Neural Architecture Search with Loss Flatness-aware Measure

NAVER

ICML International Conference On Machine Learning

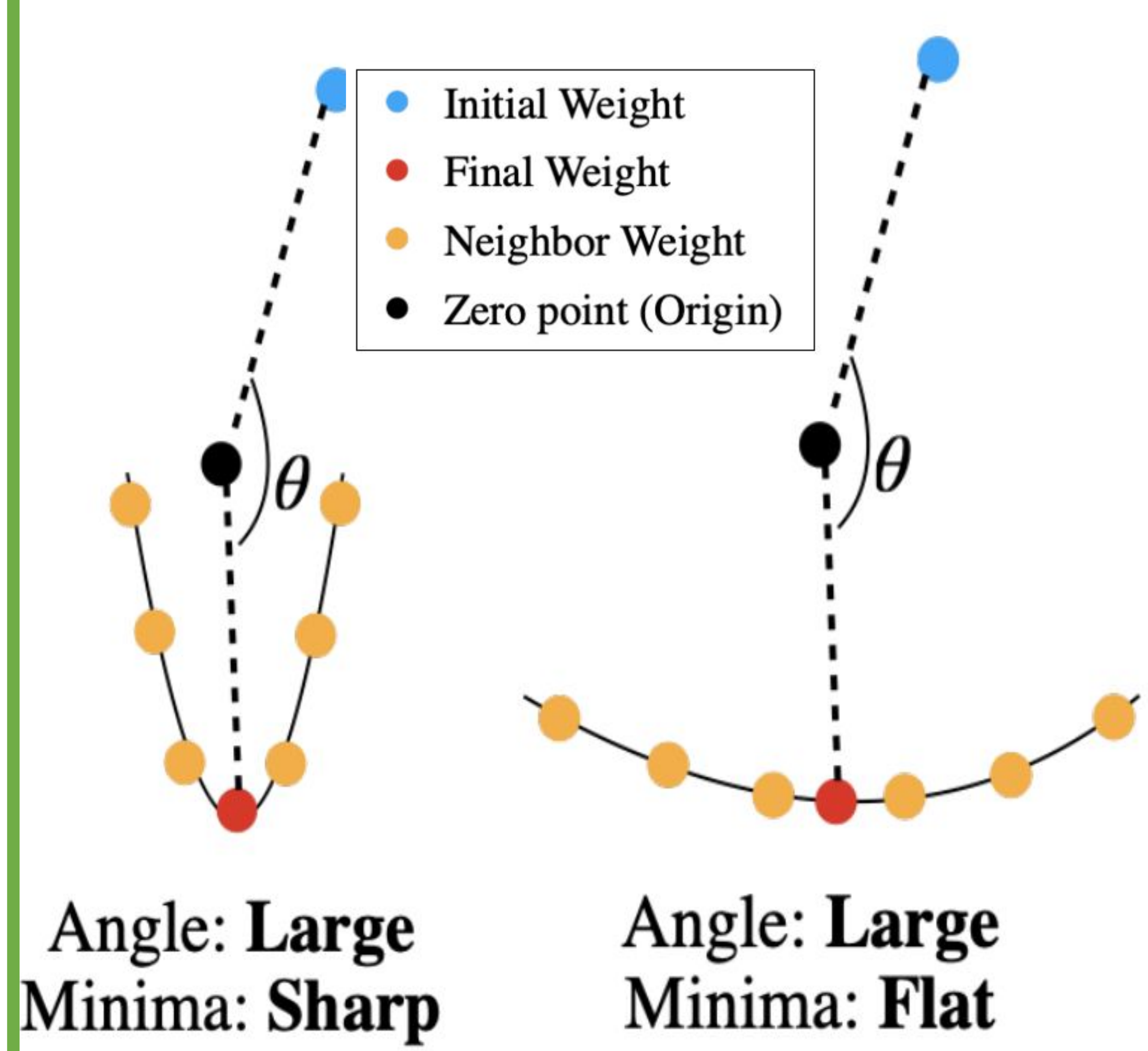
DyNN

Joonhyun Jeong[1] Joonsang Yu[1 2] Dongyoon Han[2] YoungJoon Yoo[1 2]
[1]Image Vision, NAVER CLOVA [2]NAVER AI Lab
{joonhyun.jeong, joonsang.yu, dongyoon.han, youngjoon.yoo}@navercorp.com

