Speeding Up Joint Architecture Hyperparameter Search with Priors

Muhammad Ali, Giacomo Mossio, Omar Swelam
AutoML Final Project

Summary

Joint Neural Architecture and Hyperparameter Search (JAHS) is challenging due to the extensive search space. We show that using expert priors for architecture and hyperparameters speeds up the search by 25%, providing better results under time and epochs constraints.

Approaches

1. **Sequential CNN architectures**
   - Search space expanded to **7 convolutions with 512 channels** for higher capacity models.
   - Performing **NAS and HPO resulted in overfitting**. To have a regularized approach we use the NAS-Bench-201 search space (Dong et al., 2020).
2. **NAS-Bench-201 with HPO**
   - ASHA for JAHS with and without priors.
   - Additional experiment with architecture search space constricted to **10 best-performing** architectures from NAS-Bench-201 pretrained on CIFAR-100.
   - Cifar-100 was chosen due to compatible image resolution and their classification being concerned with subclasses making it more relevant in terms of subclass classification.
3. **Cell search space w/ HPO (Shallow and Wide arch)**
   - Considering the walltime and epoch constraints, we introduce a **2-stage, 3-cell** search space with **64 initial channels**. To better fit the data within the constraints.
   - **Architecture priors** extracted from frequency of edge operation of **100 best-performing architectures** in NAS-Bench-201 on CIFAR-100.
   - **Hyperparameters priors** are expert priors based on previous trials done on top NAS-201 models.
   - **Search space:** learning rate, independent dropout probabilities, weight decay, trivial augment & cell operations.

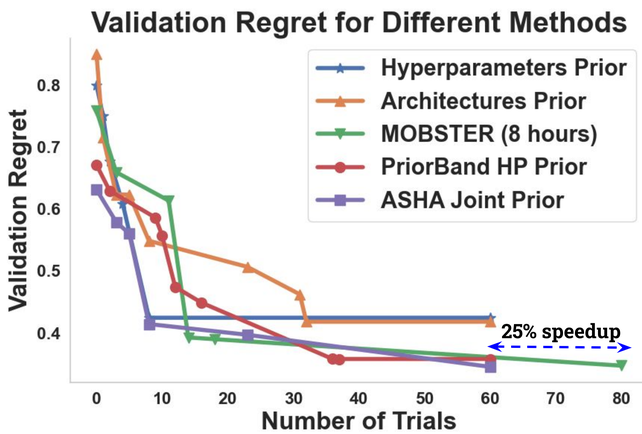
Multi-fidelity algorithms used:

1. **ASHA** (Li et al., 2020): As we use a single worker, this acts as Successive Halving.
2. **MOBSTER** (Klein et al., 2020): Model-based asynchronous multi-fidelity Bayesian optimization.
3. **PriorBand** (Mallick et al., 2023): Bayesian optimization Hyperband combining random sampling, prior sampling & incumbent-based sampling.

Experiment Setup

- Multi-fidelity optimization with epochs as budget type. Compared to image size and dataset subset, epochs had higher correlation coefficients and gave greater val accuracy.
- 20 maximum training epochs per configuration.
- **Three runs of 2 hours each, on three standardized seeds**. Two-fold cross validation.
- Mean test accuracy and standard deviation are reported.
- Test accuracy of an **additional 6 hour run** are also included.
- Hardware used: **GTX 1060 3 GB.**
- Algorithm implementations from **NEPS** library.

Speedup Results



- **Inclusion of priors accelerates the convergence** when having limited time and computation constraints.
- Joint prior method **ASHA-Prior,** provides a nearly **25%** speed-up compared to multi-fidelity BO (MOBSTER).
- PriorBand with HP priors provides similar speedup to ASHA with joint priors.
- All experiments are performed with a limit of 6 hours except for MOBSTER to compare the convergence speedup.

Hyperparameters Analysis

DeepCAVE Local Parameter Importance has shown:

- **For sequential CNN** the most important factors:
  - **HP:** 1) Learning Rate 2) Dropout rate 3) Batch size
  - **Architecture:** 1) Channels in conv_0 2) no. of conv layers
- For the **NAS-Bench-201 Top 10 trials**, focusing on **model selection** from the best 10 models emerged as the most crucial HP.

Performance Results

Method	6h Test Acc	2h Tests Acc
Sequential CNN		
Baseline	63.3	61.9±3.2
NAS w/ default HP	**76.4**	**62.2±3.3**
NAS-Bench-201 search space (16 initial channels)		
ASHA NAS Random HP	54.1	49.81±6.8
MOBSTER	61.4	52.9±5.7
ASHA Joint Priors (1)	64.1	57.9±4.9
Top 10 Archs HPO	**74**	**71.4±1.6**
Cell search space (64 initial channels)		
MOBSTER	68	59±4.9
ASHA HP Priors Only	73.7	59.4±12.3
ASHA Arch Priors Only	76.7	64.8±7.6
PriorBand- HP Priors	73.9	64.6±5.4
ASHA Joint Priors (2)	**79.1**	**68.6±1.3**

- Using **ASHA, HP Priors Only and Arch Priors Only** both outperform MOBSTER (without priors).
- **Architecture priors** provide a **larger performance improvement** than **HP priors**.
- PriorBand outperforms ASHA when run with HP priors.
- We expect PriorBand with joint priors to outperform ASHA with joint priors. PriorBand implementation in NEPS could not be tested due to its incompatibility with our pipeline.

Ablation Study

Method	6h Test Acc	2h Tests Acc
Prior Sampling Fraction		
ASHA- Priors 0.5	79.8	65.2±7.2
ASHA- Priors 0.8	**79.1**	**68.6±1.3**

- The fraction of samples from prior controls the **ratio of** configurations sampled from **prior** distribution by ASHA-Prior.
- **Higher fraction reduces standard deviation** across seeds and results in greater test accuracy after 2h.
- Given that the 6h test accuracy for the 0.5 fraction is higher, this motivates exploring the influence of exploration done by successive halving. Scheduling **different prior fractions followed by BO** using the run history might further achieve better convergence results.