

AIPerf: Automated machine learning as an AI-HPC benchmark

1^{*} Zhixiang Ren

Peng Cheng Laboratory
renzhx@pcl.ac.cn

2nd Yongheng Liu

Peng Cheng Laboratory
yongheng.liu@pcl.ac.cn

3rd Tianhui Shi

Department of Computer Science and Technology
Tsinghua University
sth19@mails.tsinghua.edu.cn

4th Lei Xie

Department of Computer Science and Technology
Tsinghua University
xie-l18@mails.tsinghua.edu.cn

5th Yue Zhou

Peng Cheng Laboratory
zhouy@pcl.ac.cn

6th Jidong Zhai

Department of Computer Science and Technology
Tsinghua University
zhaijidong@tsinghua.edu.cn

7th Youhui Zhang

Department of Computer Science and Technology
Tsinghua University
zyh02@tsinghua.edu.cn

8th Yunquan Zhang

Institute of Computing Technology
Chinese Academy of Sciences
zyq@ict.ac.cn

^{*} Wenguang Chen

Department of Computer Science and Technology
Tsinghua University
cwg@tsinghua.edu.cn

Abstract—The plethora of complex artificial intelligence (AI) algorithms and available high performance computing (HPC) power stimulates the expeditious development of AI components with heterogeneous designs. Consequently, the need for cross-stack performance benchmarking of AI-HPC systems emerges rapidly. The current HPC benchmarks can not reflect AI computing power without representative workloads and the current AI benchmarks have fixed problem size therefore limited scalability. To address these issues, we propose an end-to-end benchmark suite utilizing automated machine learning (AutoML) that represents real AI scenarios. More importantly, AutoML is auto-adaptive to various scales of machines with an extreme computational cost therefore a desired workload. We implement the algorithms in a highly parallel and flexible way to ensure the efficiency and optimization potential on diverse systems with customizable configurations. The major metric to quantify the performance is floating-point operations per second (FLOPS) that is measured in an analytical and systematic approach. We verify the benchmark’s stability at discrete timestamps and the linear scalability on various numbers of machines equipped with up to 400 AI accelerators. With flexible workload size as well as single metric measurement, our benchmark can scale from small clusters to large AI-HPC and rank them easily. The source code, specifications and detailed procedures are publicly accessible on GitHub¹.

Index Terms—artificial intelligence, high performance computing, benchmark, automated machine learning

I. INTRODUCTION

Artificial intelligence (AI), machine learning (ML) and deep learning (DL) have drawn tremendous attention in recent years.

DL requires a training process [39], which is essentially a multi-dimensional fitting, to automatically adjust the weights (parameters) of the neural network. As the learnable data grows at an unprecedented rate, the high performance computing (HPC) machines are needed for the large AI model to harness the big data and extract the complex abstractions [58]. The hybrid HPC models with AI surrogates reveal a collection of unique and novel opportunities for scientific breakthrough and unforeseeable discoveries [31], as well as business innovations and other societal benefits. The increase in algorithmic advances of AI algorithms, available computing power and data collections, as well as the demand for scalable and data-driven solutions stimulate the convergence of AI and HPC machines [18]. The convergence [57] still faces multiple challenges, like the effective and parallel implementation of algorithms on large scale clusters, high bandwidth as well as low latency communications between distributed workers, and high-speed interconnections to the network file system etc., therefore a comprehensive evaluation is needed to address all aspects of the system. Also, though a unified fashion is preferred, the relentless pace of the development demands evolving and diverse architectural designs. The HPC system needs to incorporate the support for various AI workloads on top of inconsistent accelerators and software frameworks for an AI-HPC adaption. Consequently, the need for an open and reliable benchmark suite to evaluate the cross-stack performance of heterogeneous AI-HPC systems emerges rapidly, as shown in Figure 1.

¹ AIPerf: <https://github.com/AI-HPC-Research-Team/AIPerf>

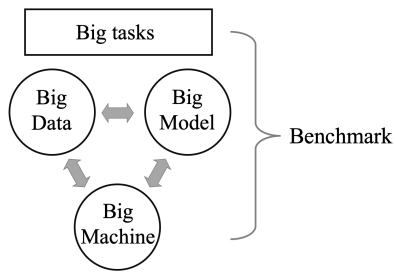


Fig. 1. The converge of AI and HPC with the growth of model, machine, data, as well as potential tasks. The benchmark should cover the system heterogeneity and reflect the cross-stack performance.

There are three major challenges for AI-HPC benchmarking. First, the benchmark workload needs to represent the real problems regarding hardware utilization, set-up cost and computing patterns etc. Second, the benchmark workload is preferably auto-adaptive to various scales of machines without extra human effort. Third, using simple and fewer metrics to measure the whole system performance on AI applications. Unfortunately, the current HPC and AI benchmarks do not address these challenges. Popular HPC benchmarks like LINPACK can not reflect the cross-stack performance on AI without representative workloads. The AI benchmarks like MLPerf [41] have a fixed workload size. This is fallacious since the increased computing power tends to be utilized attack larger problems instead of the same problem with less time. The fixed problems can also not adapt to different scales of machines automatically. Automated machine learning (AutoML) can search and optimize the AI models more automatically and it is getting increasing attention in the AI community. As an representative AI application, AutoML contains nearly all the critical components regarding the primary computing operations (e.g. sparse matrices multiplication), calculation precision (in FP-32 or lower) and workflow in the real AI scenarios. More importantly, AutoML is auto-adaptive to various scales of machines and requires extreme computational cost, therefore a desired workload. Considering all the advantages, we choose AutoML to be the benchmark workload to tackle the first two challenges. We utilize floating-point operations per second (FLOPS), which is conventionally favored in the HPC community, as our benchmark score to quantitatively measure the machine computing power on AI. The FLOPS is measured in an analytical and systematic method to account for both training and inference processes. With the auto-adaptive workload and single metric measurement, our benchmark can easily rank the various size of machines from small clusters to large AI-HPC.

In summary, our main contributions are the following:

- We propose AutoML as a representative and auto-adaptive workload of various scales of machines to establish an end-to-end benchmark suite.
- Our implementation is highly parallel and customizable to keep the optimization potential on diverse systems.

