNASv2. Compared to the baseline methods SPOS and GreedyNAS, our proposed method achieves better performance with less search cost. To further investigate our superiority, we even search on a larger space, which has \( \sim 10^4 \times \) architectures compared to the commonly-used MobileNetV2-SE search space, and the results show that our searched model outperforms state-of-the-art NAS models. The performance on different scales of search spaces are illustrated in Figure 1. Besides, we also compare the searching performance on a recent benchmark NAS-Bench-Macro [23] for one-shot NAS. Ablation studies show that our GreedyNASv2 effectively samples better architectures than uniform sampling and the multi-path sampler in GreedyNAS.

2. Related Work

2.1. NAS with search space shrinkage

Path-level shrinkage. To obtain a path-level shrinkage on search space, GreedyNAS [34] proposes a candidate pool to store those evaluated good paths and samples from it using an exploration-exploitation strategy. MCT-NAS [23] proposes to sample architectures with the guidance of Monte-Carlo tree search; hence the good paths can be sampled with better exploration and exploitation balance. However, the limited size (e.g., 1000) of candidate pool in GreedyNAS is too aggressive to train the elite paths with
enough diversity, and the exponentially-increased Monte-
Carlo tree makes the MCT-NAS difficult to scale to larger
search spaces.

**Operation-level shrinkage.** Operation-level shrinkage
is also an effective way to reduce both training parameters
and the size of search space. Some methods [14, 22]

design importance metrics to identify the good operations
and drop the weak ones. For example, ABS [14] measures
the importance of each operation using the angle between
its trained weights and initialized weights; BS-NAS [22]
proposes a channel-level importance metric by measuring
a number of architectures on the validation dataset. How-
evertheless, these methods only consider operation-level statistics,
while for each specific architecture, the preferences of op-

erations may be different. On the other hand, NSENet [4]
proposes to learn the importance using additional learnable
indicators after each operation, which is learned by simulat-
ing the gradients of binary-selected architectures. However,
this simulation of gradients introduces approximation errors
and also increases memory consumption.

In this paper, we perform both path-level and operation-
level shrinkage using a path filter. The path filter is con-
structed by a binary classifier, which efficiently filters the
weak paths and generalizes well to the whole search space;

hence we can filter the weak paths more greedily. Furthe-

more, we can perform operation-level shrinkage without ex-
tra costs by measuring the learned operation embeddings
in the path filter. This operation merging strategy holds nat-
urally since the operations with similar embeddings would
have similar predictions and thus similar performance.

**2.2. Positive-unlabeled learning**

Positive-Unlabeled (PU) classification is a problem of
training a binary classifier from only positive and unlabeled
data [9, 17]. Many effective methods [1, 8, 16] are proposed
to train a good binary classifier in PU learning. Specifi-
cally, uPU [8] rewrites the classification risk in terms of
the distributions over positive and unlabeled samples, and
obtains an unbiased estimator of the risk without negative
samples. To overcome the overfitting problem in uPU, a
non-negative risk estimator is proposed in nnPU [16]. One
recent approach VPU [1] proposes a variational principle
for PU learning without involving class prior estimation or
any other intermediate estimation problems. In this paper,
we implement VPU to learn our path filter.

**3. Revisiting Multi-path Sampler**

In single path one-shot NAS [11, 23, 34], the search space
is treated as an over-parameterized supernet \( \mathcal{N} \), in which
the searching layers are stacked sequentially, and each layer
is required to select one operation from candidate oper-
ations. Assume the supernet has \( L \) layers and \( N \) can-
didate operations \( O = \{o_i\}, \forall i = 1, 2, ..., N \), then each ar-

chitecture can be represented by a tuple with size \( L \), i.e.,
\[ \alpha = (o^{(1)}, o^{(2)}, ..., o^{(L)}) \], where \( o^{(j)} \in O, \forall j = 1, 2, ..., L \).

