Differentiable Architecture Search Meets Network Pruning at Initialization: A More Reliable, Efficient, and Flexible Framework

- 作者：Miao Zhang，Wei Huang，Steven Su et.al.
- 机构：Aalborg University，UTS，Monash University，RMIT University
- 会议：arxiv
- 地址：https://arxiv.org/abs/2106.11542
- 代码：暂无

论文主要内容

摘要

DARTS最终得到的架构参数 α 的*magnitudes*很难代表operation的重要性。文章利用网络剪枝中的*synaptic saliency*指标来无需训练的获得opeartion的重要性。提出*training free differentiable architecture search (FreeDARTS)*。freeDarts能在不同搜索空间下超越baseline，并且对内存、计算友好（可以避免搜索的depth gap）。

贡献

1. 快，nasbench-201上0.6s；DARTS space上1.5s
2. zero-cost指标挪用到架构参数上，就算是zero-cost的NAS也得经过sampling-eval-sampling的循环。这篇文章计算完指标就能得到最终架构，一步到位。

研究内容

Motivation

- 背景
  - Darts训练的超网获得的架构参数 α 与opeartion的重要程度关系不大
  - Darts的内存消耗大（导致往往搜索proxy任务，比如在浅的网络搜索架构，用到深的网络结构depth gap）
  - 最近有工作说明无label的NAS也能获得相当的性能

- 抛出问题：能否在微分的范式下，不经任何训练得到operation的重要性？
  - FreeDARTS

方法

介绍

一些zero-cost指标：snip、grasp、synflow（label-agnostic）

主要源自purning at initiallzation，用来描述网络每个权重参数的重要程度

$$\mathcal{S}_{snip}(\theta) = \left| \frac{\partial \mathcal{L}}{\partial \theta} \odot \theta \right|.$$

$$\mathcal{S}_{grasp}(-\theta) = -(H\frac{\partial \mathcal{L}}{\partial \theta}) \odot \theta,$$

$$\mathcal{R}_{\mathrm{SF}} = \mathbf{1}^T \left(\prod_{l=1}^{L} \left| \theta^{[l]} \right| \right) \mathbf{1},$$

$$\mathcal{S}_{\mathrm{SF}}(\theta) = \frac{\partial \mathcal{R}_{\mathrm{SF}}}{\partial \theta} \odot \theta.$$

$\mathcal{R}$ 是label-agnostic的loss

方法

DARTS超网的架构参数为 α

$$\mathcal{F}_{snip}(\alpha) = \left| \frac{\partial \mathcal{L}}{\partial \alpha} \odot \alpha \right|.$$

$$\mathcal{F}_{grasp}(-\alpha) = -(H\frac{\partial \mathcal{L}}{\partial \alpha}) \odot \alpha.$$

$$\mathcal{R}_{\mathrm{SF}}(\alpha) = \mathbf{1}^T \left(\prod_{l=1}^{L} \left| W(\theta, \alpha)^{[l]} \right| \right) \mathbf{1},$$

$$\mathcal{F}_{\mathrm{SF}}(\alpha) = \frac{\partial \mathcal{R}_{\mathrm{SF}}}{\partial \alpha} \odot \alpha.$$

$$f(x) = \sum_{o=1}^{|\mathcal{O}|} \alpha_{L,o} W^{L,o} \ldots \sum_{o=1}^{|\mathcal{O}|} \alpha_{1,o} W^{1,o} x \ ,$$

$$\mathcal{F}_{\mathrm{SF}}(\alpha) = \left[\mathbf{1}^T \left(\prod_{l=i+1}^{L} \sum_{o=1}^{|\mathcal{O}|} \left| \alpha_{i,o} * W^{[l,o]} \right| \right) \right] |\alpha_{i,j}| \left[\left(\prod_{l=1}^{i-1} \sum_{o=1}^{|\mathcal{O}|} \left| \alpha_{i,o} * W^{[l,o]} \right| \right) \mathbf{1} \right],$$

