Single Path One-Shot Neural Architecture Search with Uniform Sampling

https://arxiv.org/abs/1904.00420

https://github.com/megvii-model/SinglePathOneShot

NAS

Definition: Neural Architecture Search (NAS).

Find the architecture that leads to the best validation accuracy (or other metrics such as efficiency.)

- **Example:** ResNet has better accuracy than VGG.
- **Example:** MobileNet is more efficient than ResNet, although MobileNet has lower accuracy.

NAS

Search Space

Hyper-parameter Types	Candidates
# of filters	{24, 36, 48, 64}
size of filters	{3×3, 5×5, 7×7}
stride	{1, 2}

Search space: The set containing all the possible architectures.

- We want to build a CNN with 20 Conv layers.
- Search space:

$$\{24, 36, 48, 64\}^{20} \times \{3\times3, 5\times5, 7\times7\}^{20} \times \{1, 2\}^{20}.$$

- Size of search space (i.e., number of possible architectures):

$$(4\times3\times2)^{20}$$

NAS

Challenge 1: Each trial is expensive.

- Training a CNN from scratch takes hours or days, if a single GPU is used.

Challenge 2: The search space is too big.

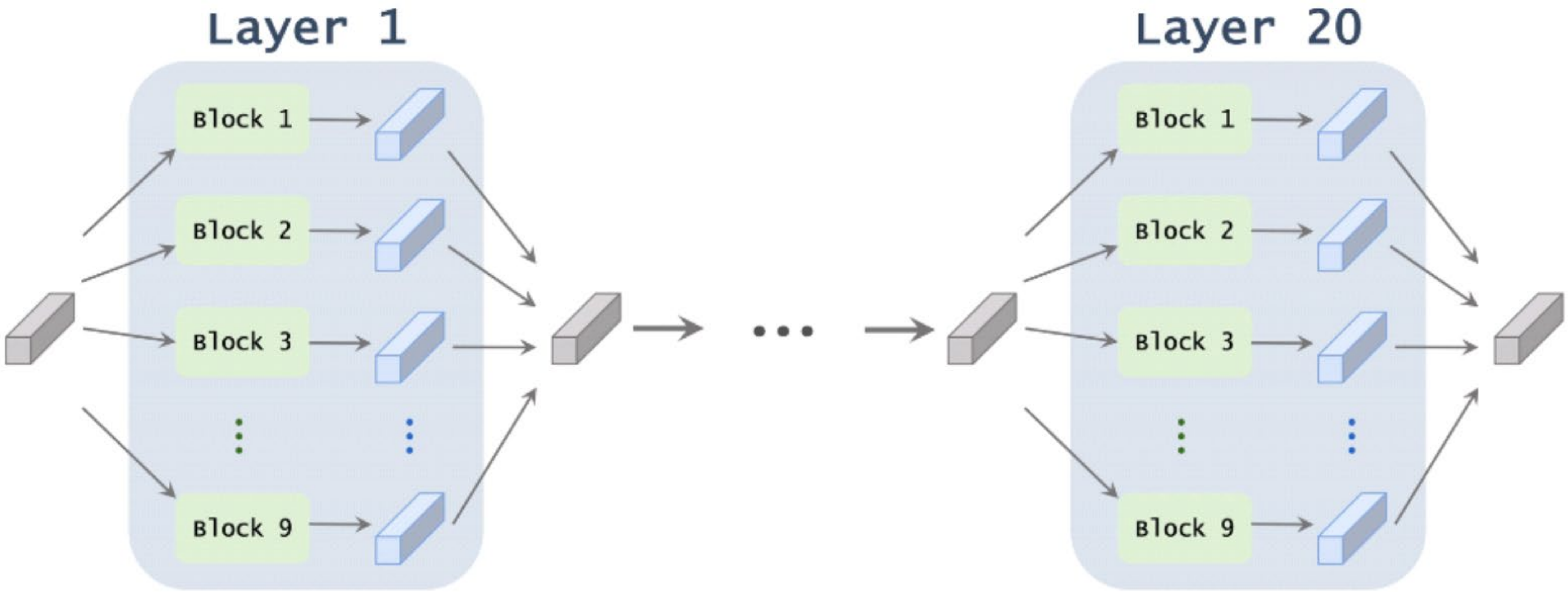
- Number of possible architectures:

