GreedyNAS: Towards Fast One-Shot NAS with Greedy Supernet

CVPR 2020

黄涛

见习研究员

商汤3D&AR-身份认证与视频感知组

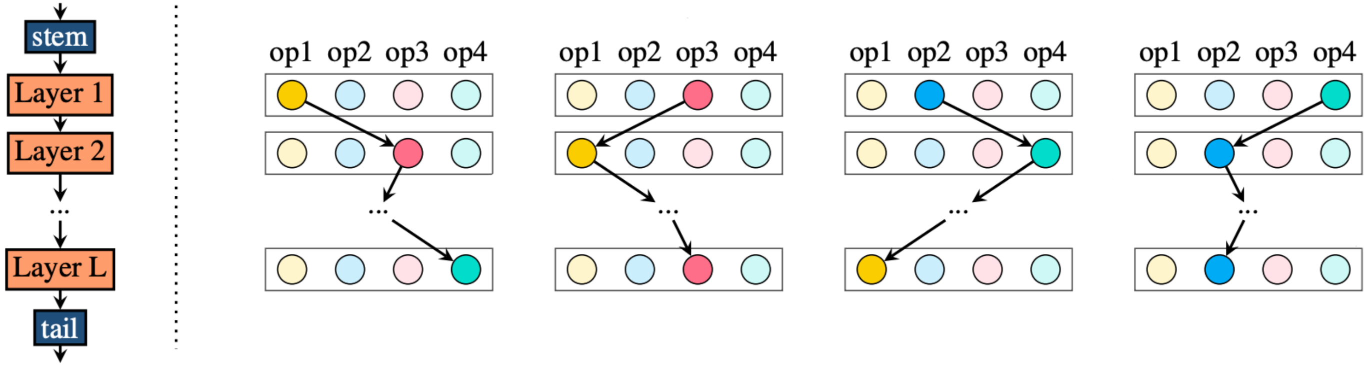
华中科技大学

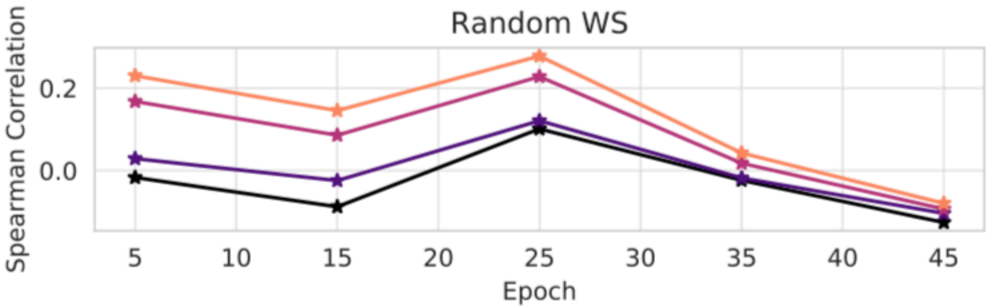
计算机科学与技术学院 大四

Motivation

Supernet: a fundamental performance estimator of different architectures (paths).

Target Assumption: the supernet should estimate the performance accurately for all paths, and thus all paths are treated equally and trained simultaneously.





Correlation between the one-shot validation error and the corresponding NAS-Bench-101 test error. (arXiv: 2001.10422)

Issues:

1. It is harsh to evaluate accurately on such a huge-scale search space (e.g. 7^{21}).
2. Training architectures with inferior quality would disturb the weights of those potentially-good paths.
3. Training on those weak paths involves unnecessary update of weights, and slows down the training efficiency.