Main Contributions

- **A new search proxy**, **FBS**: Flatness of local minima as a measure for searching generalizable architectures.
- **Baseline Improvements:** FBS further boosts generalizability of conventional search metrics.
- **Task-generalizability:** FBS searches generalizable architectures on downstream tasks such as object detection.

Motivation



(a) ABS (b) FBS (Ours)

- Investigate an **open question**:

  Flatness -> Generalizability (?)

  ➡ Can quantifying flatness acquire generalizable architectures?

- **Insufficient Generalization**:

Comparison	Kendall's Tau		
	CIFAR-10	CIFAR-100	ImageNet16-120
Angle & Flatness	0.4302	0.4724	0.4097
Accuracy & Flatness	0.7923	0.7568	0.7620

- Conventional search metrics have a **large headroom** for better generalization in terms of flatness.

- Especially, Angle-Based Searching (ABS) shows significantly low correlation with flatness.

Method

- **Objective**: In an entire architecture search space A, find the **maximal flat** architecture $\boldsymbol{a^*}$.

$$a^* = \underset{a \in A}{\operatorname{argmax}} F_{val}(W_A^*(a)).$$

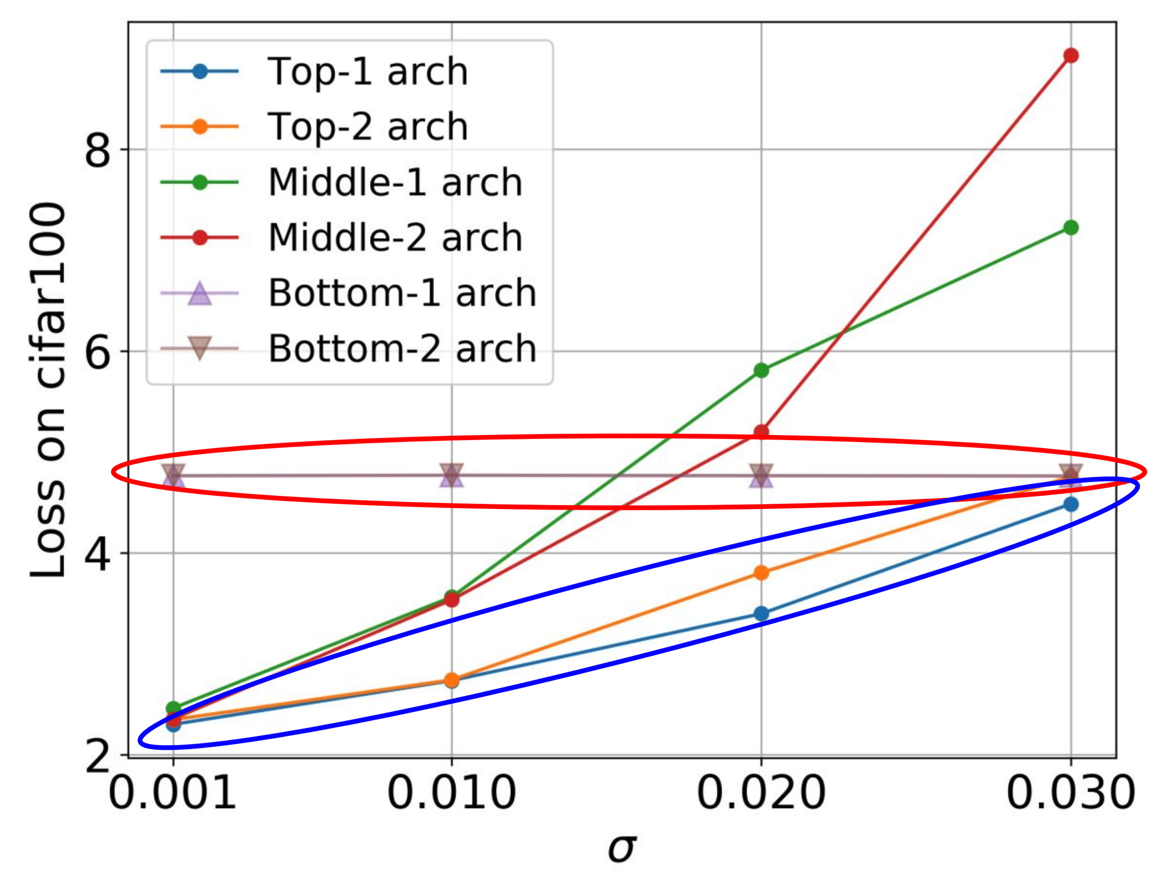
➡ How to measure **flatness of local minima**, F_{val} ?

