The Next Generation of Benchmarks for Automated Deep Learning

Research Institution
[1]University of Freiburg
[2]Leibniz Supercomputing Centre

Principal Investigator
Frank Hutter[1]

Researchers
Archit Bansal[1], Danny Stoll[1], David Brayford[2]

Project Partners
–

SuperMUC Project ID: pn68xi

Introduction

In recent years, Deep Learning (DL) has been incredibly successful in different areas, from speech recognition and automatic image analysis to prototypes of autonomous driving and world-class level algorithms for playing computer games and board games, such as "Go". The reasons for this success are: (1) the availability of massive training data; (2) advanced DL techniques that can accurately model learning problems by overparameterized large neural networks, called Deep Neural Networks (DNN); and (3) advances in computing software/hardware that made training DNNs affordable. This popularity was promoted by novel DNN architectures for image classification/segmentation tasks, such as ResNet, InceptionNet, DenseNet and U-Net. However, widespread use in other areas is still hindered due to: (1) the lack of sufficient data; and (2) the complex development process of new problem-specific DNN architectures. Our research focuses on the latter. The process of designing DNN architectures is still mostly based on experience and trial-and-error. Inexperienced users are often unable to unlock the power of DL as they struggle to navigate the search space of possible architectures: number of layers, proper operations (e.g, convolution, pooling and skip connection), etc. Our work aims to overcome these issues by investigating Neural Architecture Search (NAS) - a key area in automated DL. The underlying idea is to view the problem of designing a DNN architecture as an optimization problem. A growing body of work now shows that automatically generated DNNs can outperform manually created ones. More generally, NAS is also the natural next step after deep learning: while deep learning automatically learns representations, with NAS we now automatically learn the architectures that allow us to learn those representations. Therefore, NAS has become one of the hottest and most intensely researched directions in deep learning, with an exponentially growing number of publications on the topic. NAS is, however, remarkably costly as it might involve training hundreds of thousands of DNN architectures, taking from days to weeks or even months. Even more concerning is the carbon footprint of such simulations due to the enormous power consumption. In order to fix this, PI Frank Hutter has spearheaded the creation of so-called tabular NAS benchmarks, which train a large number of DNNs architectures and record their validation performance in a table, in order to allow very fast evaluations of NAS algorithms after the one-time cost of creating the tabular NAS benchmark. This line of work has been extremely successful, but current NAS benchmarks are still limited in several ways:

- No logging of side information.
- No variation of hyperparameters.
- Limitation to a single or very few datasets.
- Limitation to simple search spaces.
- Focus on expensive evaluations, rather than cheap multi-fidelity evaluations used by modern NAS optimizers.

In our work we use the SuperMUC-NG cluster to create the next generation of NAS benchmarks to overcome all of these limitations. This next generation of benchmarks will facilitate research on NAS, while reducing the environmental impact of conducting such research and lowering the barrier to entry in this field of research. An exciting aspect of our work for the HPC field is that, due to our focus on cheap multi-fidelity evaluations (reduced network size, reduced image resolutions, etc), in contrast to the predominant use of GPUs in other areas of deep learning, our experiments are actually much more cost-efficient to perform on a large cluster of CPUs; thus, we carry them out on SuperMUC-NG. More information about our research group is available at our website [1] and about our research projects at the project website [2].

Results and Methods

Our on-going project has already resulted in the creation of the largest known benchmark dataset in the field to date, containing 54 million data points each consisting of over 20 metrics about the performance characteristics of the trained models. Our benchmark will facilitate research on automated Deep Learning in

Property	Reference	Ours
# Architectures	~ 1,000	15,625
# Hyperparameters	2 x discrete	2 x discrete, 2 x continuous
# Fidelities	0	4
# Metrics	9	20
# Datasets	1	3
# Trained DL models	48,000	810,000
# Data points	192,000	140,000,000
Extensive logging of side information	No	Yes

Table 1: This table compares the current status of our ongoing project to the most recent work in the field, titled NAS-HPO-Bench-II (October, 2021), that shares a number of common properties with our work but is much more limited in its scope.