As a result, the search space \( \mathcal{A} \) is of size \( |\mathcal{A}| = N^L \). With
a pre-defined supernet, the NAS procedure is split into two
stages: supernet training and search. During training, the
supernet is optimized by alternately sampling paths and up-
dating their corresponding weights. Thereafter, the optimal
path can be determined as the one with the highest accuracy
on a hold-out validation set.

Although the supernet shares weights with all architec-
tures, it still has \( \sim N \times \) parameters than a common path.
For example, the benchmark MobileNetV2-SE search space
has 13 operations and 46M parameters for supernet, while a
path only has \( \sim 5 \)M parameters. With such a large supernet,
it is harsh to optimize all the architectures well and evaluate
them accurately. Therefore, instead of sampling paths uni-
formly [11], GreedyNAS [34] proposes a multi-path sam-
pling with rejection to greedily sample those potentially-


![Figure 3](image_url)

**Figure 3.** Confidence \( P(Q) \) of sampling at least \( k \) good (weak)
paths out of \( m = 10 \) paths with proportion \( p \) of good paths.
However, this is not enough. Since we target at the optimal architecture, the candidate elite paths are supposed to be way less than the weak ones, which means that $p \ll 0.5$ holds naturally, and thus we can have an actually shrunk space to boost the searching. Frustratingly, the confidence will degrade accordingly; for example, with $p = 0.1$, the previous confidence $\mathbb{P}$ will be only 0.16%. Though GreedyNAS leverages a candidate pool to recycle paths, many elite paths could be missed since it heavily relies on the limited number of stored paths (e.g., 1000).

### 3.2. Turning tables with weak paths

Sampling good paths is frustratingly ineffective in a more greedy space since the confidence collapses as a failure. In contrast, since the good prior $p$ is low, the search space will be glutted with weak paths, and the probability of sampling a weak path $q := 1 - p$ is thus large accordingly. Similarly, based on Theorem 1, in multi-path sampling the confidence of sampled weak paths goes even larger, denoted as $Q := \sum_{j=k}^{m} C_{j}^{m} q^{j}(1 - q)^{m-j}$. For example, with $q = 0.9$ ($p = 0.1$), the probability of sampling at least 5 weak paths out of 10 is very high ($Q = 99.99\%$). See Figure 3 for more details.

Now the tables have been turned. If we can sample weak paths with high confidence, we can easily rule them out from the entire search space and implement a more greedily searching on the shrunk space, thus boosting the searching performance as well. Then the question goes to: how can we leverage the sampled weak paths to identify a shrunk space composed of good paths? Intuitively, we encourage learning a path filter to encode the characteristics of sampled weak paths and identify the label of a given new path.

Nevertheless, during multi-path sampling, we only have confidence about weak paths, can we still learn a discriminative path filter as a decent binary classifier to predict labels of paths? The answer is affirmative; in the sequel, we will cast the learning as a typical Positive and Unlabeled (PU) problem.

### 4. Greedier Sampling with a Path Filter

After multi-path sampling, we now have confident weak paths; nevertheless, remaining paths are difficult to specify whether they are weak or good, since the corresponding confidence will be low. As a precaution, we regard the remaining paths (together with unsampled paths) as unlabeled examples, as they may contain both weak and good paths.

#### 4.1. Learning path filter as PU prediction

Here we want to learn a path filter with the identified weak paths (positive examples) and remaining paths (unlabeled examples). Formally, let us first consider a binary classification problem where the architectures $\mathbf{a} \in \mathcal{A}$ and class labels $y \in \{-1, +1\}$ are distributed according to a joint distribution $\mathbb{D}(\mathbf{a}, y)$, and the paths with positive label +1 denote weak paths to be discarded. In GreedyNASv2, we have positive dataset $\mathcal{P} = \{a_{1}, ..., a_{M}\}$ and unlabeled dataset $\mathcal{U} = \{a_{M+1}, ..., a_{N}\}$ sampled from the search space. The learning of path filter is thus cast as a Positive and Unlabeled (PU) learning problem, where a binary predictor $\Phi$ is learned based on $\mathcal{P}$ and $\mathcal{U}$ so that the class labels of unseen architectures can be accurately predicted.

