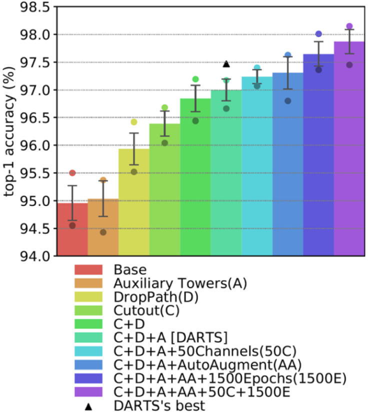
AutoML: Neural Architecture Search (NAS)

Issues and Best Practices in NAS Research

Bernd Bischl   Frank Hutter   Lars Kotthoff
Marius Lindauer   Joaquin Vanschoren

Issues in NAS Research & Evaluations

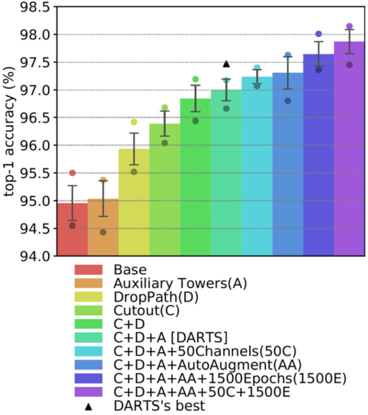
- Most NAS methods are extremely difficult to reproduce and compare [Li and Talwalkar. 2019]
- For benchmarks used in almost all NAS papers:
  - Training pipeline matters much more than neural architecture



[Yang et al. 2020]

Issues in NAS Research & Evaluations

- Most NAS methods are extremely difficult to reproduce and compare [Li and Talwalkar. 2019]
- For benchmarks used in almost all NAS papers:
  - Training pipeline matters much more than neural architecture
- The final benchmark results reported in different papers are typically incomparable
  - Different training code (often unavailable)
  - Different search spaces
  - Different evaluation schemes

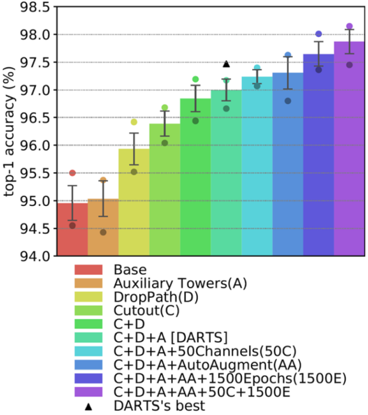


[Yang et al. 2020]

Issues in NAS Research & Evaluations

- Most NAS methods are extremely difficult to reproduce and compare [Li and Talwalkar. 2019]
- For benchmarks used in almost all NAS papers:
  - Training pipeline matters much more than neural architecture
- The final benchmark results reported in different papers are typically incomparable
  - Different training code (often unavailable)
  - Different search spaces
  - Different evaluation schemes

→ We emphasize concepts, not published performance numbers



[Yang et al. 2020]

Building a Scientific Community for NAS

- Benchmarks
  - NAS-Bench-101 [Ying et al. 2019]
  - NAS-Bench-201 [Dong and Yang. 2020]
  - NAS-Bench-1Shot1 [Zela et al. 2020]

Building a Scientific Community for NAS

- Benchmarks
  - NAS-Bench-101 [Ying et al. 2019]
  - NAS-Bench-201 [Dong and Yang. 2020]
  - NAS-Bench-1Shot1 [Zela et al. 2020]
- Best Practice Checklist for Scientific Research in NAS
  [Lindauer and Hutter. 2020]

Building a Scientific Community for NAS

- Benchmarks
  - NAS-Bench-101 [Ying et al. 2019]
  - NAS-Bench-201 [Dong and Yang. 2020]
  - NAS-Bench-1Shot1 [Zela et al. 2020]
- Best Practice Checklist for Scientific Research in NAS
  [Lindauer and Hutter. 2020]
- Unifying open-source implementation of modern NAS algorithms
  [Zela et al. 2020]
  - Finally enables empirical comparisons without confounding factors

Building a Scientific Community for NAS

- Benchmarks
  - NAS-Bench-101 [Ying et al. 2019]
  - NAS-Bench-201 [Dong and Yang. 2020]
  - NAS-Bench-1Shot1 [Zela et al. 2020]
- Best Practice Checklist for Scientific Research in NAS [Lindauer and Hutter. 2020]
- Unifying open-source implementation of modern NAS algorithms [Zela et al. 2020]
  - Finally enables empirical comparisons without confounding factors
- First NAS workshop at ICLR 2020