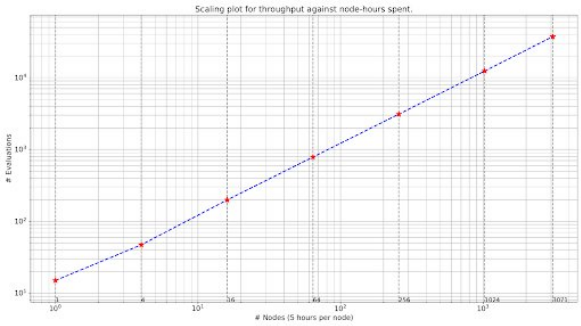

Figure 1: Scaling graph created from scaling studies performed for the block operation.

several ways:

- Since our search space consists of both architectures and hyperparameters, our benchmark supports research for NAS, Hyperparameter Optimization (HPO) and joint NAS and HPO.
- Our dataset contains 5,400 unique fidelity settings spread over 4 different fidelity parameters that control the cost of training a neural network (most benchmarks are limited to 0 or 1 fidelity parameter), a unique feature of our benchmark that would support researchers who want to employ Multi-Fidelity Optimization techniques in NAS and HPO.
- Our dataset also records 20 metrics and heterostructures that could be used by researchers to employ Multi-Objective Optimization techniques for their research.
- Since our dataset also spans a growing number of datasets (3 thus far, more are in progress), researchers will also have an opportunity to employ Transfer Learning techniques with our benchmark.
- Additionally, we have a collection of nearly 10 TB of checkpoint data, which can, e.g., be used when benchmarking algorithms that employ so-called Zero-Cost Proxies, another recent direction in NAS research.

See Table 1 for a comparison with a reference benchmark published in October, 2021. We employed a job setup that is known to be quite possibly the most efficient way to utilize parallel compute resources for running any calculation, known as an “Embarrassingly Parallel” job setup. This allows us to scale-up the number of computations we perform in parallel, thereby increasing cluster usage efficiency, almost infinitely. With this job setup we have, thus far, managed to train 810k unique DL model configurations that consist of unique combinations of architecture, hyperparameter and fidelity choice on three different datasets. To collect this data, we used a total of 27M core hours. Figure 1 illustrates the near linear scaling efficiency of the parallel processing strategy employed. This information was used to by the LRZ to determine if the workflow could be deployed on a significant portion of SuperMUC-NG during the “block operation”. In the “block operation” the project was provided exclusive use of SuperMUC-NG for 2 days. During this time significant amount of compute resource were used and the benchmarks generated are being used by another project that produced interesting results, which will go into future publications [3,4] and a Master’s thesis [5]. Our work offers a staggering amount of cost-savings in the form of DNN evaluations that can be replaced by queries to our benchmark: On the one hand, the average amount of compute-time needed to evaluate the most important class of models in our benchmark is greater than 45 core-hours. On the other hand, a single query to our benchmark produces equivalent data at the cost of less than 0.1 core-seconds. A single research paper may need to evaluate many hundreds of thousands of such models, and using our benchmark will thus both help democratize research in this field and save very substantial compute time & carbon emissions. As a result, other related NAS benchmarks have been cited by hundreds of research papers, and we expect the same for our new benchmark.

Ongoing Research / Outlook

Currently, we use the remaining compute resources of our project to create benchmarks for more datasets, as the inclusion of only one or very few target datasets is one of the limitations of existing NAS benchmarks. For this, we are able to utilize the massive number of parallel compute nodes available on the SuperMUC-NG to run simulations for up to 100k Deep Learning trainings in parallel. The greatest limitation we face is the availability of RAM on a per-node basis, which bottlenecks the number of parallel computations that we can perform. Nonetheless, the SuperMUC-NG provides us with parallel compute resources well beyond what we see in related work (see Table 1). Finally, we plan to create NAS benchmarks based on complex hierarchical search spaces, which are in contrast to the simple search spaces of existing benchmarks. As the NAS research community is only starting to tackle hierarchical NAS, cheap and rigorous ways to benchmark in this area are of dire need. Within our remaining compute resources we aim to create a proof-of-concept for such benchmarks and will apply for a follow-up project focusing on these hierarchical NAS benchmarks.

References and Links

[1] https://ml.informatik.uni-freiburg.de/
[2] https://www.automl.org/
[3] A. Bansal, D. Stoll, M. Janowski, A. Zela, F. Hutter, "JAHS-Bench-201: A Foundation For Research On Joint Architecture And Hyperparameter SearchŁ". Submitted for peer-review at the 36th Conference on Neural Information Processing Systems, Datasets and Benchmarks Track. June, 2022.
[4] https://github.com/automl/jahs_bench_201
[5] Janowski et al., paper in preparation.