As an introduction of PU learning, we first investigate the expected risk (classification loss) on the whole dataset of the commonly supervised learning (PN learning) as

$$R(\Phi) = \pi_{\mathcal{P}} \mathbb{E}_{\mathcal{P}}[l_{+}(\Phi(\mathbf{a}))] + (1 - \pi_{\mathcal{P}}) \mathbb{E}_{\mathcal{N}}[l_{-}(\Phi(\mathbf{a}))],$$  \hspace{1cm} (2)

where $\pi_{\mathcal{P}} = \mathbb{P}(y = +1)$ denotes the class prior of positive data, $\mathcal{N}$ refers to negative dataset, and $l_{+}$ and $l_{-}$ denote classification losses with

$$\mathbb{E}_{\mathcal{P}}[l_{+}(\Phi(\mathbf{a}))] = \frac{1}{|\mathcal{P}|} \sum_{\mathbf{a} \in \mathcal{P}} l(\Phi(\mathbf{a}), +1),$$

$$\mathbb{E}_{\mathcal{N}}[l_{-}(\Phi(\mathbf{a}))] = \frac{1}{|\mathcal{N}|} \sum_{\mathbf{a} \in \mathcal{N}} l(\Phi(\mathbf{a}), -1),$$  \hspace{1cm} (3)

which are the expectations of $l_{+}(\Phi(\mathbf{a}))$ on the positive dataset $\mathcal{P}$ and $l_{-}(\Phi(\mathbf{a}))$ on the negative dataset $\mathcal{N}$. 

11895
Nevertheless, the negative dataset $\mathcal{N}$ is unavailable in our PU learning setting. To train the model with positive and unlabeled data, the classical method uPU [8] encourages an unbiased formulation to the PU learning by rewriting the expectation of negative classification loss $E_{\mathcal{N}}[\log(1 - \Phi(a))]$ to

$$(1 - \pi_{\mathcal{P}})E_{\mathcal{N}}[\log(1 - \Phi(a))] = E_{\mathcal{U}}[\log(1 - \Phi(a))] - \pi_{\mathcal{P}}E_{\mathcal{P}}[\log(1 - \Phi(a))],$$

and thus Eq.(2) can be adapted to

$$R(\Phi) = \pi_{\mathcal{P}}E_{\mathcal{P}}[\log(1 - \Phi(a))] - \pi_{\mathcal{P}}E_{\mathcal{P}}[\log(1 - \Phi(a))].$$

However, such a method easily leads to severe overfitting, especially on deep neural classifiers. In our paper, to alleviate the above weakness, we leverage the learning of the selected operations, where $\Phi(a)$ is fed into $E$ to get the hidden states $H$ of the selected operations, where $A \in \mathbb{R}^{L \times H}$. Then we use a bi-directional LSTM to get the feature $f_a \in \mathbb{R}^H$ of the architecture. Finally, the architecture feature $f_a$ is fed into a binary classifier (two-layer perceptions with intermediate ReLU activation) to obtain the prediction.

Following the multi-path sampling strategy in GreedyNAS, each time we train the path filter, we randomly sample $m$ paths and evaluate them using the loss on a small validation set, which contains 1000 images sampled from the validation set. We sort those sampled paths with their losses in ascending order, and label the last $p$ percentage of paths as “weak paths” to build the positive dataset $\mathcal{P}$, while the unlabeled dataset $\mathcal{U}$ is constructed by uniformly sampling $10 \times p \times m$ paths from the search space. With the learning objective in Eq.(6), we train the path filter with datasets $\mathcal{P}$ and $\mathcal{U}$ every $t$ epochs in the training of supernet to ensure its accuracy. Once the path filter is trained, it can be used to predict a batch of uniformly-sampled paths, and filter those paths with positive labels, and the remained paths are treated as the potentially-good paths and used in optimization.

