NAS-Bench-301 and the Case for Surrogate Benchmarks for Neural Architecture Search

Julien Siems*, Lucas Zimmer*, Arber Zela, Jovita Lukasik, Margret Keuper, Frank Hutter

Contributions

- **First surrogate NAS benchmark:**
  - Allows *cheap* benchmarking of NAS optimizers
  - Covers *realistic* space (10^{18} architectures)
- We demonstrate that surrogate benchmarks can outperform tabular benchmarks
- **Dataset** of 60k architecture evaluations
- New insights into performance of *Local Search* on DARTS search space.

Existing NAS-Benchmarks

NAS-Bench-101 [Ying et al. 2019]

- Exhaustively evaluated cell-search space on CIFAR-10 for 4 epoch budgets.
  - **~423k unique architectures**

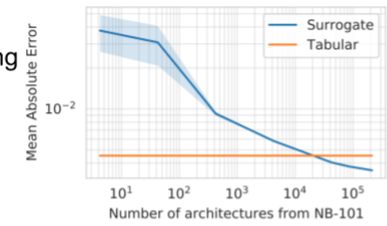
NAS-Bench-201 [Dong and Yi 2020]

- Cell-search space with fixed cell connectivity evaluated on CIFAR-10/CIFAR-100/TinyImage net.
  - **~6k unique architectures**

➡ Benchmarks too *small*: Even Random Search competitive

Tabular vs. Surrogate

- Architecture evaluations are noisy
- Fit surrogate model on NB-101
- Compare to tabular benchmark with one evaluation w.r.t. MAE to remaining evaluations



➡ Surrogate smoothes out noise

➡ Surrogate yields strong predictive performance, even when trained on subset

NAS-Bench-301 - Dataset

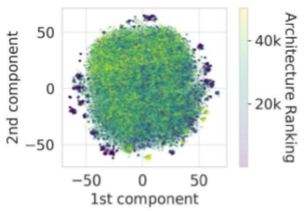
Data Collection

Objective: Cover space efficiently

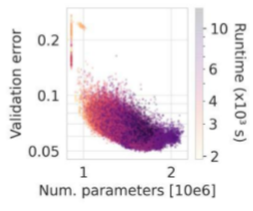
	NAS methods	# eval
	RS (Bergstra & Bengio, 2012)	23746
Evolution	DE (Awad et al., 2020)	7275
	RE (Real et al., 2019)	4639
BO	TPE (Bergstra et al., 2011)	6741
	BANANAS (White et al., 2019)	2243
	COMBO (Oh et al., 2019)	745
One-Shot	DARTS (Liu et al., 2019b)	2053
	PC-DARTS (Xu et al., 2020)	1588
	DrNAS (Chen et al., 2020)	947
	GDAS (Dong & Yang, 2019)	234

➡ *Public Dataset:* **~60k**

Data Visualization



→ Good coverage



→ Wide performance range

NAS-Bench-301 - Surrogate Models

Data Fit

Model	Test R^2	Test sKT
LGBoost	**0.892**	0.816
XGBoost	0.832	**0.817**
GIN	0.832	0.778
NGBoost	0.810	0.759
μ-SVR	0.709	0.677
MLP (Path enc.)	0.704	0.697
RF	0.679	0.683
ϵ-SVR	0.675	0.660

Noise Modelling

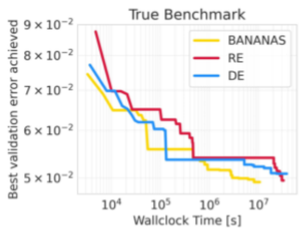
- Deep ensembles
- Evaluate 500 architectures with 5 different seeds
- Reproduce results from Tabular vs. Surrogate experiment

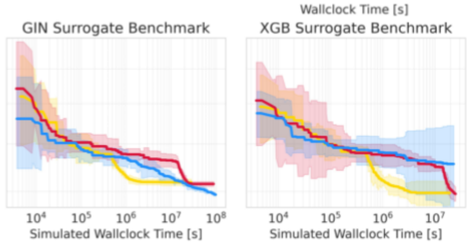
Model	MAE 1, [2,3,4,5]	Mean σ	KL div.
Tabular	1.38e−3	undef.	undef.
GIN	**1.13e-3**	0.6e−3	**16.4**
LGB	1.33e−3	0.3e−3	68.9
XGB	1.51e−3	0.3e−3	134.4

➡ **Use XGB, LGB & GIN**

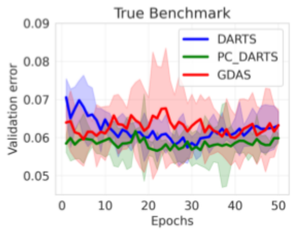
NAS-Bench-301 - Benchmark

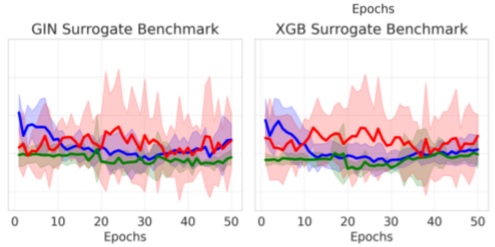
Blackbox Optimizers





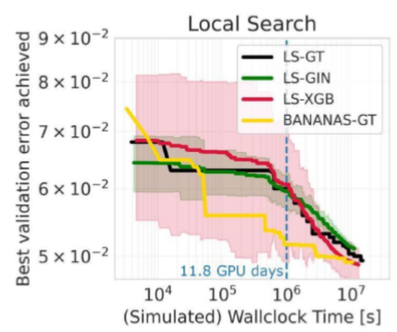
One-Shot Optimizers





NAS-Bench-301 - Case Study

- *Local Search* result by White et al. 2020:
  - Not competitive on DART Search Sp.
  - Tested with small compute budget
- Fast NB-301 analysis suggests it *is competitive.*
  - LS-GIN, LS-XGB
- Verified by extensive empirical evaluation
  - LS-GT



➡ Demonstration of how NAS-Bench-301 can be used to cheaply check research hypothesis.

References

- Ying, Chris, et al. "Nas-bench-101: Towards reproducible neural architecture search." International Conference on Machine Learning. 2019.
- Dong, Xuanyi, and Yi Yang. "NAS-Bench-201: Extending the Scope of Reproducible Neural Architecture Search." International Conference on Learning Representations. 2019.
- White, Colin, Sam Nolen, and Yash Savani. "Local Search is State of the Art for NAS Benchmarks." arXiv preprint arXiv:2005.02960 (2020).

Benchmark Publicly Available: https://github.com/automl/nasbench301

AutoML.org

@automlfreiburg