$$(4\times3\times2)^{20} = 4\times10^{27}.$$

1.每个结构的尝试成本很高；

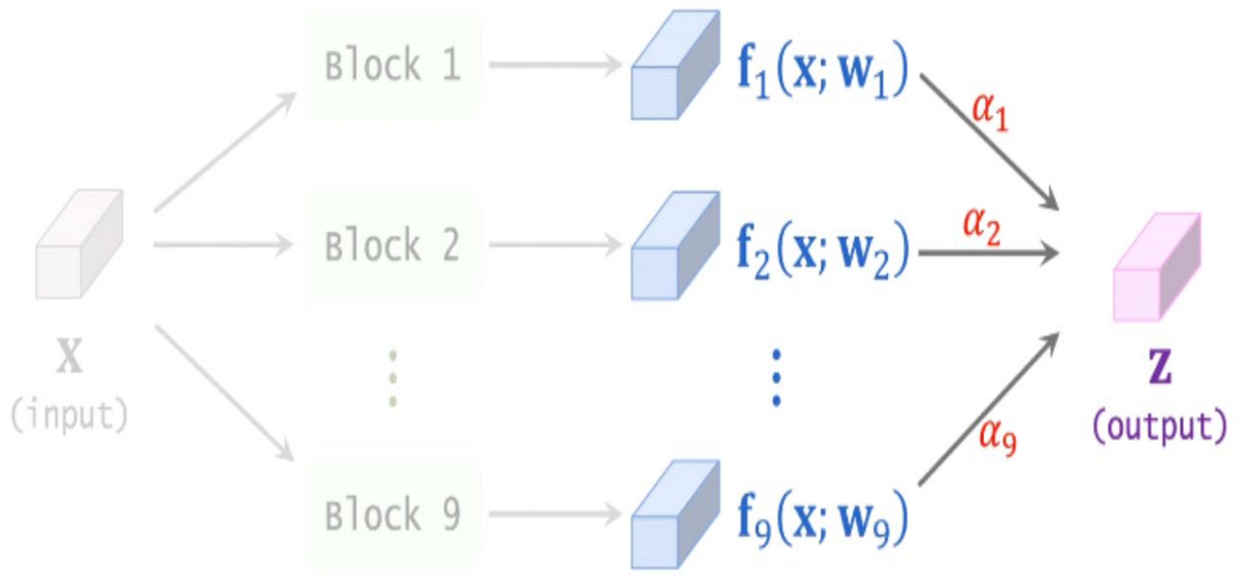
2.搜索空间很大。

FBNet



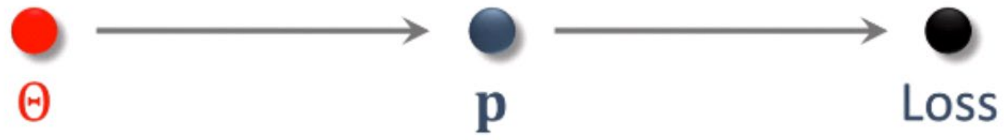
SuperNet

FBNet

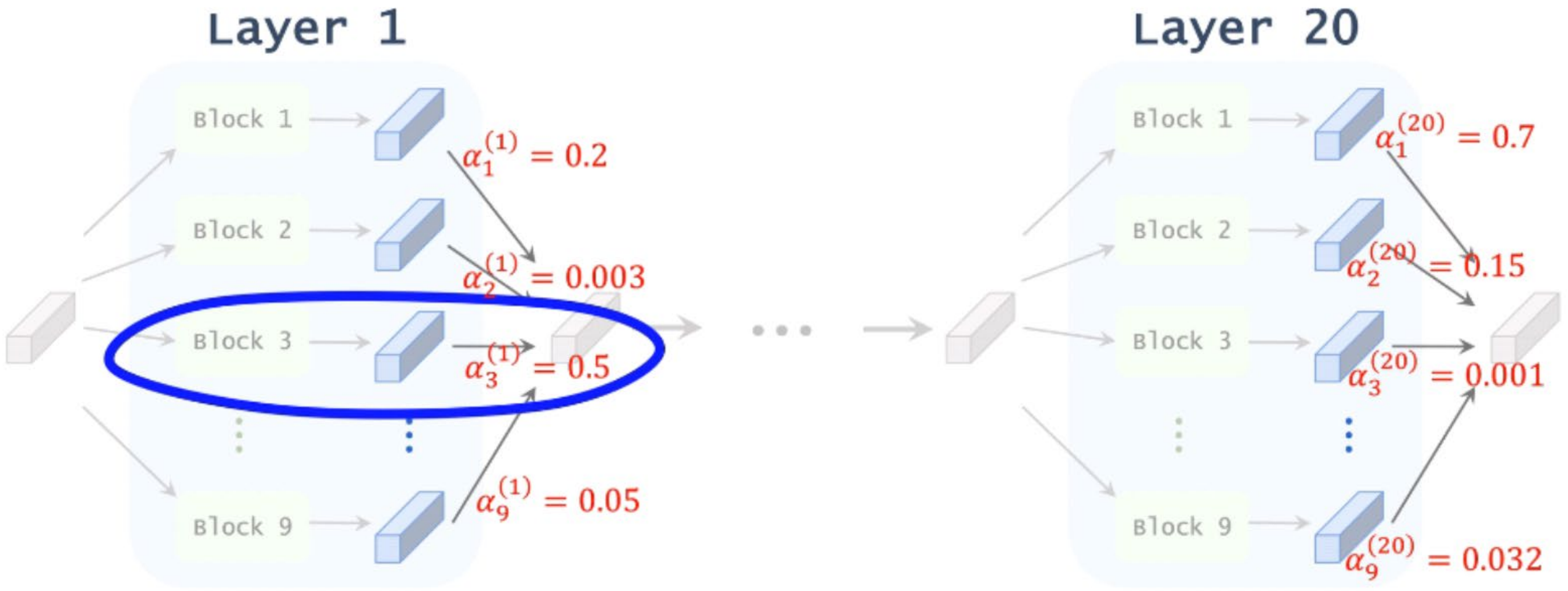


- $\mathbf{x}_1, \cdots, \mathbf{x}_n$: training images.
- $\mathbf{y}_1, \cdots, \mathbf{y}_n$: targets (aka labels).
- $\mathbf{p}(\mathbf{x}_i; \mathcal{W}, \Theta)$: a prediction made by the 20-layer super-net.
- Learn $\mathcal{W}$ and Θ from the training set by minimizing the cross-entropy loss:

$$\min_{\mathcal{W},\Theta} \ \frac{1}{n}\sum_{i=1}^{n} \mathrm{Loss}\big(\mathbf{y}_i,\ \mathbf{p}(\mathbf{x}_i; \mathcal{W}, \Theta)\big).$$



FBNet



每一层只选择α最大的block，对20层重复进行这个操作。只保留最重要的模块，丢弃其他模块。
