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.

Train fast, shallow learner (e.g., linear classifier, random forest, random kitchen sink, etc.)

Stage 1:
Train architecture till threshold accuracy

Partially trained architecture

Stage 2:
Freeze most of the architecture and continue training for a few more epochs

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)

Hypothesis: Should have escaped feature-learning regime.

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

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!

Top %

Nasbench-201 (1000 networks each randomly sampled)

Nasbench-301 (21580 networks randomly sampled)

CIFAR10

CIFAR100

ImageNet16-120

CIFAR10

CIFAR100

ImageNet16-120

syflow 0.000037
jacob_cov -0.13
snip 0.31
fishe 0.62
gras 0.25

SNIP and GRAD_NORM degrade as network trains!