Intuition: Path Filtering

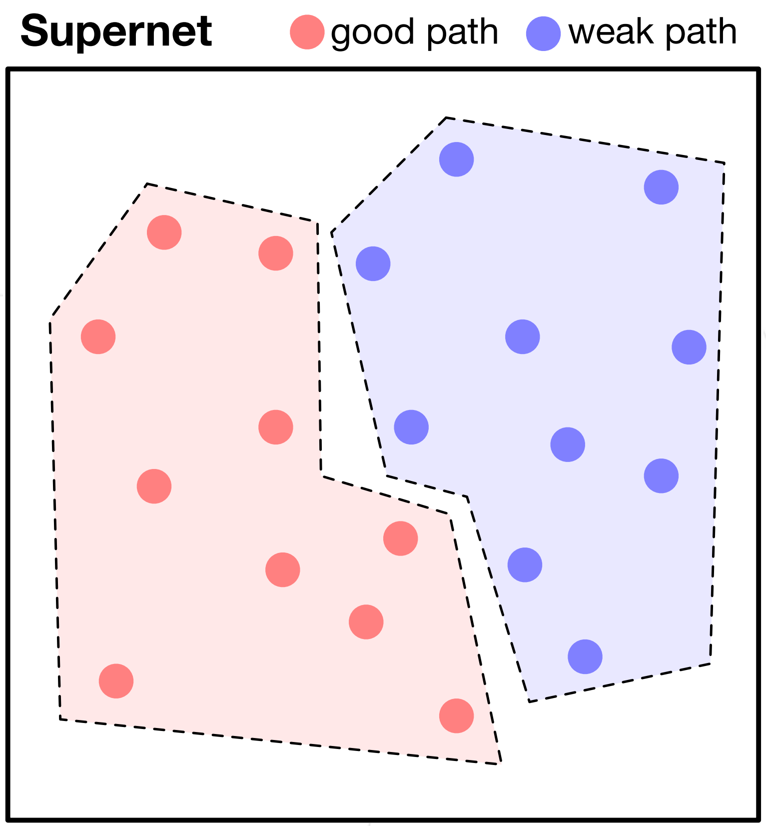
Consider a complete partition of search space $\mathcal{A}$ of two subsets $\mathcal{A}_{good}$ and $\mathcal{A}_{weak}$:

$$\mathcal{A} = \mathcal{A}_{good} \bigcup \mathcal{A}_{weak}, \ \mathcal{A}_{good} \bigcap \mathcal{A}_{weak} = \emptyset,$$

where for an Oracle supernet $\mathcal{N}_o$,

$$\mathrm{ACC}(\boldsymbol{a}, \mathcal{N}_o, \mathcal{D}_{val}) \geq \mathrm{ACC}(\boldsymbol{b}, \mathcal{N}_o, \mathcal{D}_{val})$$

holds for all $\boldsymbol{a} \in \mathcal{A}_{good}, \boldsymbol{b} \in \mathcal{A}_{weak}$ on validation dataset $\mathcal{D}_{val}$.