Stopping principle via path predictions. In training, if the supernet is trained well, the rankings of paths tend to be steady; hence GreedyNAS proposes an early stopping principle by measuring the steadiness of the candidate pool. We now propose a more accurate way by predicting more paths using the learned path filters, i.e.,

$$u := \frac{\sum_{a_i \in A_{\mathcal{r}}} 1_{\Phi_t(a_i) = \Phi_{t-1}(a_i)} > \beta}{M},$$

where $A_{\mathcal{r}}$ is a set of $M$ randomly sampled paths, $\Phi_t$ denotes the learned path filter at iteration $t$. $u$ measures the proportion of the same predictions in the last two path filters, if $u > \beta$, we believe the supernet has been trained enough, and its training can be stopped accordingly. We set $N = 10^4$ and $\beta = 0.9$ in our experiments.

Evolutionary search with path filter. We adopt evolutionary algorithm (EA) NSGA-II [6] to search architectures. Unlike SPOS [11] generates architectures randomly and GreedyNAS [34] only specifies an initial population, we use the learned path filter to filter the weak architectures generated by EA during the whole search, thus the search could be more efficient.

Algorithm 1 Training supernet with a greedy path filter.

**Input:** Supernet $\mathcal{N}$, path filter $\mathcal{P}$, max training iteration $T$, train dataset $\mathcal{D}_{tr}$, small validation dataset $\mathcal{D}_{val}$, predictor update interval $t$, number of evaluation paths $m$, weak path prior $q$, merge operation threshold $\theta$.

1: for $i = 1, \ldots, T$ do
2: $a_i \sim U(\mathcal{N})$;
3: while $i_{\text{weak}}(\mathcal{P}, a_i) \text{ do}$
4: $a_i \sim U(\mathcal{N})$;
5: end while
6: train($\mathcal{N}$, $a_i$, $\mathcal{D}_{tr}$); \(\triangleright\) train for one iteration
7: if $i \% t = 0$ then
8: sample $m$ paths $A = \{a_j\}_{j=1}^m \text{ i.i.d w.r.t. } a_j \sim U(\mathcal{N})$;
9: $s = \{\text{evaluate}(\mathcal{N}, a_j, \mathcal{D}_{val})\}_{j=1}^m$;
10: $A_{\text{weak}} = \text{lth percentile paths}$; \(\triangleright\) get last $q$ percentile paths
11: train predictor $\mathcal{P}$ with $A_{\text{weak}}$;
12: merge operations according to Section 4.2;
13: end if
14: end for
4.2. Shrinking operations with learned embeddings

The PU predictor distinguishes whether a path is good or bad using learned operation embeddings. If the embeddings of two operations \( o_a^{(i)} \) and \( o_b^{(i)} \) in layer \( i \) are totally the same, it means that for all the architectures, replacing \( o_a^{(i)} \) by \( o_b^{(i)} \) would not affect the classification results and vice versa. As a result, if two operations act similarly, we can greedily merge them and keep the less-costly one (e.g., the one with smaller FLOPs).

Cosine similarity is a commonly-used metric to measure the similarity of two vectors. Given two vectors \( x \) and \( y \), their cosine similarity \( S_c(x, y) \) is represented as

\[
S_c(x, y) = \frac{x \cdot y}{\|x\| \cdot \|y\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}.
\]

After each time we train the predictor, we will measure the cosine similarities between different operations in each layer. If the similarity between two operations is less than a pre-defined threshold \( s_{\text{thrd}} \), we then merge these two operations into one operation by keeping the one with smaller FLOPs and removing another one.

Formally, for operations \( \{o_1^{(i)}, o_2^{(i)}, ..., o_N^{(i)}\} \) in layer \( i \), it has \( C_N^2 \) combinations of pairs, we compute the cosine similarity between \( o_j^{(i)} \) and \( o_k^{(i)} \) using the learned embeddings, i.e.,

\[
S_{j,k}^{(i)} = S_c(E_{i,j}, E_{i,k}).
\]

For any layer \( i \leq L \) and operation pairs \( o_j^{(i)} \) and \( o_k^{(i)} \) (\( j < k \leq N \)), we merge them into the one with less FLOPs when they satisfy \( S_{j,k}^{(i)} > s_{\text{thrd}} \). After merging, the removed operations would never be sampled in training and search, thus reducing the training parameters in supernet. We set \( s_{\text{thrd}} = 0.8 \) in our experiments.