$$F_{val}(\theta) = \Big(\sum_{i=1}^{t-1} \frac{L(\theta + N(\sigma_{i+1})) - L(\theta + N(\sigma_i))}{\sigma_{i+1} - \sigma_i}\Big)^{-1},$$

1. Get a loss surface by **perturbing** the converged weight θ with Gaussian Noise $N(\sigma)$ for $t-1$ times.

2. Estimate flatness of the loss surface.

- Consideration on **loss depth**:



- Naive flatness-based searching selects **sub-optimal** architectures (red circle).

- For achieving top-performances (blue circle), depth of loss should be considered together as:

$$F_{val}(\theta) = \Big(\sum_{i=1}^{t-1} \Big|\frac{L(\theta + N(\sigma_{i+1})) - L(\theta + N(\sigma_i))}{\sigma_{i+1} - \sigma_i}\Big| + \boxed{\alpha \Big|\frac{L(\theta + N(\sigma_1))}{\sigma_1}\Big|}\Big)^{-1}$$

where σ_1 is the smallest perturbation degree.

Experimental Results

- **Results on ImageNet**
  - Transfer from CIFAR-100

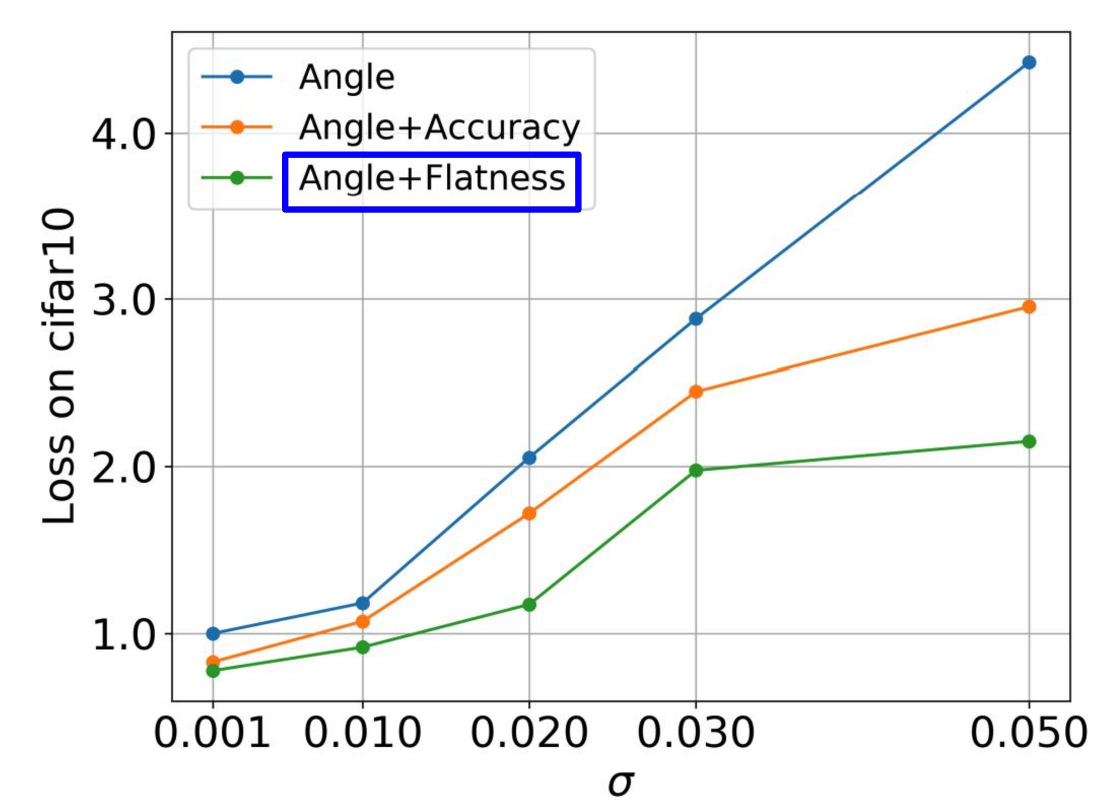
Method	Params (M)	FLOPs (G)	Top-1 Acc (%)	Top-5 Acc (%)
PC-DARTS	5.3	0.59	74.75	92.16
RLNAS	5.4	0.61	75.00	92.31
DropNAS†	5.1	0.57	75.07	92.33
P-DARTS	5.1	0.58	75.30	92.50
SPOS	5.4	0.60	75.37	92.23
GeNAS (Ours)	**5.2**	**0.58**	**76.05**	**92.64**

our **GeNAS** achieves **state-of-the-art Performance** with similar # params and FLOPS!

- **Collaborative effect** of FBS on other conventional search metric (Angle)