Best Practice Checklist for NAS Research [Lindauer and Hutter. 2020]

- Best practices for releasing code
  - Code for the training pipeline used to evaluate the final architectures
  - Hyperparameters used for the final evaluation pipeline, as well as random seeds
  - Code for the search space

Best Practice Checklist for NAS Research [Lindauer and Hutter. 2020]

- Best practices for releasing code
  - Code for the training pipeline used to evaluate the final architectures
  - Hyperparameters used for the final evaluation pipeline, as well as random seeds
  - Code for the search space
  - Code for your NAS method
  - Hyperparameters for your NAS method, as well as random seeds

Best Practice Checklist for NAS Research [Lindauer and Hutter. 2020]

- Best practices for releasing code
  - Code for the training pipeline used to evaluate the final architectures
  - Hyperparameters used for the final evaluation pipeline, as well as random seeds
  - Code for the search space
  - Code for your NAS method
  - Hyperparameters for your NAS method, as well as random seeds
- Note that the easiest way to satisfy the first three is to use existing NAS benchmarks

Definition: NAS Benchmark [Lindauer and Hutter. 2020]

A NAS benchmark consists of a dataset (with a predifiend training-test split), a search space, and available runnable code with pre-defined hyperparameters for training the architectures.

Best Practice Checklist for NAS Research [Lindauer and Hutter. 2020]

- Best practices for comparing NAS methods
  - For all NAS methods you compare, did you use exactly the same NAS benchmark, including the same dataset (with the same training-test split), search space and code for training the architectures and hyperparameters for that code?

Best Practice Checklist for NAS Research [Lindauer and Hutter. 2020]

- Best practices for comparing NAS methods
  - For all NAS methods you compare, did you use exactly the same NAS benchmark, including the same dataset (with the same training-test split), search space and code for training the architectures and hyperparameters for that code?
  - Did you control for confounding factors (different hardware, versions of DL libraries, different runtimes for the different methods)?

Best Practice Checklist for NAS Research [Lindauer and Hutter. 2020]

- Best practices for comparing NAS methods
  - For all NAS methods you compare, did you use exactly the same NAS benchmark, including the same dataset (with the same training-test split), search space and code for training the architectures and hyperparameters for that code?
  - Did you control for confounding factors (different hardware, versions of DL libraries, different runtimes for the different methods)?
  - Did you run ablation studies?

Best Practice Checklist for NAS Research [Lindauer and Hutter. 2020]

- Best practices for comparing NAS methods
  - For all NAS methods you compare, did you use exactly the same NAS benchmark, including the same dataset (with the same training-test split), search space and code for training the architectures and hyperparameters for that code?
  - Did you control for confounding factors (different hardware, versions of DL libraries, different runtimes for the different methods)?
  - Did you run ablation studies?
  - Did you use the same evaluation protocol for the methods being compared?

Best Practice Checklist for NAS Research [Lindauer and Hutter. 2020]

- Best practices for comparing NAS methods
  - For all NAS methods you compare, did you use exactly the same NAS benchmark, including the same dataset (with the same training-test split), search space and code for training the architectures and hyperparameters for that code?
  - Did you control for confounding factors (different hardware, versions of DL libraries, different runtimes for the different methods)?
  - Did you run ablation studies?
  - Did you use the same evaluation protocol for the methods being compared?
  - Did you compare performance over time?

Best Practice Checklist for NAS Research [Lindauer and Hutter. 2020]

- Best practices for comparing NAS methods
  - For all NAS methods you compare, did you use exactly the same NAS benchmark, including the same dataset (with the same training-test split), search space and code for training the architectures and hyperparameters for that code?
  - Did you control for confounding factors (different hardware, versions of DL libraries, different runtimes for the different methods)?
  - Did you run ablation studies?
  - Did you use the same evaluation protocol for the methods being compared?
  - Did you compare performance over time?
  - Did you compare to random search?

Best Practice Checklist for NAS Research [Lindauer and Hutter. 2020]

- Best practices for comparing NAS methods
  - For all NAS methods you compare, did you use exactly the same NAS benchmark, including the same dataset (with the same training-test split), search space and code for training the architectures and hyperparameters for that code?
  - Did you control for confounding factors (different hardware, versions of DL libraries, different runtimes for the different methods)?
  - Did you run ablation studies?
  - Did you use the same evaluation protocol for the methods being compared?
  - Did you compare performance over time?
  - Did you compare to random search?
  - Did you perform multiple runs of your experiments and report seeds?