Our operation shrinkage method can significantly reduce the search space without additional evaluation steps. It can be naturally combined with the path-level shrinkage to perform a greedier search. The overall supernet training strategy is summarized in Algorithm 1.

5. Experiments

5.1. Experimental setup

**Search space.** As summarized in Table 1, for comparisons with baselines [23, 34], we first search on MobileNetV2-SE search space, which consists of Identity, MobileNetV2 block [21], and optional SE modules [13]. To validate our superiority on larger search spaces, we extend the search space with MixConv [29] block, namely, MobileNetV2-SE+MixConv. Moreover, we set up an extremely-large search space MobileNetV2-SE+MixConv+Shuffle, which adds 4 ShuffleNetV2 blocks [19] following SPOS [11]. In order to validate the effectiveness of our method on larger networks, we also introduce a ResNet-like search space, which consists of the blocks in ResNet [12], ResNeXt [31], and SENet [13]. Details can be found in supplementary materials.

| Method | ACC (%) | ACC on supernet (%) |
|-------|---------|---------------------|
| SPOS [11] | 76.8 | 76.6 | 75.5 | 56.5 | 48.2 | 33.4 |
| GreedyNAS [34] | 77.1 | 76.8 | 76.5 | 57.6 | 49.3 | 35.1 |
| GreedyNASv2 | 77.3 | 77.4 | 77.5 | 58.1 | 55.5 | 43.8 |

| Method | ACC (%) | ACC on supernet (%) |
|-------|---------|---------------------|
| Small | Medium | Large | Small | Medium | Large |
| SPOS [11] | 76.8 | 76.6 | 75.5 | 56.5 | 48.2 | 33.4 |
| GreedyNAS [34] | 77.1 | 76.8 | 76.5 | 57.6 | 49.3 | 35.1 |
| GreedyNASv2 | 77.3 | 77.4 | 77.5 | 58.1 | 55.5 | 43.8 |

**Table 1.** Summary of our search spaces. Details can be found in Supplementary Materials.
Table 3. Comparisons of searched architectures with state-of-the-art NAS methods and handcraft models. Training epochs and search number are hyper-parameters in the training of supernet. We measure the training cost of supernet using 8 NVIDIA V100 GPUs. *: trained with the same strategy as GreedyNASv2-L.

| Methods                  | Top-1 (%) | Top-5 (%) | FLOPs (M) | Params (M) | Training epochs | Training cost (GPU days) | Search number |
|--------------------------|-----------|-----------|-----------|------------|-----------------|--------------------------|---------------|
| MobileNetV2 [21]         | 72.0      | 91.0      | 300       | 3.4        | -               | -                        | -             |
| EfficientNet-B0 [28]     | 76.3      | 93.2      | 390       | 5.3        | -               | -                        | -             |
| SPOS [11]                | 74.7      | -         | 328       | 3.4        | 120             | 12                       | 1000          |
| MCT-NAS-B [23]           | 76.9      | 93.4      | 327       | 6.3        | 120             | 12                       | 1000          |
| K-shot-NAS-B [27]        | 77.2      | 93.3      | 332       | 6.2        | 120             | 12                       | 1000          |
| NSENet [4]               | 77.3      | -         | 333       | 7.6        | 100             | 166.7                    | 2100          |
| GreedyNAS-B [34]         | 76.8      | 93.0      | 324       | 5.2        | 46              | 7                        | 1000          |
| GreedyNASv2-S            | 77.5      | 93.5      | 324       | 5.7        | 65              | 7                        | 500           |