- We propose an analytical and speedy approach to calculate the operation count of various neural networks with different architectures.
- We evaluate our benchmark on large clusters with big data and validate the benchmark’s linear scalability and stability.

The rest of this paper is organized as follows. In Section II, we review the existing HPC and AI benchmarks and point out their downsides for AI-HPC benchmarking. In Section III, we briefly review AutoML and the popular frameworks for AI. In Section IV, we describe the details of our algorithms, implementations and measurements. In Section V, we evaluate our benchmark on different scales of machines. We summarize our work in Section VI.

II. RELATED WORK

A. HPC Benchmarks

LINPACK² is the de facto HPC benchmark nowadays. It is essentially an algebra library that solves a dense system of linear equations that is the heart of many computational science problems. There are three reasons why LINPACK is not suitable for benchmarking AI-HPC. First, the problem size is usually manually decided and can not be scaled automatically based on the tested machines. Second, LINPACK provides little information about the set-up cost and I/O ability, which are critical data-intensive applications like AI. This is problematic since most algorithms do more data motion than arithmetic [22]. Third, the calculation is performed in FP-64, while most AI applications typically only require FP-32 or even FP-16. HPL-AI Mixed-Precision benchmark [11] is developed based on LINPACK to highlight the third issue, but it still suffers from the other two issues. Other HPC benchmarks including NASA Parallel Benchmarks³, SLALOM [21] and HINT [22] do not utilize workloads that can represent real AI scenarios, therefore share the same problems as LINPACK. Though we can not use the existing HPC benchmarks for AI-HPC, they still inspire us in the benchmark design. For example, the biggest challenge in benchmarking is to create a single workload that can capture all the features of real applications and be auto-adaptive without a fixed problem size. Also, further performance optimization with customizable configuration is encouraged, as long as the user does not specialize the program to input data. Last but not least, a single number metric is preferred for easy comparison and ranking.

B. AI Benchmarks

Fair and inclusive comparison of machine computing power on AI applications is not trivial. As the opposite of monoculture, the system’s heterogeneity, the variety of AI workloads and the stochastic nature of approaches make the benchmarking complicated. Previous AI benchmarks attempt to highlight the challenges by incorporating different hardware systems [4], [16], [34], [52], [54], software frameworks [28]

²<https://www.top500.org/project/linpack/>

³<https://www.nas.nasa.gov/publications/npb.html>

or AI algorithms [2], [3], [19], [59]. More recently, end-to-end benchmarks include [7], [14], [36], [41], [60] are developed to evaluate hardware systems and AI algorithms simultaneously. MLPerf [41], the arguably most accepted AI benchmark so far, uses time-to-accuracy to measure the co-performance of hardware and software. This metric is an indirect quantification of the computing ability comparing to FLOPS, which is our metric. Since MLPerf is composed of multiple micro-tasks, each one would result in a different measurement. Though this approach makes the benchmark more accurate on various applications, it also makes the comparison and ranking more difficult. Also, the limited workloads in MLPerf have insufficient scalability with fixed problem size. Other AI benchmarks have similar drawbacks as MLPerf. Overall, there are two reasons why the existing AI benchmarks are not suitable to be AI-HPC benchmark:

- Existing AI benchmarks have fixed problem size therefore limited scalability.
- Existing AI benchmarks do not provide a single and direct measurement to quantify performance.

III. BACKGROUND

A. Automated Machine Learning

Developing AI solutions have mostly relied on a complex model design which involves human expertise heavily and is extremely time-consuming. To explore the architecture space more efficiently and optimize the model automatically, AutoML [33]) emerges as the AI model complexity increases exponentially⁴ in recent years. It may sound surprising but AutoML is already mature enough to rival human experts to make a real impact on AI research. Overall, AutoML is inherently computing-intensive, highly scalable and representative of AI-like workflows. Considering all the unique advantages, we choose AutoML as our benchmark workload. As shown in Figure 2, AutoML contains various parts [25]. The first part is data preparation, which involves data collection and data cleaning. The second part is feature engineering, including feature selection, feature construction and feature extraction. Although data and features lay the foundations of AI performance, they depend on the application scenarios and are irrelevant to the machine computing power, therefore not considered in our benchmark. The third part is to generate the neural architecture and the optimal configuration (referred to as hyperparameters), which can have a significant impact on the performance. The two main approaches for model generation are the experts' manual design and the automated neural architecture search (NAS [17]). Without human intervention, NAS has the potential to generate novel architectures beyond imagination and can boost the performance significantly. Hyperparameter optimization (HPO [8]) is essentially the optimization of the loss function over the complex configuration space. The NAS and HPO can be implemented in a parallel manner to fully utilize the distributed resources. Finally, model evaluation measures the performance once the candidate model is generated. The

simplest method is to conduct the inference on the test dataset for enough epochs. This is prohibitively expensive since there are numerous configurations for each neural architecture. In this paper, We use warm-up and early stopping strategies [45] that stops the training once the validation loss flats and this can provide measurements quickly to a certain degree of accuracy.

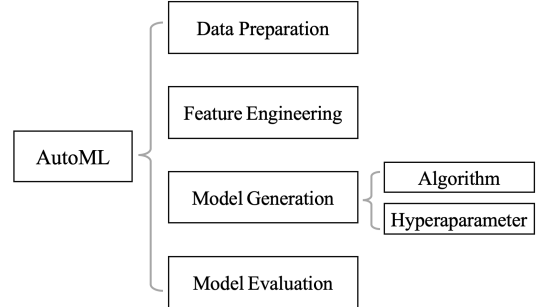


Fig. 2. An overview of the AutoML. We limit our attention to the model generation in this paper.