用这些指标来作为DARTS中opeartion的重要性

Algorithm 1 FreeDARTS

1: **input**: Initialized supernet weights W and architecture parameters α; Set of edges $\mathcal{E}$ and set of candidate operations $\mathcal{O}$.
2: **for all** operations $o \in \mathcal{O}_e$ **do**
3: Calculate the operation saliency score for each operation $\alpha_{e,o}$ based on Eq. (7), (8), or (10).
4: **end for**
5: Prune the candidate operation by one-shot based on $\mathcal{F}_{\mathrm{snip}}(\alpha_{e,o})$, $\mathcal{F}_{\mathrm{grasp}}(\alpha_{e,o})$, or $\mathcal{F}_{\mathrm{SF}}(\alpha_{e,o})$;
6: **output:** Obtain a valid architecture α^*.

主要步骤：

1. 随机初始化架构参数 α
2. 用一个batch的数据经过网络计算**每条边上的每种**operaton(α)对应的score metric
   a. 从文章中看应是一次算所有边上一个op（signle path），但会不会产生边1，边2的组合问题？
   b. 如果是整个超网的op score metric一起算，但内存会占用大

3. 根据score metric的值大小，采用每条边取argmax的op策略获得最优架构

Zero-cost NAS用来score 架构

freeDarts 用来score op的重要性

实验结果

Results

Table 2. Zero-cost NAS and FreeDARTS with different saliency metrics on NAS-Bench-201.

Method	CIFAR-10		CIFAR-100		ImageNet-16-120	
	Valid(%)	Test(%)	Valid(%)	Test(%)	Valid(%)	Test(%)
Zero-cost with SNIP	89.08±0.96	91.82±1.56	68.00±2.16	68.08±2.01	37.29±5.91	37.12±5.88
Zero-cost with Grasp	88.10±0.65	90.92±0.81	66.26±1.40	66.35±1.17	34.66±4.83	33.88±4.43
Zero-cost with SynFlow	90.19±0.66	93.45±0.28	70.55±1.61	70.73±1.36	43.24±2.52	43.64±2.42
FreeDARTS with SNIP	89.57±0.57	92.96±0.52	69.77±0.76	69.90±0.80	42.66±1.51	43.79±1.54
FreeDARTS with Grasp	90.02±0.31	93.22±0.30	70.54±0.67	70.52±0.58	44.41±1.08	44.80±1.37
FreeDARTS with SynFlow	**90.39±0.30**	**93.52±0.11**	**70.96±0.51**	**70.78±0.08**	**44.68±1.67**	**45.28±1.73**

NAS-Bench-201:

Table 1. Comparison results with NAS baselines on NAS-Bench-201.

Method	CIFAR-10		CIFAR-100		ImageNet-16-120		Search Cost
	Valid(%)	Test(%)	Valid(%)	Test(%)	Valid(%)	Test(%)	(GPU sec.)
SETN [10]	84.04±0.28	87.64±0.00	58.86±0.06	59.05±0.24	33.06±0.02	32.52±0.21	31010
GDAS [11]	89.88±0.33	93.40±0.49	70.95±0.78	70.33±0.87	41.28±0.46	41.47±0.21	28925.91
DARTS (1st) [25]	39.77±0.00	54.30±0.00	15.03±0.00	15.61±0.00	16.43±0.00	16.32±0.00	10889.87
DARTS (2nd) [25]	39.77±0.00	54.30±0.00	15.03±0.00	15.61±0.00	16.43±0.00	16.32±0.00	29901.67
PC-DARTS [40]	89.96±0.15	93.41±0.30	67.12±0.39	67.48±0.89	40.83±0.08	41.31±0.22	10023.
SNAS [39]	90.10±1.04	92.77±0.83	69.69±2.39	69.34±1.98	42.84±1.79	43.16±2.64	32345
Random baseline	83.20±13.28	86.61±13.46	60.70±12.55	60.83±12.58	33.34±9.39	33.13±9.66	-
NASWOT [27]	89.16±1.13	92.45±1.12	68.53±2.01	68.66±2.02	41.13±3.94	41.35±4.08	30.01
Zero-Cost NAS [1]	90.19±0.66	93.45±0.28	70.55±1.61	70.73±1.36	43.24±2.52	43.64±2.42	115.2
Zero-Cost-PT NAS [38]	-	93.75±0.00	-	71.11±0.00	-	41.43±0.00	647.5
TE-NAS* [5]	89.99±0.40	93.28±0.25	69.19±0.63	69.62±0.71	43.72±2.06	44.29±1.97	1558
FreeDARTS	90.39±0.30	93.52±0.11	70.96±0.51	70.70±0.08	44.68±1.67	45.28±1.73	**0.6**
optimal	91.61	94.37	74.49	73.51	46.77	47.31	-