| Methods                  | Top-1 (%) | Top-5 (%) | FLOPs (M) | Params (M) | Training epochs | Training cost (GPU days) | Search number |
|--------------------------|-----------|-----------|-----------|------------|-----------------|--------------------------|---------------|
| RegNetX-4.0GF [20]       | 78.6      | -         | 4230      | 25.0       | -               | -                        | -             |
| ResNet-50* [12]          | 78.8      | 94.6      | 4089      | 25.6       | -               | -                        | -             |
| SE-ResNetX-50 [13]       | 78.9      | 94.5      | 4233      | 27.6       | -               | -                        | -             |
| SKNet-50 [18]            | 79.2      | -         | 4470      | 27.5       | -               | -                        | -             |
| SE-ResNet-50* [13]       | 80.5      | 94.8      | 4094      | 30.6       | -               | -                        | -             |
| GreedyNASv2-L            | 81.1      | 95.4      | 4098      | 26.9       | 57              | 9                        | 500           |

architectures using a RMSProp optimizer with a batch size 96 on each of 8 GPU, a step learning rate scheduler which warms up for 3 epochs then decays 0.97 every 2.4 epochs is adopted with initial value 0.048. While for ResNet-like model, we train it using SGD optimizer with weight decay $10^{-4}$ and batch size 1536, the initial learning rate is set to 0.6 and decays for 240 epochs with a cosine scheduler. We use a data augmentation pipeline of Autoaugment [5], random cropping, and clipping. We use a train and test image size of $224 \times 224$. Besides, an exponential moving average on weights is also adopted with a decay 0.9999.

5.2. Results on ImageNet

Comparisons with NAS methods. We first compare our GreedyNASv2 with the baseline methods SPOS [11] and GreedyNAS [34] on MB-SE, MB-SE+MixConv, and MB-SE+MixConv+Shuffle search spaces based on our implementations. We use a constraint of 330M FLOPs and report the evaluation accuracies of the searched architectures on retraining and supernet in Table 2. We can see that, on all sizes of search spaces, our GreedyNASv2 can obtain higher ACCs than the other two methods. Moreover, the performance of SPOS on medium and large spaces drops significantly, showing that it would be difficult for SPOS to train promising supernets on such huge spaces. While our GreedyNASv2 obtains similar performance, and even achieves the best performance on large space. We compare our obtained model GreedyNASv2-S on MB-SE+MixConv+Shuffle search space with state-of-the-art NAS methods in Table 3.

Search for larger networks. To evaluate our generalization, we conduct search on a ResNet-style search space Res-50-SE. As the results shown in Table 3, our GreedyNASv2 achieves significant improvement compared to the baseline ResNet, ResNeXt, and SENet models. Note that we train our GreedyNASv2-L with simple SGD optimizer and an additional Autoaugment [5] data pipeline. However, its performance still outperforms the state-of-the-art training strategies with more sophisticated optimization and strong data augmentation in TIMM [30], which achieves 80.4% ACC on ResNet-50.

5.3. Results on NAS-Bench-Macro

MCT-NAS [23] proposes a NAS benchmark named NAS-Bench-Macro for single path one-shot NAS methods, which consists of 6561 architectures and their isolated evaluation results on CIFAR-10 dataset. We leverage this benchmark to validate the effectiveness of GreedyNASv2.

Performance of path filter with ground-truth training data. To validate the pure performance of our PU learning method, we conduct experiments to train the path filter with the ground-truth labels in the benchmark. Concretely, we split the architectures to 10% of good paths and 90% of weak paths according to their evaluation accuracies, then samples 1%, 10%, 50%, and 100% data as a train set. We use the whole set to validate the classification performance of the learned path filter. We use precision and recall as evaluation metrics. We first train the path filter with our PU learning settings using only weak paths and randomly sampled unlabeled paths. For comparisons, we also adopt the PN learning (traditional supervised learning) by using both weak labels and good labels. As the results reported in
Table 4. Performance of our PU learning method compared with PN learning on NAS-Bench-Macro [23].

| Method | 10% data | 50% data | 100% data |
|--------|---------|---------|----------|
| Pre.   | Recall  | Pre.   | Recall  | Pre.   | Recall  |
| PU     | 97.21  | 98.30  | 98.81  | 98.30  | 98.81  |
| PN     | 79.52  | 88.45  | 95.55  | 98.15  | 98.25  |

Figure 5. Histogram of percentile rank of sampled paths on NAS-Bench-Macro search space. The average percentile rank of uniform sampling, GreedyNAS, and GreedyNASv2 are 90.6%, 18.1%, and 6.4%, respectively.