B. Frameworks

a) *Deep Learning Frameworks*: DL frameworks provide user-friendly API and transform programs in high-level languages into an internal representation of certain functionalities. The low-level efficient libraries, e.g. cuDNN, are invoked to execute primary operations like matrix multiplication. Multiple solutions with desired performance exist [20], therefore implementation and customized setups vary while maintaining similar results. The difference is critical as the training process is stochastic and approximate intrinsically. An open-source framework with enough community support would be a decent candidate for building the benchmark. According to GitHub⁵, the most popular deep learning frameworks are TensorFlow [1], Keras [13] and PyTorch [44]. TensorFlow is an open-source library for low-level numerical calculation with static computational graphs where operations are written as high-performance C++ binaries with high-level Python abstractions. Keras is a high-level library wrapper that is built on top of frameworks like TensorFlow and provides off-the-shelf but often inflexible models. PyTorch utilizes dynamic computation graphs that is modifiable at run-time but the "Pythonic" nature makes it less efficient for benchmarking purposes. After carefully comparing different frameworks [43], [55], we choose TensorFlow in our benchmark evaluation for the following reasons:

- TensorFlow is the most popular open-source deep learning framework so far, with a large and active community supported by Google for quick updates and frequent releases.
- TensorFlow is efficient, user-friendly and easy-to-debug (with TensorBoard) regarding the numerical computations for both research and deployment.

⁴OpenAI: <https://openai.com/blog/ai-and-compute/>

⁵Deep learning frameworks: <https://github.com/topics/deep-learning>

- TensorFlow supports various systems with high performance and scalability.

b) AutoML Frameworks: Various work has been done to develop user-friendly AutoML frameworks [53], [61] including Neural Network Intelligence (NNI), Tree-based Pipeline Optimization Tool (TPOT) and auto-sklearn. NNI⁶ is a popular open-source toolkit that automates the DL model design process. One key feature is the rich collection of algorithms to generate neural architectures and optimizing hyperparameters, as well as a simple interface for more user-defined algorithms. Other frameworks focus on the AutoML pipeline optimization, especially data pre-processing and feature engineering, which is irrelevant for benchmarking the computing power. Therefore we choose to build our own benchmark suite on top of NNI.

IV. METHODOLOGY

A. Neural Architecture Search

Notable successes of neural architecture designs [24], [27], [29], [30], [50], [51] in the past few years have drawn enormous attention in AI research community. The manual design of neural architecture requires tremendous human effort, sometimes even domain knowledge, in an ad-hoc fashion. In contrast, the architectures are automatically generated by selecting and combining primary operations (e.g. convolution) with NAS approaches which can be categorized into three abstraction levels [17]: search space, search strategy and performance estimation strategy. The major search strategies (algorithms) [25] include random search [9], reinforcement learning [62], evolutionary [48], Bayesian optimization [42] and gradient-based method [40]. Research around NAS is typically exploring three dimensions of abstractions simultaneously using various algorithms to search for different combinations of building blocks. In the spirit of transfer learning and knowledge inheritance, [12] proposed network transformation that transforms a pre-trained parent network to a more complex child network while preserving the input and output consistency. The knowledge represented by the neural architecture is transformed from the parent network to the child network. [56] first dubbed "network morphism" that can perform multiple transformation operations including width, depth, kernel size and skip operation. [37] proposed an open-source framework (Auto-Keras), which is part of NNI, to perform network morphing guided by Bayesian optimization. Though every method has its own advantages, we choose the implementation in [56] as our baseline for developing the benchmark. We choose residual network [24] (ResNet-50) as the initial model since ResNet-50 is one of the de facto showcase models in the current DL community. We modify the morphism so that each transformation step adds a block (convolutional layer, batch normalization [35] and activation function all together) instead of just one layer. In addition, We adapt this implementation to suit benchmarking in a parallel and distributed way which is explained later.

⁶<https://www.microsoft.com/en-us/research/project/neural-network-intelligence/>

B. Hyperparameter Optimization

HPO problems can be viewed as the identification of optimal model configurations of all related hyperparameters. Similar to NAS, HPO has three abstractions [33]: search space, search approach and evaluation method. Various search approaches can be applied to select the best hyperparameter combinations including grid search [38], random search [8], Bayesian optimization [10] and heuristic search like evolutionary [49] etc. In our case, the search space is defined by the hyperparameters that are more directly related to the computational cost including the batch size and kernel size to reduce the randomness for benchmarking purposes. We use the stochastic gradient descent (SGD) with momentum [47] as the optimizer since it requires less memory and be more efficient. We evaluated the different optimization approaches and then compare the validation accuracy on the test dataset. The results of multiple experiments on CIFAR10 show that Bayesian optimization (TPE) slightly outperforms other methods. Similar to NAS, we use this fixed algorithm to optimize the batch size and kernel size simultaneously. In our benchmark workflow, the HPO is performed separately after the NAS process on each worker.

C. Workflow

As mentioned, we choose NNI (V1.5) as a baseline to adapt to our benchmark suite. The original NNI framework is implemented with a "master-slave" architecture and performs the NAS and HPO on the master server, which is the bottleneck on large clusters. Also, not all operations in AutoML run on AI accelerators, like model generation and data movement. Consequently, the AI accelerator idles because of the potential bottleneck on CPU or disk I/O. In addition, the model generation is time-consuming and can be implemented with thread parallelism on CPUs. To address these problems and fully appreciate all computing resources in a balanced way, we need to effectively distribute the computations and use proper parallelism [46] on both CPU and AI accelerator. Therefore, we modify the NNI framework in various aspects as shown in Figure 3, including performing the model generation and training on slave nodes asynchronously, utilize slave node CPUs parallelly to generate new architectures and perform training parallelly with all available AI accelerators on each slave node. we utilize data parallelism with synchronous strategy and all-reduce so that all AI accelerators can train on different partitions of data and results in individual gradients, which are then aggregated all-together at each step. We summarize our benchmark workflow as follows:

- User accesses the master node through Secure Shell (SSH), collects information about slave nodes and creates a SLURM configuration script.
- Master dispatches workloads with SLURM to slave nodes corresponding to the requested and available resources, parallelly and asynchronously.
- The slave nodes receive the workloads and perform architecture searching and model training parallelly.

- The CPUs on slave nodes search for new architectures based on the current historical model list, which contains detailed model information and accuracy on the test dataset, then store the architecture in the buffer (e.g. network file system) for later training.
- The AI accelerators on slave nodes load the candidate architecture and data, utilize data parallelism to train along with HPO and then store the results in the historical model list.
- The running terminates once the condition is satisfied (e.g. reaching user-defined time). The final results are calculated based on the recorded metrics and then reported.