如果将supernet看作一个图，选择的神经网络结构就是其中的一条路径，从输入通往输出。

SPOS

联合优化问题

- 超网中的权重是紧密耦合的，尚不清楚子网的权重继承为何是有效的。
- 使用同时优化的方式也给网络架构参数和超网参数引入了耦合。

SPOS

提出了均匀采样的single path one-shot方法，可以克服现有one-shot方法的缺点。其简单的形式允许更大的搜索空间，包括block搜索、通道搜索、比特宽度搜索等。采用进化算法来进行搜索，可以满足低延迟等约束。

SPOS

第一阶段 优化supernet

$$W_{\mathcal{A}} = \underset{W}{\operatorname{argmin}} \, \mathcal{L}_{\text{train}} \left(\mathcal{N}(\mathcal{A}, W) \right).$$

$\mathcal{A}$代表网络搜索空间，W代表超网权重，$\mathcal{N}(\mathcal{A}, W)$代表超网中编码的搜索空间。

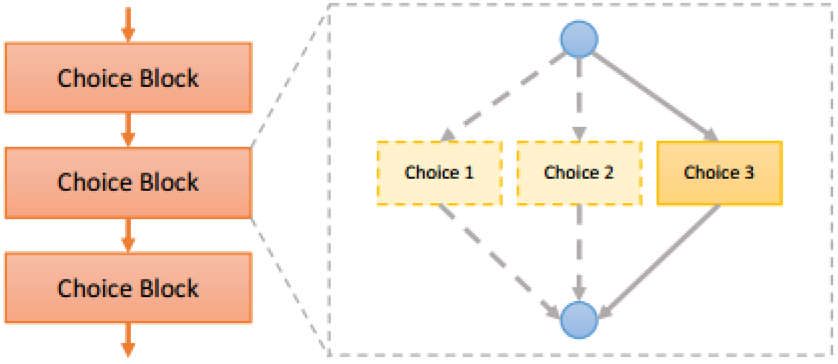
第二阶段 网络架构search

$$a^* = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \, \text{ACC}_{\text{val}} \left(\mathcal{N} \left(a, W_{\mathcal{A}}(a) \right) \right)$$