Table 4, the PU learning even achieves better performance than PN learning. This might be because paths are densely distributed in the space, an absolute threshold for partitioning “P” and “N” data might involve many uncertain paths, while the PU learning could handle this uncertainty well by treating unlabeled data more safely.

**Performance of path filter on supernet.** We also validate the performance of the path filter learned in supernet training. Unlike the previous experiment using the ground-truth labels, the labels in supernet training are generated by evaluating sampled architectures with a validation set; hence, they could have some noises. The learned path filter obtains 93.74% precision and 98.21% recall on the whole search space, comparing to the best performance of using ground-truth labels (98.81% precision and 98.62% recall), the small decrease in precision is acceptable since our method only needs to greedily focus on a proportion of potentially-good paths instead of locating all the good ones.

**Average percentile ranks of the sampled architectures during training.** We collect the percentile ranks of the sampled architectures during training. As shown in Figure 5, our method samples more paths with smaller percentile ranks compared to baselines, which means that our trained supernet is greedier towards those good paths.

**Correlation between validation and retraining accuracies.** Since GreedyNASv2 greedily filters out weak paths and focuses on the potential good paths, the correlation between validation accuracies on supernet and their retraining accuracies will be boosted in terms of those good paths identified by the path filter. We measure the rank correlations on those paths within top 10% percentile ranks on NAS-Bench-Macro, and the Kendall’s Tau of SPOS, GreedyNAS, and GreedyNASv2 are 23.9%, 41.5%, and 50.3%, respectively. This indicates our effectiveness since discriminating among good paths are fairly challenging.

**5.4. Ablation studies**

**Visualization of learned operation similarities.** As summarized in Figure 6 and Table 5, we visualize the learned operation similarities at the first searching layer of MB-SE supernet. It shows that the operations with the same kernel size are more likely to have similar embeddings. For ID operation, it has negative similarities to all the other operations since it performs poorly on down-sampling layers. For the whole supernet, there are a total of 35 (≈ 13%) out of 13 × 21 = 273 operations merged.

6. Conclusion

We propose GreedyNASv2, a NAS method with greedy path-level and operation-level shrinkage of search space. Unlike the previous works, our method achieves a greedier search with a greedy path filter, which is trained with highly-confident “weak” paths and unlabeled paths using positive-unlabeled (PU) learning. By dint of the learned embeddings in our path filter, we can further perform operation-level shrinkage by aggregating similar operations with similar embeddings, and the search can be more efficient and accurate. Extensive experiments show that our GreedyNASv2 achieves better performance compared to our baselines in various scales of search spaces.

**Acknowledgements**

This work was supported in part by the Australian Research Council under Project DP210101859 and the University of Sydney SOAR Prize.
References

[1] Hui Chen, Fangqing Liu, Yin Wang, Liyue Zhao, and Hao Wu. A variational approach for learning from positive and unlabeled data. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14844–14854. Curran Associates, Inc., 2020. 3, 5

[2] Yukang Chen, Tong Yang, Xiangyu Zhang, Gaofeng Meng, Xinyu Xiao, and Jian Sun. Detnas: Backbone search for object detection. *Advances in Neural Information Processing Systems*, 32:6642–6652, 2019. 1

[3] Xiangxiang Chu, Bo Zhang, and Ruijun Xu. Fairnas: Re-thinking evaluation fairness of weight sharing neural architecture search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12239–12248, 2021. 1

[4] Yuanzheng Ci, Chen Lin, Ming Sun, Boyu Chen, Hongwen Zhang, and Wanli Ouyang. Evolving search space for neural architecture search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6659–6669, 2021. 3, 7

[5] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019. 7

[6] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 6(2):182–197, 2002. 5, 6

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, Ieee, 2009. 6

[8] Marthinus Du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In *International conference on machine learning*, pages 1386–1394. PMLR, 2015. 3, 5