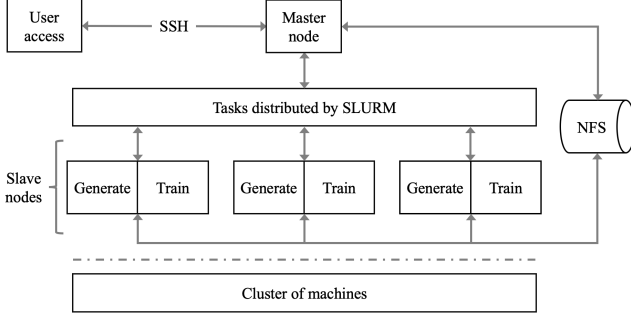


Fig. 3. Schematic diagram of the benchmark workflow. The details are explained in the main text.

D. Measurement

FLOPS is the most cited performance metric to reflect the overall computing ability of HPC⁷ as yet. Our benchmark utilizes FLOPS as the major metric (score) to directly describe the computing power of AI accelerators. Since the processing time can be easily recorded, we only need to measure the total floating-point operations (FLOPs). We require all computation to be conducted with floating points of at least FP-16 precision, therefore we use FLOPs and operation count interchangeably in this paper. Toolkit like NVIDIA profiling tools (nvprof⁸) can record the executed operation count by kernel replay, which is exceptionally slow. This method is also limited to NVIDIA hardware and is not suitable for various platforms. Inspired by LINPACK, we treat the operation counting as a mathematical problem to calculate the operation needed to finish the complex computation in the training and validation processes without any optimization. For a given dataset and model with specific hyperparameters, the theoretical operation needed to finish the training or validation is predetermined. If the hardware or software has any special optimization, the operation count is reduced or the execution is faster, therefore higher FLOPS eventually.

To calculate the operation count analytically, we need to understand the training and validation process. DL libraries like TensorFlow use computational graphs to represent the

computations and guide the workflow. A computational graph is a directed acyclic graph where nodes represent variables or operations and edges represent function arguments (data dependency). Each computation is essentially a node so that variables feed values into operations and operations feed the outputs into other operations. Computational graphs can compose complex models with simple functions and enable automatic differentiation to train the neural networks. Backpropagation [26] is a reverse mode automatic differentiation [6] which applies the chain rule efficiently and recursively to compute gradients of inputs and parameters and other intermediates along with computational graphs. As shown in Algorithm 1, backpropagation has two parts: forward pass (FP) that compute results of operations and save intermediate values needed for gradients computation in memory and backward pass (BP) that apply the chain rule to compute the gradients of the loss function with respect to the inputs (multiply Jacobian matrices by gradients).

Algorithm 1 Backpropagation [26]

Forward Pass:

1. Define the computational graph where each node represent a variable (parameters and intermediates).
2. Visit each node in topological order to compute the variables with corresponding operations and store the values at the nodes.

Backward Pass:

3. Initialize the loss gradients $\frac{dL}{dy}$ and all local partial derivatives $\frac{dy}{dx_i}$.
 4. Visit each node in reverse topological order to compute the loss gradients w.r.t. local variables with chain rule: $\frac{dL}{dx_i} = \frac{dL}{dy} \times \frac{dy}{dx_i}$.
- Return: $\frac{dL}{dx_i}$ for all variables.

The total operation count is the sum of that in FP and BP, which includes operations to calculate the gradients and the operations to update the parameters with gradient descent. Most computations in neural networks are matrices multiplication, which is dot products $y = w[0] \times x[0] + w[1] \times x[1] + \dots + w[n-1] \times x[n-1]$ that has n multiply-accumulate (MACC) and corresponding to roughly $2n$ operations. The gradient descent procedure can be described as repeat $\theta_i := \theta_i + \alpha \frac{dL}{d\theta_i}$ until convergence, so the operation needed is equivalent to one MACC for one parameter in one BP. We break down the original and morphed models into several components (layers) and analytically compute the operation count needed of each layer in the FP, as listed in Table I. The detailed descriptions of each layer are in [24], [35].

The analytical analysis of operation computing is more complicated in the BP process. The convolution in FP can be described as $O_{ij} = \sum_{m=1}^{k-1} \sum_{n=1}^{k-1} X(i-m, j-n) F(m, n)$, where O_{ij} is the output, $X(i-m, j-n)$ is the input and $F(m, n)$ is the filter (kernel). The partial derivatives of local parameters ($\frac{\partial O}{\partial F}$) and local intermediates ($\frac{\partial O}{\partial X}$) can be easily derived and

⁷Top500 Project: <https://www.top500.org/>

⁸CUDA Toolkit Documentation: <https://docs.nvidia.com/cuda/profiler-users-guide/index.html>

TABLE I

THE ANALYTICAL OPERATION COUNTS OF EACH LAYER (PER IMAGE) IN THE FP. FOR CONVOLUTIONAL LAYER, THE INPUT IMAGE DIMENSION IS $H_i \times W_i \times C_i$, THE OUTPUT DIMENSION IS $H_o \times W_o \times C_o$ AND THE KERNEL (FILTER) SIZE IS $K \times K$. FOR DENSE LAYER, THE INPUT IS C_i AND OUTPUT IS C_o . FOLLOWING THE CONVENTION IN [32], THE OPERATION WEIGHT OF MACC IS 2, THE WEIGHT OF ADD/SUBTRACT/MULTIPLY/COMPARISON IS 1, THE WEIGHT OF DIVIDE/SQRT IS 4 AND THE WEIGHT OF SPECIAL OPERATION LIKE EXPONENTIAL IS 8. THE OPERATION IS ONLY AN APPROXIMATION.

