FEAR: Ranking Architectures by their Feature-Extraction Capabilities

Debadeepta Dey, Shital Shah, Sebastien Bubeck, Microsoft Research

Bottleneck in Discrete NAS Methods

Evaluating individual architectures
is the most expensive step in discrete NAS methods!

Open Question: How many epochs to evaluate to accurately rank candidates?

Current approaches:

- Pick a small number of epochs. Hope it is enough!
- Training-less measures.
  - Lift-and-shift pruning-at-initialization schemes.
  - Zero-cost measures [Abdelfattah et al. 2021]
    - SNIP [Lee et al. 2018]
    - GRASP [Wang et al. 2020]
    - SYNFLOW [Tanaka et al. 2020]
- NAS-without-training [Mellor et al. 2020]

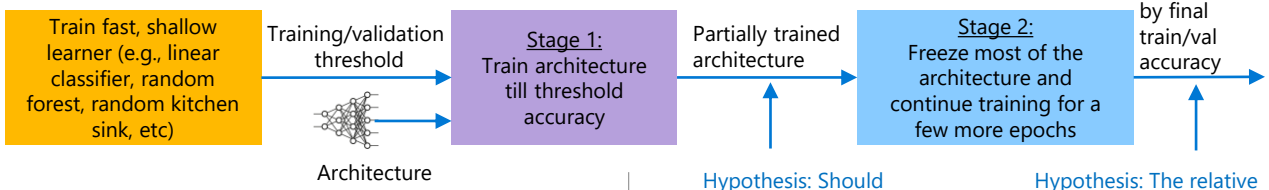
Deeper Dive in Training-less Measures

Synthetic dataset:

- Random 32x32 RGB images.
- 10-way multi-class classification.
- Labeled via 10 randomly initialized two-layer NNs.
- Drop-in replace CIFAR10.

Method	Spearman Corr.
synflow	−0.000437
jacob_cov	−0.13
snip	−0.31
fisher	−0.42
grasp	−0.25
synflow_bn	0.18
FEAR	0.55

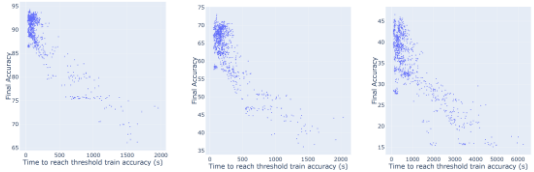
SNIP and GRAD_NORM degrade as network trains!
At init: JACOB_COV: 0.69, 1 epoch: -0.02!
At init: GRASP: 0.63, 1 epoch: -0.41!



Motivation:

- Two-stage training of NNs: [Hu et al. 2020][Chizat and Bach 2020][Nakkiran et al. 2019][Allen-Zhu and Li 2020]
  - Stage 1: Network uses initialization as a kernel embedding and does kernel regression.
  - Stage 2: Actively trains the features from phase 1 to learn a classifier.
- Lower layers train quickly.
- Weak architectures are slow to train (no late bloomers)

Nasbench-201 (1000 networks each randomly sampled)

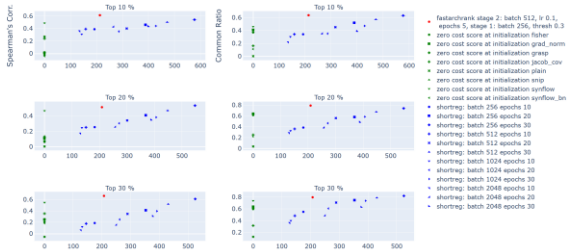


CIFAR10 CIFAR100 ImageNet16-120

Nasbench-301 (21580 networks randomly sampled)

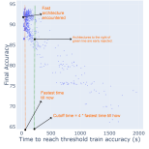
Hypothesis: Should have escaped feature-learning regime.

Hypothesis: The relative performance of architectures depends on the power of the features learnt in stage 1.



Top %	FastArchRank (spe, s)	Nearest Pareto (spe, s)	FastArchRank (common, s)	Nearest Pareto (common, s)
10	$\mathbf{0.61, 213.07}^{\pm 6.33}$	$0.42, 264.68^{\pm 3.52}$	$\mathbf{0.63, 213.04}^{\pm 6.33}$	$0.63, 264.68^{\pm 3.52}$
20	$\mathbf{0.51, 208.84}^{\pm 4.36}$	$0.25, 257.12^{\pm 2.43}$	$\mathbf{0.79, 208.84}^{\pm 4.36}$	$0.79, 257.12^{\pm 2.43}$
30	$\mathbf{0.66, 207.19}^{\pm 3.63}$	$0.15, 249.86^{\pm 2.03}$	$\mathbf{0.79, 207.19}^{\pm 3.63}$	$0.79, 249.86^{\pm 2.03}$
40	$\mathbf{0.70, 205.13}^{\pm 3.16}$	$0.27, 253.67^{\pm 2.18}$	$\mathbf{0.85, 205.13}^{\pm 3.17}$	$0.85, 244.96^{\pm 1.85}$
50	$\mathbf{0.76, 201.65}^{\pm 2.78}$	$0.26, 247.37^{\pm 1.97}$	$\mathbf{0.90, 201.65}^{\pm 2.78}$	$0.90, 239.41^{\pm 1.66}$
100	$\mathbf{0.93, 313.52}^{\pm 9.31}$	$0.90, 329.79^{\pm 2.19}$	$1.0, 313.52^{\pm 9.31}$	$\mathbf{1.00, 281.30}^{\pm 2.42}$

(b) Natsbench CIFAR 100



	FEAR (4.0 * fastest) top1 (%), duration (s)	shortreg (50 epochs) top1 (%), duration (s)
CIFAR10	$\mathbf{93.97}^{\pm 0.08}$, $\mathbf{90560}^{\pm 845}$	$94.10^{\pm 0.10*}$, $347643^{\pm 2287}$
CIFAR100	$\mathbf{72.04}^{\pm 0.29}$, $\mathbf{142550}^{\pm 3106}$	$72.08^{\pm 0.30}$, $347640^{\pm 1674}$
ImageNet16-120	$\mathbf{45.97}^{\pm 0.17}$, $\mathbf{21824}^{\pm 4674}$	$45.64^{\pm 0.21}$, $528454^{\pm 4901}$