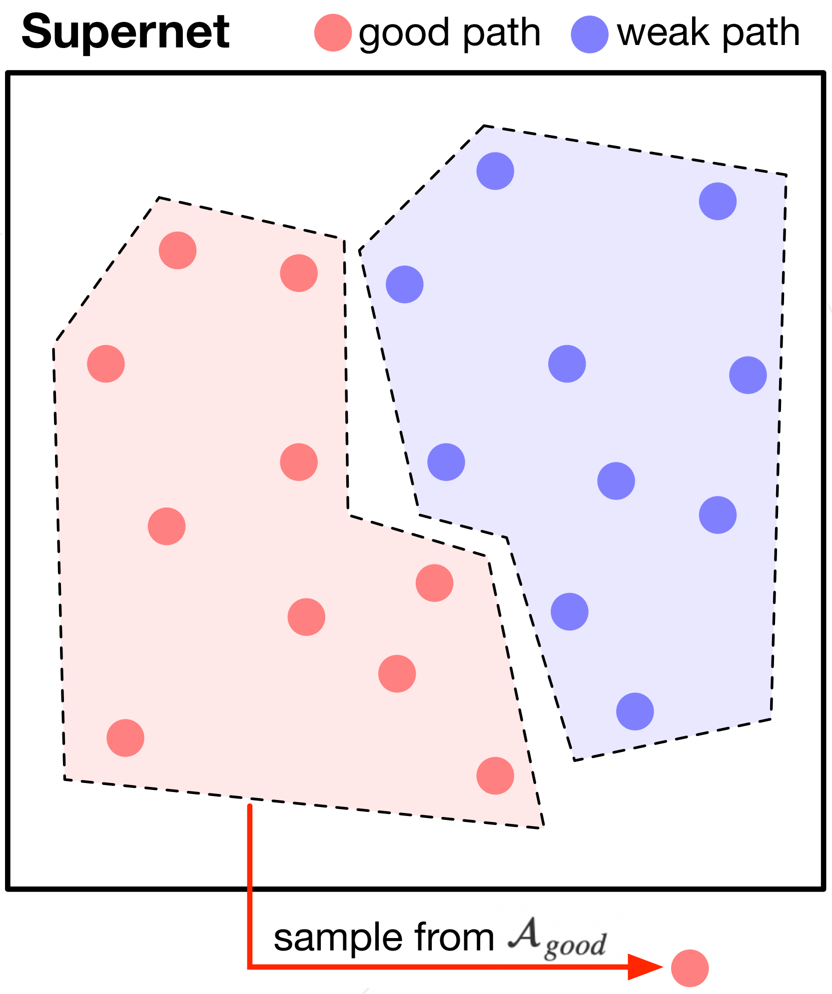
Intuition: Path Filtering

Idea: just sample from the potentially-good paths $\mathcal{A}_{good}$ instead of all paths $\mathcal{A}$:

$$p(\boldsymbol{a}; \mathcal{N}_o, \mathcal{D}_{val}) = \frac{1}{|\mathcal{A}_{good}|} \mathbb{I}(\boldsymbol{a} \in \mathcal{A}_{good}).$$

Problems:

- Q: Oracle supernet is unknown.
  A: greedily use current supernet as a proxy.
- Q: How can we accurately identify whether a path is from $\mathcal{A}_{good}$ or $\mathcal{A}_{weak}$ (computation cost of evaluating all paths in $\mathcal{A}$ is unacceptable)?
  A: multi-path sampling with rejection.

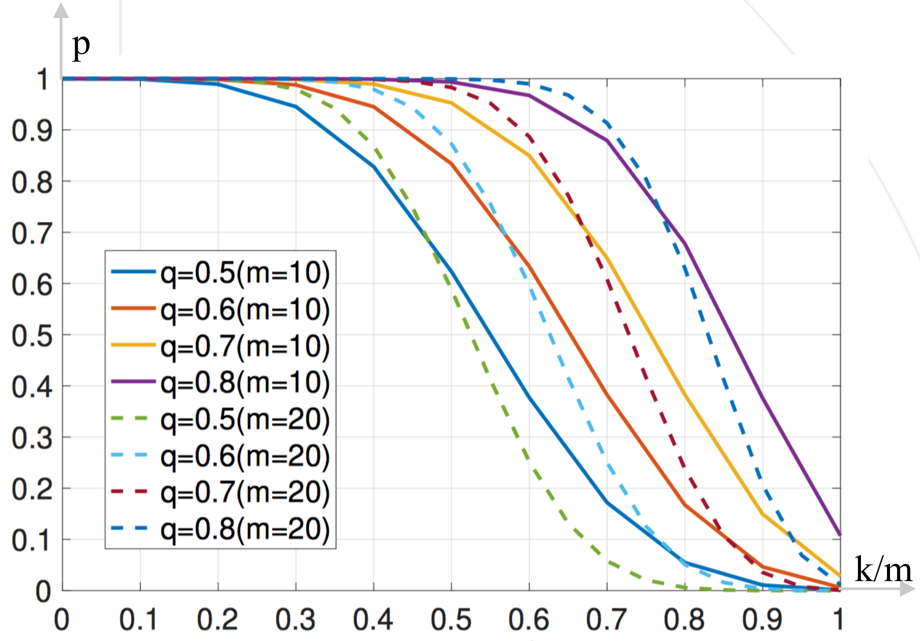


Solution: Multi-path Sampling with Rejection

Theorem: *If m paths are sampled uniformly i.i.d. from $\mathcal{A}$, then it holds that at least k (k ≤ m) paths are from $\mathcal{A}_{good}$ with probability*

$$\sum_{j=k}^{m} \mathbb{C}_m^j q^j (1-q)^{m-j},$$

where $q = |\mathcal{A}_{good}|/|\mathcal{A}|$.



With q = 0.6, it has 83.38% confidence to say at least 5 out of 10 paths are from $\mathcal{A}_{good}$.

Solution: just rank the sampled m paths using validation data $\mathcal{D}_{val}$, keep the Top-k paths and reject the remaining paths.