Layer	Operation in the FP
Convolutional layer	$MACC = K \times K \times C_i \times H_o \times W_o \times C_o$
Dense layer	$MACC = C_i \times C_o$
Batch normalization	$MACC = Add = Div = H_i \times W_i \times C_i$
ReLU	$Comparison = H_o \times W_o \times C_o$
Add layer	$Add = H_o \times W_o \times C_o$
Max-pooling layer	$Comparison = K \times K \times H_o \times W_o \times C_o$
Global-pooling layer	$Add = H_i \times W_i \times C_i; Div = C_i$
Softmax layer	$Exp = Add = Div = C_o$

are used in gradient calculation. Applying the chain rule, we have the parameters' gradients and the intermediates' gradients by multiplying the loss gradients with the local gradients as Equation 1. By substituting the derivatives ($\frac{\partial O}{\partial F}$ and $\frac{\partial O}{\partial X}$), we can express the backpropagation as Equation 2. Therefore, the total operation needed to calculate all gradients is roughly twice as that in FP. The total parameter is convolution layer (without bias) is $K \times K \times C_i \times C_o$, so the operation needed to update all parameters with gradient descent method is $2 \times K \times K \times C_i \times C_o$. Consider all steps we can have the total operation in BP shown in Table II. Since the K , C_i and C_o are typically small values in convolutional layers, the total operations in BP in roughly twice that of FP.

$$\begin{cases} \frac{\partial L}{\partial F_i} = \sum_{k=1}^m \frac{\partial L}{\partial O_k} \times \frac{\partial O_k}{\partial F_i} \\ \frac{\partial L}{\partial X_i} = \sum_{k=1}^m \frac{\partial L}{\partial O_k} \times \frac{\partial O_k}{\partial X_i} \end{cases} \quad (1)$$

$$\begin{cases} \frac{\partial L}{\partial F} = Convolution (Input X, Loss gradient \frac{\partial L}{\partial O}) \\ \frac{\partial L}{\partial X} = Full Convolution (Flipped F, Loss gradient \frac{\partial L}{\partial O}) \end{cases} \quad (2)$$

For dense layer $Y = W^T X + B$, the intermediates' gradients can be obtained by multiplying the loss gradients ($\frac{\partial L}{\partial Y}$) with the Jacobian matrices of intermediates ($\frac{\partial Y}{\partial X}$). Similarly, the weights' gradients is $\frac{\partial L}{\partial W} = \frac{\partial L}{\partial Y} \times \frac{\partial Y}{\partial W}$. In both cases, the operation needed is the same as that in FP. The bias gradient is $\frac{\partial L}{\partial B} = \frac{\partial L}{\partial Y} \times \frac{\partial Y}{\partial B} = \frac{\partial L}{\partial Y}$ since $\frac{\partial Y}{\partial B} = 1$, therefore resulting no extra operation. The total parameter in a dense layer (with bias) is $(C_i + 1) \times C_o$ and the total operation needed in the BP of the dense layer is shown in Table II. Unlike the convolutional layer, the operation of the dense layer in BP is more than tripled of that in FP. The operation in BP of the rest layers, including batch normalization, activation function (ReLU), element-wise add layer, max-pooling, global-pooling and softmax layer are all ignorable for practical purposes. We confirmed our analytical method by comparing

the results of ResNet-50 on ImageNet [15] with TensorFlow profiler⁹ (only computing operation in the FP) and NVIDIA profiling tools (compute operation in both FP and BP). In our analytical method, we do not consider any hardware or software optimization that would result in such an effect. The operation count from this analytical approach is only related to the neural architecture, hyperparameters configuration and data (like image resolution). The optimizations that result in less operation will speed up the training or validation processes therefore higher final FLOPS. The details of other verification of our FLOPS measure approach are elaborated in Appendix VII.

TABLE II

THE ANALYTICAL OPERATION COUNTS OF EACH LAYER (PER IMAGE) IN THE BP. THE MEANINGS OF SYMBOLS ARE THE SAME AS IN TABLE I. THE TOTAL OPERATION NEEDED FOR CALCULATING THE GRADIENTS AND FOR UPDATING PARAMETERS ARE SUMMED.

Layer	Operation in the BP
Convolutional layer	$MACC = 2 \times (K \times K \times C_i \times H_o \times W_o \times C_o) + (K \times K \times C_i \times C_o)$
Dense layer	$MACC = 2 \times C_i \times C_o + (C_i + 1) \times C_o$

TABLE III

THE ANALYTICAL OPERATION COUNTS OF EACH LAYER (PER IMAGE) IN FP AND BP TOGETHER. MOST OPERATIONS IN RESNET-50 HAPPEN IN CONVOLUTIONAL LAYER.

Layer	FP	BP	BP/FP	Total
Convolutional	7.71E09	1.52E10	1.9755	2.29E10
Dense	4.10E06	1.23E07	3.0005	1.64E07
Batch normalization	7.41E07	1.91E03	0.00003	7.41E07
ReLU	9.08E06	0	0	9.08E06
Max-pooling	1.81E06	0	0	1.81E06
Average-pooling	1.00E05	0	0	1.00E05
Add	5.52E06	0	0	5.52E06
Softmax	2.10E04	0	0	2.10E04
Total	7.81E09	1.52E10	1.9531	2.31E10

Measuring AI-HPC quantitatively is not trivial due to the diversity, sometimes even conflicting, of workloads and metrics. One single metric like FLOPS alone may not be sufficient to reflect the AI-HPC computation capabilities considering both the hardware and software. For example, the data parallelism algorithm that is frequently applied in distributed machine learning will speed up the whole process at a cost of lower average AI accelerator utilization and FLOPS. While one can present all relevant metrics separately, we intend to provide a metric to informatively characterize the system's overall performance. In general, an efficient AI-HPC would perform more computation and result in higher accuracy in less time. The empirical results [20] show that the accuracy on the validation dataset increases monotonically and then plateaus over time. In other words, the error (1 - accuracy) decreases slower and slower over time. We would like to compensate for this effect with an increasing changing rate of the metric. Therefore the absolute value of the partial derivative of the metric with respect to the error should increase with decreasing error.

⁹<https://www.tensorflow.org/guide/profiler>

On the other hand, the partial derivative of the metric with respect to FLOPS should be independent of FLOPS to make the computation contribute to the metric uniformly. We use this metric as a regulated score in our benchmark, besides the FLOPS, to quantitatively measure the cross-stack performance of an AI-HPC. According to the above conditions, we design our regulated score as Equation 3:

$$\text{Regulated Score} = -\ln(\text{Error}) \times \text{FLOPS} \quad (3)$$

where $\text{Error} \in (0,1)$ and the negative sign keeps the $\ln(\text{Error})$ be positive. Consequently, the regulated score increases faster with lower error and increases linearly with FLOPS.