"*" indicates the results reproduced with the same *seeds*. We use the Synflow metric in this experiment. Our best single run achieves **93.66%**, **70.78%**, and **46.53%** test accuracy on three datasets, respectively.

中间部分是trianing-free NAS

NAS-Bench-1shot1

Table 3. Statistic search results (test error) on NAS-Bench-1shot1.

Method	CIFAR-10 Average (%) Valid	CIFAR-10 Average (%) Test	CIFAR-10 Best (%) Valid	CIFAR-10 Best (%) Test	Search Cost
GDAS	6.8±0.1	6.1±0.2	6.7	5.9	11425s
PC-DARTS	6.7±0.1	6.2±0.2	6.6	5.9	14760s
DARTS (1st)	6.8±0.05	6.1±0.2	6.6	5.9	8280s
DARTS (2nd)	6.8±0.05	6.2±0.05	6.6	6.2	19800s
Random	24.4±32.8	24.1±33.2	7.8	7.5	N/A
FreeDARTS	7.8±2.4	7.3±2.4	**6.0**	**5.3**	**1.1s**

DARTS space

Table 4. Comparison results with state-of-the-art NAS approaches on DARTS search space.

Method	Test Error (%) CIFAR-10	CIFAR-100	ImageNet	Param (M)	+× (M)	Train Free	Label Agnostic	Search Cost (GPU day or sec.)
PARSEC [4]	2.86±0.06	-	26.3	3.6	509	×	×	0.6d
SNAS [39]	2.85±0.02	20.09	27.3 / 9.2	2.8	474	×	×	1.5d
BayesNAS [49]	2.81±0.04	-	26.5 / 8.9	3.4	-	×	×	0.2d
MdeNAS [48]	2.55	17.61	25.5 / 7.9	3.6	506	×	×	0.16d
GDAS [11]	2.93	18.38	26.0 / 8.5	3.4	545	×	×	0.2d
PDARTS [8]	2.50	16.63	24.4 / 7.4	3.4	557	×	×	0.3d
PC-DARTS [40]	2.57±0.07	17.11	25.1 / 7.8	3.6	586	×	×	0.3d
DrNAS [7]	2.54±0.03	16.30	24.2 / 7.3	4.0	644	×	×	0.4d
DARTS (1st) [25]	2.94	17.76	-	2.9	513	×	×	1.5d
DARTS (2nd) [25]	2.76±0.09	17.54	26.9 / 8.7	3.4	574	×	×	4d
TE-NAS [5]	2.63	17.83	26.2 / 8.3	3.8	610	✓	×	0.17d
Zero-Cost-PT [38]	2.68±0.17	17.53	24.4 / 7.5	4.7	817	✓	✓	0.018d
FreeDARTS	2.78±0.06	18.03	26.1 / 8.2	3.6	634	✓	✓	**1.5s**
FreeDARTS†	**2.50±0.05**	17.08	25.4 / 7.8	3.6	577	✓	✓	**1.5s**
FreeDARTS‡	2.67±0.04	**16.35**	**24.4 / 7.3**	4.1	655	✓	✓	**1.5s**

"Param" is the model size on CIFAR-10, while "+×" is calculated on ImageNet dataset. "d" is the GPU days and "s" is the GPU seconds. We only consider the Synflow based metric, which is label agnostic, for our FreeDARTS in this search space.

- FreeDARTS † 堆20个cell消除depth gap
- FreeDARTS ‡ 在Imagenet上搜索

Table 5. Search results on with different settings.