Exploration and Exploitation Training with Candidate Path Pool

We further introduce a candidate path pool to store the discovered good paths, and sample from it,

$$\boldsymbol{a} \sim (1-\epsilon)\cdot U(\mathcal{A}) + \epsilon \cdot U(\mathcal{P}),$$

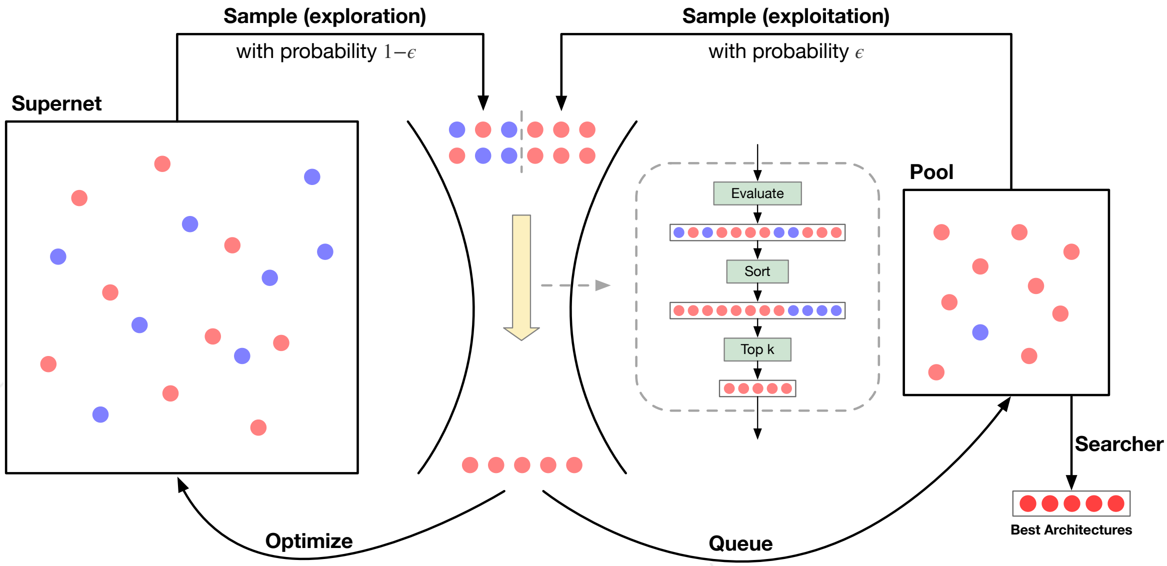
Advantages:

1. boosting the training efficiency
2. increasing the probability of sampling good paths $q = \epsilon + (1-\epsilon)|\mathcal{A}_{good}|/|\mathcal{A}|$, e.g. from 83.38% to 99.36% for 5/10 with ϵ = 0.5
3. stopping principle via candidate pool

   Stop by observing the steadiness of pool:

$$\pi := \frac{|\mathcal{P}_t \bigcap \mathcal{P}|}{|\mathcal{P}|} \leq \alpha$$

4. searching by initializing with candidate pool



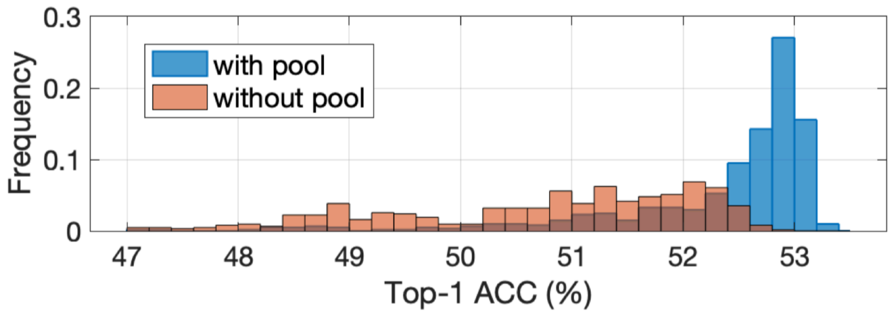


Figure 3: Histogram of accuracy of searched paths on supernet by evolutionary searching method (with or without candidate pool).

Problem: It is computationally expensive for evaluating paths using full validation dataset during training.

Solution: Using a small portion of validation dataset (1k images) for evaluation.