E. Fixed and Customizable Configuration

There are several rules in our benchmark for a fair comparison across various platforms. With a "pencil-and-paper" manner [5], our benchmark also has customizable configurations that allow users to optimize the performance. First, the benchmark should run on a "master-slave" architecture. The master is deployed on a strong server without any AI accelerator to dispatch tasks and collect all results from the slave nodes. The slave node is composed of one or multiple servers equipped with AI accelerator(s) and can be deployed with or without a container environment. Both scale-up (multiple AI accelerators on each slave node) and scale-out (one AI accelerator on each slave node) configurations are supported. Second, the algorithms and search space used for AutoML are fixed, i.e. network morphism for NAS and Bayesian optimization for HPO, with aforementioned operations and hyperparameters. The HPO only starts at the fourth round of training on each slave node since the earlier rounds are trained insufficiently, which is also referred to as the warm-up process in this paper. A predicted accuracy, instead of the actual one, is used in the warm-up process. There is also a maximum limit on epoch and patience, which is the number of epochs to wait before early stop if no progress on the validation dataset. Third, the dataset is fixed to be ImageNet which has 1281167 and 50000 224*224 RGB images for training and validation, respectively. We keep the back-end DL framework and most hyperparameters open to further optimization. This would partially relieve the performance dependency on manual designs and be more independent of the software part of the system. The data can be formatted in different ways corresponding to the DL framework. For example, the data loading with TFRecord is more efficient for TensorFlow. Forth, our benchmark requires the minimum running time to be 1 hour, the minimum precision to be FP-16 and the maximum error to be 30%. The average value of FLOPS during the valid time period (from 1 hour to the ending time) is considered as the final score and the user can terminate the running once the result is converged. The summarized configurations are shown in Table IV.

TABLE IV
FIXED AND CUSTOMIZABLE CONFIGURATIONS. THE CUSTOMIZABLE SETUPS ARE PREDETERMINED EITHER EMPIRICALLY OR EXPERIMENTALLY WITH DEFAULT VALUES.

Configuration	Fixed and customizable setups
Server arrangement	Fixed: master-slave
NAS method	Fixed: network morphism
HPO method	Fixed: Bayesian optimization
Dataset	Fixed: ImageNet
DL Framework	Default: TensorFlow
Initial architecture	Fixed: ResNet-50
Initial weight	Default: method in [23]
Batch size	Default: 448 (on NVIDIA V100 32 GB)
Optimizer	Default: gradient descent with momentum
Learning rate	Default: 0.1 with decay=0.1/60
Loss function	Default: categorical cross entropy
Maximum epoch	Default: 60
Parallelism	Default: synchronous all-reduce
Minimum precision	Fixed: FP-16
Maximum error	Fixed: 30%
Minimum running time	Fixed: 1 hour

V. EVALUATION

A. Setup

In our preliminary test, We verified our benchmark design (regarding algorithm and implementation) on our local machine with 4 NVIDIA 1080Ti based on CIFAR10 dataset. For the formal evaluation presented here, we perform it on a large cluster consisting of multiple servers each with 2 CPUs and 8 AI accelerators (see Table V for hardware specifications). As a modern practice in AI research, we perform the evaluation in containers with the allocated resources and pre-assigned services for the consistency of the testing environment. we utilize Kubernetes to deploy the docker containers that wrap in all the dependencies including the operating system, libraries and workload codes, etc. to provide the running environment. We use each physical server with the same hardware specifications as either a master or a slave node for simplicity. The detailed information of the evaluation environment is shown in Table VI.

TABLE V
HARDWARE SPECIFICATIONS.

Components	Specifications
Processor	Intel skylake 6151 (18 cores)
AI Accelerator	NVIDIA Tesla NVLink V100 (32 GB)
Memory	2667MHz DDR4 512 GB
Storage	NVMe 5 TB
Ethernet network	InfiniBand 100 Gb/s

TABLE VI
EVALUATION ENVIRONMENT SPECIFICATIONS.

Allocated resources	30 CPU cores, 128 GB memory, 8 NVIDIA V100
Environment	Ubuntu 16.04, docker 18.09.9, SLURM 15.08.7 TensorFlow V2.2, CUDA V10.1, Python 3.5

B. Performance

In this paper, we limit our evaluation of two major characteristics of the benchmark: stability and scalability. We run

the benchmark on various scales of machines from 10 to 50 slave nodes with up to 400 GPUs. All the intermediate results including the generated architectures, hyperparameter configurations, accuracy at each epoch and timestamps are recorded in log files. Once the benchmarking process is finished, we run the data analysis toolkit to calculate the score along with other complementary results utilizing all the recorded information and then create a report. Within 12 hours on various scales of AI accelerators, the scores are converged and increase steadily as shown in Figure 4. The highest scores (see labels) scales linearly with the number of AI accelerators and the scaling factor is almost the same as that of the machines at different scales. The achievable error of the automatically generated models is limited by the actual running time in our evaluation. Previous AutoML research like [62] searched 12800 neural architectures before finding the optimal one, comparing to roughly 200 architectures in our largest scale of evaluation with 50 nodes in 12 hours. The potential of AutoML can only be appreciated by longer searching time with more machines as it is tremendously computing greedy. The regulated scores in Figure 4 also converge since it is essentially just FLOPS multiplied with the model performance as a coefficient. For AI systems at the same machine scale but with different software optimizations, the regulated score can reflect the hardware and software co-performance, therefore we also provide it as a complementary result.

To ensure the stability, we monitor the GPU performance during the benchmarking process. We use NVIDIA System Management Interface (nvidia-smi¹⁰) to track the GPU utilization to show the percentage of time during which one or more kernels are occupied, along with the GPU memory utilization during the same time period. We developed a toolkit to extract real-time information with 30 seconds sampling interval during the entire running time. As shown in Figure 5, the GPU utilization and memory occupancy are both high while the training phase with the default benchmark configuration (for NVIDIA V100).