Flatness (%)	Params (M)	FLOPs (G)	Top-1 Acc (%)	Top-5 Acc (%)
0	5.43	0.61	75.00	92.31
20	5.45 (+0.02)	0.60 (-0.01)	75.22 (+0.22)	92.39 (+0.08)
43	5.57 (+0.14)	0.61 (+0.00)	75.58 (+0.58)	92.44 (+0.13)
76	5.41 (-0.02)	0.60 (-0.01)	75.63 (+0.63)	92.54 (+0.23)
89	5.41 (-0.02)	0.60 (-0.01)	75.72 (+0.72)	92.46 (+0.15)

As proportion of **flatness increases**, test accuracy **consistently increases** without change of # params and FLOPs.



Flatness enables Angle to have better generalizability with much **smoother test-loss surface**.

- **Generalizability on Object Detection**
  - Transfer results on MS-COCO

Method	Params (M)	FLOPs (G)	AP	AP_{50}	AP_{70}	AP_S	AP_M	AP_L
PC-DARTS	5.3	0.59	35.56	55.50	37.45	19.85	38.80	47.70
RLNAS	5.4	0.61	35.98	55.78	38.22	20.80	39.72	47.90
SPOS	5.4	0.60	36.04	56.30	38.08	20.01	39.49	47.76
DropNAS	5.1	0.57	36.39	56.14	38.45	**21.88**	39.82	48.20
GeNAS (ours)	5.2	0.58	**37.05**	**56.92**	**39.19**	20.70	**40.68**	**49.74**

GeNAS can excavate **well-generalizable** architecture on object detection task, compared to other NAS methods.

Conclusion

Our GeNAS framework provides...

- A **promising proxy** for predicting generalizability of a model.
- **Superior generalizability** than conventional search metrics on various tasks and datasets.