Method	CIFAR-10 Test Error (%) Best	Mean	Param (M)
GDAS	2.93	3.22±0.31	2.83±0.07
DARTS (1st)	2.94	3.22±0.45	2.02±0.41
DARTS (2nd)	2.62	3.02±0.26	2.83±0.07
FreeDARTS	2.75	2.92±0.18	3.82±0.26
FreeDARTS†	2.45	2.78±0.28	3.49±0.13
FreeDARTS‡	2.63	2.91±0.31	3.89±0.23

We report the best and mean test error after several searches "Param" is the average model size and after several searches "Mean" is the average test error of searched architectures.

DARTS更偏好无参数的op（skip-connect，pooling）

NAS-Bench-301

Table 6. Statistic search results (test error) on NAS-Bench-301.

Method	Average	Best	Ground-True
GDAS [11]	6.52±0.62 (%)	5.38%	3.07±0.16 (%)
PC-DARTS [40]	6.42±0.43 (%)	5.46%	2.57±0.07 (%)
DARTS (2nd) [25]	6.74±0.58 (%)	5.87%	2.76±0.09 (%)
Random	7.11±0.58 (%)	6.21%	3.29±0.15 (%)
FreeDARTS(SNIP)	6.60±0.47 (%)	5.71%	2.69±0.08 (%)
FreeDARTS(GraSP)	6.72±0.48 (%)	5.74%	2.78±0.09 (%)
FreeDARTS(Synflow)	6.65±0.52 (%)	5.50%	2.50±0.05 (%)

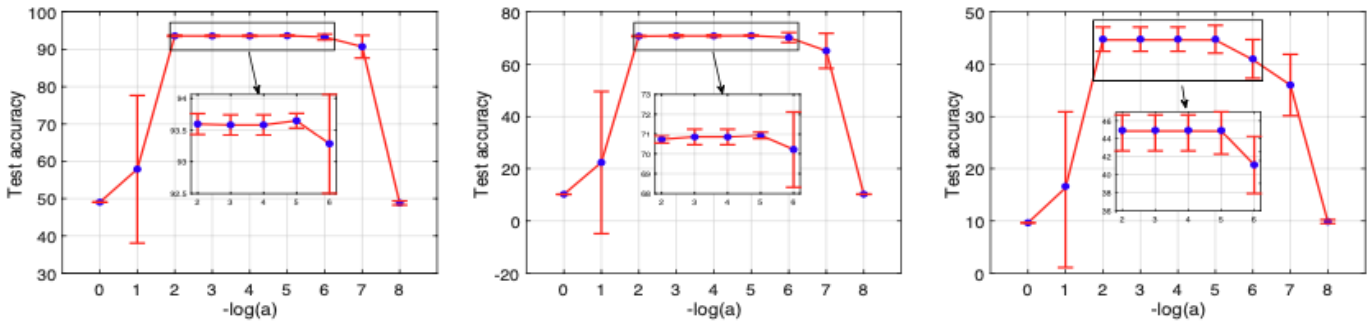
Table 7. Search results (test error) with the sample size on NAS-Bench-301.

Method	Average (10)	Best (10)	Average (100)	Best (100)	Average (1000)	Best (1000)
Random	7.11±0.58 (%)	6.21%	6.85±0.58 (%)	5.71%	6.89±0.55 (%)	5.61%
FreeDARTS(SNIP)	6.70±0.47 (%)	5.71%	6.77±0.51 (%)	5.65%	6.67±0.50 (%)	5.50%
FreeDARTS(GraSP)	6.72±0.48 (%)	5.74%	6.71±0.48 (%)	5.64%	6.65±0.49 (%)	5.55%
FreeDARTS(Synflow)	6.65⊥0.52 (%)	5.50%	6.66⊥0.53 (%)	5.49%	6.62⊥0.50 (%)	5.34%

超参研究

- α 初始化策略：a*randn

- a是 $\frac{\partial \mathcal{L}}{\partial \alpha}$ 和 α 之间的trade-off
- a在1e-5~1e-2范围效果好



(a) CIFAR-10 (b) CIFAR-100 (c) ImageNet

Figure 1. Hyperparameter analysis of FreeDARTS on the NAS-Bench-201 benchmark dataset.