Best Practice Checklist for NAS Research [Lindauer and Hutter. 2020]

- Best practices for comparing NAS methods
  - For all NAS methods you compare, did you use exactly the same NAS benchmark, including the same dataset (with the same training-test split), search space and code for training the architectures and hyperparameters for that code?
  - Did you control for confounding factors (different hardware, versions of DL libraries, different runtimes for the different methods)?
  - Did you run ablation studies?
  - Did you use the same evaluation protocol for the methods being compared?
  - Did you compare performance over time?
  - Did you compare to random search?
  - Did you perform multiple runs of your experiments and report seeds?
  - Did you use tabular or surrogate benchmarks for in-depth evaluations?

Best Practice Checklist for NAS Research [Lindauer and Hutter. 2020]

- Best practices for reporting important details
  - Did you report how you tuned hyperparameters, and what time and resources this required?

Best Practice Checklist for NAS Research [Lindauer and Hutter. 2020]

- Best practices for reporting important details
  - Did you report how you tuned hyperparameters, and what time and resources this required?
  - Did you report the time for the entire end-to-end NAS method (rather than, e.g., only for the search phase)?

Best Practice Checklist for NAS Research [Lindauer and Hutter. 2020]

- Best practices for reporting important details
  - Did you report how you tuned hyperparameters, and what time and resources this required?
  - Did you report the time for the entire end-to-end NAS method (rather than, e.g., only for the search phase)?
  - Did you report all the details of your experimental setup?

Best Practice Checklist for NAS Research [Lindauer and Hutter. 2020]

- Best practices for reporting important details
  - Did you report how you tuned hyperparameters, and what time and resources this required?
  - Did you report the time for the entire end-to-end NAS method (rather than, e.g., only for the search phase)?
  - Did you report all the details of your experimental setup?

- It might not always be possible to satisfy all these best practices, but being aware of them is the first step . . .

Best Practice Checklist for NAS Research [Lindauer and Hutter. 2020]

- Best practices for reporting important details
  - Did you report how you tuned hyperparameters, and what time and resources this required?
  - Did you report the time for the entire end-to-end NAS method (rather than, e.g., only for the search phase)?
  - Did you report all the details of your experimental setup?

- It might not always be possible to satisfy all these best practices, but being aware of them is the first step . . .

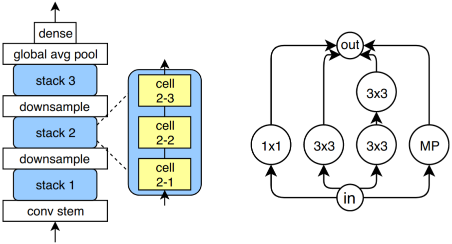
- We believe the community would benefit a lot from:
  - Clean NAS benchmarks for new applications
    - Including all details for the application. No need to also develop a new method.
  - Open-source library of NAS methods to compare methods without confounding factors
    - First version already developed: NASlib [Zela et al, under review]

NAS-Bench-101: The First NAS Benchmark [Ying et al. 2019]

- Dataset: CIFAR-10, with the standard training/test split
- Runnable open-source code provided in Tensorflow
- Cell-structured search space consisting of all directed acyclic graphs (DAGs) on V nodes, where each possible node has L operation choices.

NAS-Bench-101: The First NAS Benchmark [Ying et al. 2019]

- Dataset: CIFAR-10, with the standard training/test split
- Runnable open-source code provided in Tensorflow
- Cell-structured search space consisting of all directed acyclic graphs (DAGs) on V nodes, where each possible node has L operation choices.
- To limit the number of architectures, NAS-Bench-101 has the following constraints:
  - $L = 3$ operators:
    - 3×3 convolution - 1×1 convolution - 3×3 max-pooling
  - $V \leq 7$ nodes
  - A maximum of 9 edges



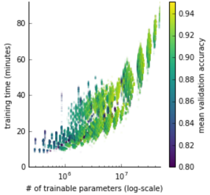
NAS-Bench-101: The First Tabular NAS Benchmark [Ying et al. 2019]

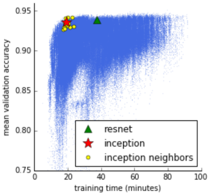
- Tabular benchmark: we exhaustively trained and evaluated all possible models on CIFAR-10 to create a tabular (look-up table) benchmark
- Based on this table, anyone can now run NAS experiments in seconds without a GPU.