VI. CONCLUSION

The rise of the convergence of AI and HPC reveals new challenges in benchmarking the state-of-the-art and future large scale clusters for AI purposes. We review the current HPC and AI benchmarks and qualitatively analyze why they do not address all the challenges. We choose AutoML, a highly scalable and representative AI application, as our benchmark workload and implement the algorithms in a highly parallel manner. We also propose an analytical approach that is independent of DL frameworks and other software implementations to estimate the FLOPS during the training and validation processes. We use average FLOPS as the benchmark score to quantitatively measure the machine computing power on AI applications. We evaluate the benchmark on different scales of AI accelerators with a large dataset and verify the benchmark’s linear scalability and stability. The scalability and

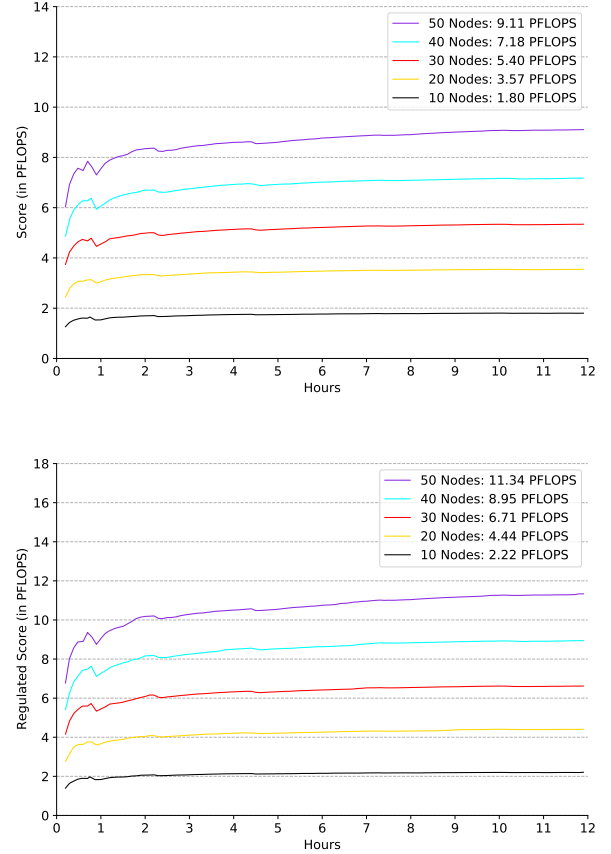


Fig. 4. The benchmark scores and the regulated scores (both in PFLOPS) over time of evaluations with different scales of machines. The FLOPS is converged and increase steadily after the initial warm-up phase and we report the average value of FLOPS during the valid time period (from 1 hour to the ending time) as the final benchmark score, shown in the labels. The results show the robustness and linear scalability of our benchmark. The drops at around 1 hour and other timestamps come from the low GPU utilization during the inter-phase between the training stages. The regulated scores have similar behaviors as the scores.

simple metric design allow us to compare machines from small clusters to large AI-HPC easily.

ACKNOWLEDGMENT

We thank Peng Cheng Laboratory for providing the evaluation resources. We also appreciate the valuable advice from Prof. Yong Dou and the technical support from Jun Long, Hongji Chen, Haiyu Fu and Ya Ouyang.

m

VII. APPENDIX: FLOPS CALCULATION

We compare our analytical approach of FLOPs computing with the TensorFlow profiler (tf.profiler¹¹) and NVIDIA profiling tool (nvprof). The tf.profiler can only count operations in the FP. The nvprof can trace the GPU activity and use the kernel replay to ensure all requested profile data including

¹¹https://www.tensorflow.org/api_docs/python/tf/compat/v1/profiler/ProfileOptionBuilder

¹⁰<https://developer.nvidia.com/nvidia-system-management-interface>

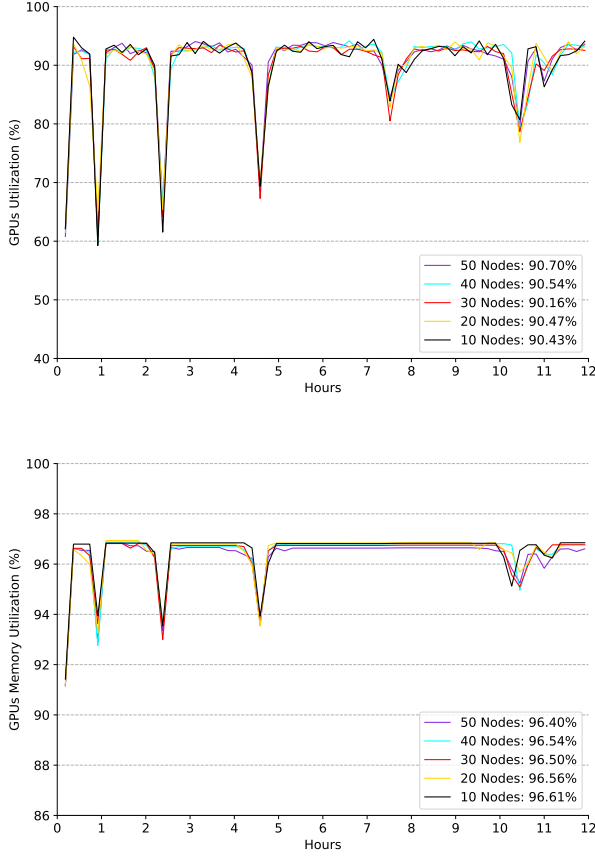


Fig. 5. The GPUs and their memory utilization of evaluations with different scales of machines. The average values are shown in the labels. The utilization drops during the inter-phase between the training stages due to the data loading and computational graph compilation etc.

operation counts of adds, multiplies, multiply-accumulates, and special operations. The profiling process with nvprof is prohibitively expensive therefore we need an approach to speed-up the process.

TABLE VII

THE COMPARISON OF OPERATION COUNTS OF EACH LAYER OF RESNET-50 ON IMAGENET (PER EPOCH, BATCH SIZE=1) WITH DIFFERENT APPROACHES. THE DIFFERENCE OF FPs IN TRAINING AND VALIDATION STAGE COMES FROM THE DATA SIZE.