Table 3: Rank correlation coefficient of 1000 paths measured by the loss (ACC) of 1K validation images and ACC of 50K validation images w.r.t. different types of supernets.

Spearman rho			Kendall tau		
random	uniform(ACC)	greedy	random	uniform(ACC)	greedy
0.155	0.968(0.869)	**0.997**	0.113	0.851(0.699)	**0.961**

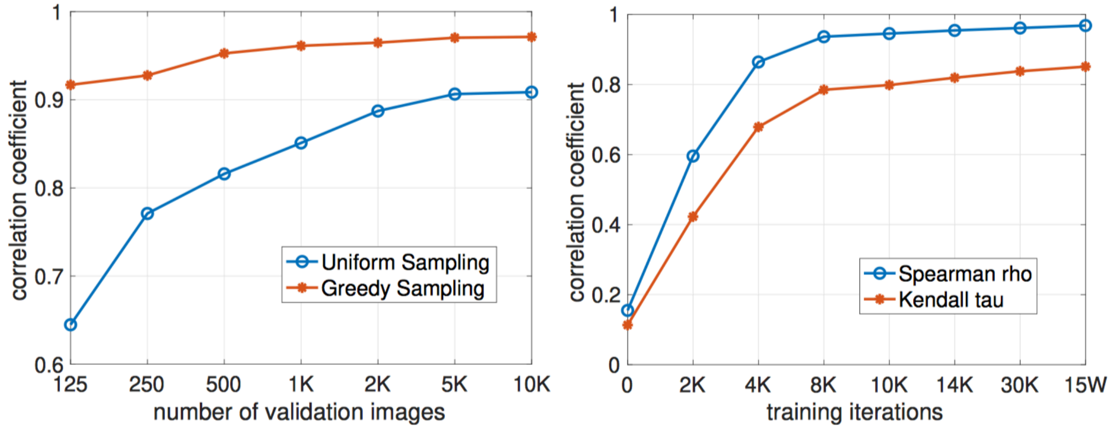


Figure 4: Rank correlation coefficient of 1000 paths measured by the loss of N validation images and ACC of the whole 50K validation images. Left: Comparison (Kendall tau) of supernet by uniform and greedy sampling w.r.t. different number N of evaluation images. Right: $N = 1K$ w.r.t. different training iterations of supernet by uniform sampling.

Experimental Results

- Searching Results with Same Search Space on ImageNet

Methods	performance			supernet training efficiency		
	Top-1 (%)	FLOPs	latency	#optimization	#evaluation	corrected #optimization
Proxyless-R (mobile)	74.60	320M	79 ms	-	-	-
Random Search	74.07	321M	69 ms	$1.23M \times 120$	-	147.6M
Uniform Sampling	74.50	326M	72 ms	$1.23M \times 120$	-	147.6M
FairNAS-C	74.69	321M	75 ms	$1.23M \times 150$	-	184.5M
Random Search-E	73.88	320M	91 ms	$1.23M \times 73$	-	89.8M
Uniform Sampling-E	74.17	320M	94 ms	$1.23M \times 73$	-	89.8M
GreedyNAS	**74.85**	320M	89 ms	$1.23M \times 46$	$2.40M \times 46$	89.7M
GreedyNAS	**74.93**	324M	78 ms	$1.23M \times 46$	$2.40M \times 46$	89.7M

- Comparison with state-of-the-art NAS methods on ImageNet

Methods	Top-1 (%)	FLOPs (M)	latency (ms)	Params (M)	training (Gdays)	search (Gdays)
SCARLET-C	75.6	280	67	6.0	10	12
MnasNet-A1	75.2	312	55	3.9	$288^{\ddagger}$	-
GreedyNAS-C	**76.2**	284	70	4.7	7	< 1
FairNAS-C	74.7	321	75	4.4	10	2
SCARLET-B	76.3	329	104	6.5	10	12
GreedyNAS-B	**76.8**	324	110	5.2	7	< 1
SCARLET-A	76.9	365	118	6.7	10	12
EfficientNet-B0	76.3	390	82	5.3	-	-
DARTS	73.3	574	-	4.7	$4^{\dagger}$	-
GreedyNAS-A	**77.1**	366	77	6.5	7	< 1