NAS-Bench-101: The First Tabular NAS Benchmark [Ying et al. 2019]

- Tabular benchmark: we exhaustively trained and evaluated all possible models on CIFAR-10 to create a tabular (look-up table) benchmark
- Based on this table, anyone can now run NAS experiments in seconds without a GPU.

- Around 423k unique cells
  - 4 epoch budgets: 4, 12, 36, 108
  - 3 repeats
  - around 5M trained and evaluated models
  - 120 TPU years of computation
  - the best architecture mean test accuracy: 94.32%

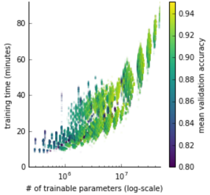




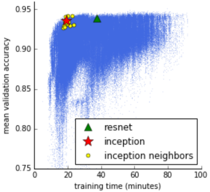
NAS-Bench-101: The First Tabular NAS Benchmark [Ying et al. 2019]

- Tabular benchmark: we exhaustively trained and evaluated all possible models on CIFAR-10 to create a tabular (look-up table) benchmark
- Based on this table, anyone can now run NAS experiments in seconds without a GPU.

- Around 423k unique cells
  - 4 epoch budgets: 4, 12, 36, 108
  - 3 repeats
  - around 5M trained and evaluated models
  - 120 TPU years of computation
  - the best architecture mean test accuracy: 94.32%

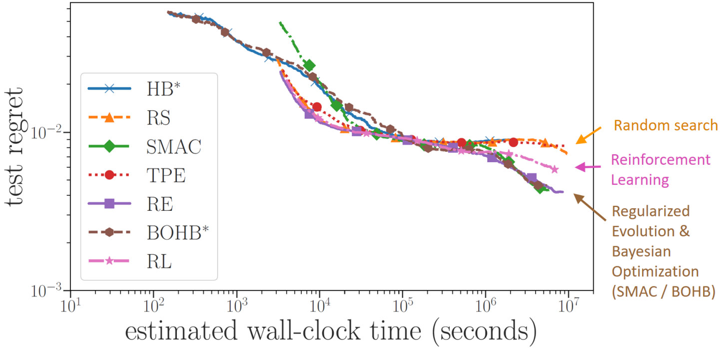


- Given an architecture encoding A, budget E_{stop} and trial number, one can query from NAS-Bench-101 the following quantities:
  - training/validation/test accuracy
  - training time in seconds
  - number of trainable model parameters



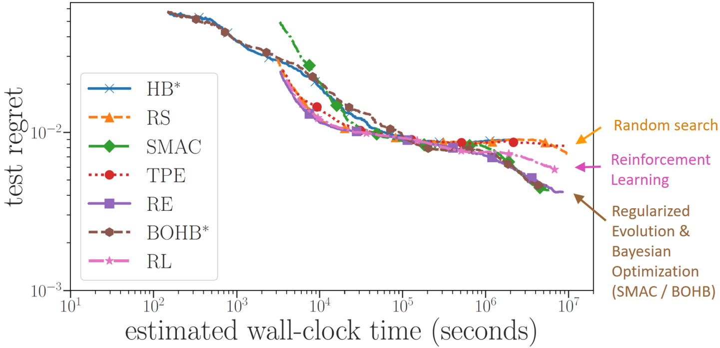
Evaluation of Blackbox NAS Methods on NAS-Bench-101 [Ying et al. 2019]

- RL outperforms random search
- BO and regularized evolution perform best, better than RL



Evaluation of Blackbox NAS Methods on NAS-Bench-101 [Ying et al. 2019]

- RL outperforms random search
- BO and regularized evolution perform best, better than RL



- Note that the BO method SMAC [Hutter et al. 2011] predated RL for NAS [Zoph and Le. 2017] by 6 years
  - Only now, benchmarks like NAS-Bench-101 allow for efficient comparisons

Questions to Answer for Yourself / Discuss with Friends

- Repetition:
  For the most common NAS search space, how important is the NAS component compared to the importance of the training pipeline used?

- Repetition:
  Why do we need proper benchmarking of NAS algorithms?

- Repetition:
  What does a NAS benchmark consist of?

- Repetition:
  List all best practices for NAS you remember.