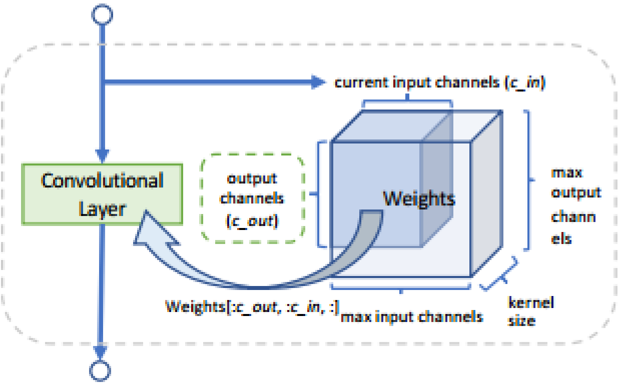
a代表被采样的子网架构，它会继承超网的权重$W_{\mathcal{A}}(a)$, 然后在这个过程中挑选验证集上准确率最高的子网结构。

SPOS

Supernet实现细节



(a)



(b)

Single path supernet

Channel number search

SPOS

Search实现细节

Algorithm 1: Evolutionary Architecture Search

Input: *supernet weights* $\mathrm{W}_{\mathcal{A}}$, *population size* P, *architecture constraints* $\mathcal{C}$, *max iteration* $\mathcal{T}$, *validation dataset* D_{val}

Output: *the architecture with highest validation accuracy under architecture constraints*

$\mathrm{P}_0 := Initialize\_population(P, \mathcal{C})$; $\mathrm{Topk} := \emptyset$;

$n := P/2$; Crossover number

$m := P/2$; Mutation number

$prob := 0.1$; Mutation probability

for $i = 1 : \mathcal{T}$ **do**

- $\mathrm{ACC}_{i-1} := Inference(\mathrm{W}_{\mathcal{A}}, \mathrm{D}_{val}, \mathrm{P}_{i-1})$;
- $\mathrm{Topk} := Update\_Topk(\mathrm{Topk}, \mathrm{P}_{i-1}, \mathrm{ACC}_{i-1})$;
- $\mathrm{P}_{crossover} := Crossover(\mathrm{Topk}, n, \mathcal{C})$;
- $\mathrm{P}_{mutation} := Mutation(\mathrm{Topk}, m, prob, \mathcal{C})$;
- $\mathrm{P}_i := \mathrm{P}_{crossover} \cup \mathrm{P}_{mutation}$;

end

Return the architecture with highest accuracy in Topk;

实验

Table 3. Results of building block search. *SPS* – single path supernet

model	FLOPs	top-1 acc(%)
all choice_3	324M	73.4
all choice_5	321M	73.5
all choice_7	327M	73.6
all choice_x	326M	73.5
random select (5 times)	~320M	~73.7
SPS + random search	323M	73.8
ours (fully-equipped)	319M	**74.3**

- 随机搜索效率低
- 进化算法搜索效率高，有限时间内搜索到最高精度

实验

Table 4. Results of channel search. * Performances are reported in the form "x (y)", where "x" means the accuracy retrained by us and "y" means accuracy reported by the original paper

Model	FLOPs/Params	Top-1 acc(%)
all choice_3	324M/3.1M	73.4
rand sel. channels (5 times)	∼ 323M/3.2M	∼ 73.1
choice_3 + channel search	329M/3.4M	**73.9**
rand sel. blocks + channels	∼ 325M/3.2M	∼ 73.4
block search	319M/3.3M	74.3
block search + channel search	328M/3.4M	**74.7**
MobileNet V1 (0.75x) [8]	325M/2.6M	68.4
MobileNet V2 (1.0x) [18]	300M/3.4M	72.0
ShuffleNet V2 (1.5x) [14]	299M/3.5M	72.6
NASNET-A [36]	564M/5.3M	74.0
PNASNET [11]	588M/5.1M	74.2
MnasNet [21]	317M/4.2M	74.0
DARTS [12]	595M/4.7M	73.1
Proxyless-R (mobile)* [4]	320M/4.0M	74.2 (74.6)
FBNet-B* [23]	295M/4.5M	74.1 (74.1)

同时搜索block和channel的精度更高，超过了其他同类型方法。