[9] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–220, 2008. 2, 3

[10] Jianyuan Guo, Kai Han, Yunhe Wang, Chao Zhang, Zhaohui Yang, Han Wu, Xinghao Chen, and Chang Xu. Hit-detector: Hierarchical trinity architecture search for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11405–11414, 2020. 1

[11] 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. In *European Conference on Computer Vision*, pages 544–560. Springer, 2020. 1, 3, 5, 6, 7

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 6, 7

[14] Yiming Hu, Yuding Liang, Zichao Guo, Ruosi Wan, Xiangyu Zhang, Yichen Wei, Qingyi Gu, and Jian Sun. Angle-based space shrinking for neural architecture search. In *European Conference on Computer Vision*, pages 119–134. Springer, 2020. 3

[15] Tao Huang, Shan You, Yibo Yang, Zhuozhuo Tu, Fei Wang, Chen Qian, and Changshui Zhang. Explicitly learning topology for differentiable neural architecture search. *arXiv preprint arXiv:2011.09300*, 2020. 1

[16] Ryuichi Kiryo, Gang Niu, Marthinus C du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1674–1684, 2017. 3

[17] Fabien Letouzey, François Denis, and Rémi Gilleron. Learning from positive and unlabeled examples. In *International Conference on Algorithmic Learning Theory*, pages 71–85. Springer, 2000. 2, 3

[18] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 510–519, 2019. 7

[19] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018. 6

[20] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10428–10436, 2020. 7

[21] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 6, 7

[22] Zan Shen, Jiang Qian, Bojin Zhuang, Shaojun Wang, and Jing Xiao. Bs-nas: Broader-and-shrinking one-shot nets with searchable numbers of channels. *arXiv preprint arXiv:2003.09821*, 2020. 3

[23] Xiu Su, Tao Huang, Yanxi Li, Shan You, Fei Wang, Chen Qian, Changshui Zhang, and Chang Xu. Prioritized architecture sampling with monte-carlo tree search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10968–10977, 2021. 1, 2, 3, 6, 7, 8

[24] Xiu Su, Shan You, Tao Huang, Fei Wang, Chen Qian, Changshui Zhang, and Chang Xu. Locally free weight sharing for network width search. In *International Conference on Learning Representations*, 2020. 1

[25] Xiu Su, Shan You, Fei Wang, Chen Qian, Changshui Zhang, and Chang Xu. Bcnet: Searching for network width with bilaterally coupled network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2175–2184, 2021. 1
[26] Xiu Su, Shan You, Jiyang Xie, Mingkai Zheng, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Vision transformer architecture search. arXiv preprint arXiv:2106.13700, 2021.

[27] Xiu Su, Shan You, Mingkai Zheng, Fei Wang, Chen Qian, Changshui Zhang, and Chang Xu. K-shot NAS: learnable weight-sharing for NAS with k-shot supernets. In ICML, volume 139 of Proceedings of Machine Learning Research, pages 9880–9890. PMLR, 2021.

[28] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning, pages 6105–6114. PMLR, 2019.

[29] Mingxing Tan and Quoc V. Le. Mixconv: Mixed depthwise convolutional kernels. In BMVC, page 74. BMVA Press, 2019.

[30] Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. arXiv preprint arXiv:2110.00476, 2021.

[31] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1492–1500, 2017.

[32] Yibo Yang, Hongyang Li, Shan You, Fei Wang, Chen Qian, and Zhouchen Lin. Ista-nas: Efficient and consistent neural architecture search by sparse coding. arXiv preprint arXiv:2010.06176, 2020.

[33] Yibo Yang, Shan You, Hongyang Li, Fei Wang, Chen Qian, and Zhouchen Lin. Towards improving the consistency, efficiency, and flexibility of differentiable neural architecture search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6667–6676, 2021.

[34] Shan You, Tao Huang, Mingmin Yang, Fei Wang, Chen Qian, and Changshui Zhang. Greedynas: Towards fast one-shot nas with greedy supernet. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1999–2008, 2020.

[35] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. arXiv preprint arXiv:1611.01578, 2016.