Procedure	tf.profiler	nvprof	analytical
FP (training)	9.97E15	1.02E16	1.00E16
BP (training)	-	2.10E16	1.95E16
BP / FP (training)	-	2.0603	1.9533
Total (training)	-	3.12E16	2.95E16
FP (Validation)	3.89E14	3.98E14	3.90E14
Total (training + validation)	-	3.16E16	2.99E16

Fortunately, we can utilize the iterative nature of DL computation and sample the profiling process based on a small partition of data. This is only an approximation since the FLOPs varies with the hyperparameter configurations.

Table VII shows the FLOPs of ResNet-50 layers on ImageNet with the 3 approaches. The FLOPs in FP are consistent between our analytical approach and tf.profiler and the FLOPs in BP are consistent among all three approaches.

REFERENCES

- [1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.
- [2] Robert Adolf, Saketh Rama, Brandon Reagen, Gu-Yeon Wei, and David Brooks. Fathom: Reference workloads for modern deep learning methods. In *2016 IEEE International Symposium on Workload Characterization (IISWC)*, pages 1–10. IEEE, 2016.
- [3] AIIA. Dnn benchmark. <http://www.aiia.org.cn/benchmark/zh/index.html>.
- [4] Baidu. Deepbench. <https://github.com/baidu-research/DeepBench>.
- [5] David H Bailey, Eric Barszcz, John T Barton, David S Browning, Robert L Carter, Leonardo Dagum, Rod A Fatoohi, Paul O Frederickson, Thomas A Lasinski, Rob S Schreiber, et al. The nas parallel benchmarks. *The International Journal of Supercomputing Applications*, 5(3):63–73, 1991.
- [6] Atılım Günes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *The Journal of Machine Learning Research*, 18(1):5595–5637, 2017.
- [7] Tal Ben-Nun, Maciej Besta, Simon Huber, Alexandros Nikolaos Ziogas, Daniel Peter, and Torsten Hoefer. A modular benchmarking infrastructure for high-performance and reproducible deep learning. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 66–77. IEEE, 2019.
- [8] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(Feb):281–305, 2012.
- [9] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(Feb):281–305, 2012.
- [10] James S Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in neural information processing systems*, pages 2546–2554, 2011.
- [11] Erin Carson and Nicholas J Higham. Accelerating the solution of linear systems by iterative refinement in three precisions. *SIAM Journal on Scientific Computing*, 40(2):A817–A847, 2018.
- [12] Tianqi Chen, Ian Goodfellow, and Jonathon Shlens. Net2net: Accelerating learning via knowledge transfer. *arXiv preprint arXiv:1511.05641*, 2015.
- [13] François Chollet et al. Keras. <https://keras.io>, 2015.
- [14] Cody Coleman, Deepak Narayanan, Daniel Kang, Tian Zhao, Jian Zhang, Luigi Nardi, Peter Bailis, Kunle Olukotun, Chris Ré, and Matei Zaharia. Dawnbench: An end-to-end deep learning benchmark and competition. *Training*, 100(101):102, 2017.
- [15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [16] Shi Dong and David Kaeli. Dnnmark: A deep neural network benchmark suite for gpus. In *Proceedings of the General Purpose GPUs*, pages 63–72. 2017.
- [17] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *arXiv preprint arXiv:1808.05377*, 2018.
- [18] Geoffrey C Fox. Perspectives on high-performance computing in a big data world. In *Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing*, pages 145–145, 2019.
- [19] Wanling Gao, Fei Tang, Lei Wang, Jianfeng Zhan, Chunxin Lan, Chunjie Luo, Yunyou Huang, Chen Zheng, Jiahui Dai, Zheng Cao, et al. Aibench: an industry standard internet service ai benchmark suite. *arXiv preprint arXiv:1908.08998*, 2019.
- [20] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
- [21] J Gustafson et al. Slalom: The first scalable supercomputer benchmark, supercomputing reo.(nov. 1990). *Parallel performance of applications on supercomputers*, 1273.

- [22] John L Gustafson and Quinn O Snell. Hint: A new way to measure computer performance. In *Proceedings of the Twenty-Eighth Annual Hawaii International Conference on System Sciences*, volume 2, pages 392–401. IEEE, 1995.
- [23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [25] Xin He, Kaiyong Zhao, and Xiaowen Chu. Automl: A survey of the state-of-the-art. *arXiv preprint arXiv:1908.00709*, 2019.
- [26] Robert Hecht-Nielsen. Theory of the backpropagation neural network. In *Neural networks for perception*, pages 65–93. Elsevier, 1992.
- [27] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [28] HPE. Deep learning benchmarking suite. <https://github.com/HewlettPackard/dlcookbook-dlbs>.
- [29] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.
- [30] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [31] EA Huerta, Asad Khan, Edward Davis, Colleen Bushell, William D Gropp, Daniel S Katz, Volodymyr Kindratenko, Seid Koric, William TC Kramer, Brendan McGinty, et al. Convergence of artificial intelligence and high performance computing on nsf-supported cyberinfrastructure. *arXiv preprint arXiv:2003.08394*, 2020.
- [32] Janice E Huss and James A Pennline. A comparison of five benchmarks. 1987.
- [33] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. *Automated Machine Learning*. Springer, 2019.
- [34] Andrey Ignatov, Radu Timofte, William Chou, Ke Wang, Max Wu, Tim Hartley, and Luc Van Gool. Ai benchmark: Running deep neural networks on android smartphones. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.
- [35] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- [36] Zihan Jiang, Lei Wang, Xingwang Xiong, Wanling Gao, Chunjie Luo, Fei Tang, Chuanxin Lan, Hongxiao Li, and Jianfeng Zhan. Hpc ai500: The methodology, tools, roofline performance models, and metrics for benchmarking hpc ai systems. *arXiv preprint arXiv:2007.00279*, 2020.
- [37] Haifeng Jin, Qingquan Song, and Xia Hu. Auto-keras: An efficient neural architecture search system. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1946–1956, 2019.
- [38] Hugo Larochelle, Dumitru Erhan, Aaron Courville, James Bergstra, and Yoshua Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th international conference on Machine learning*, pages 473–480, 2007.
- [39] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- [40] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.
- [41] Peter Mattson, Christine Cheng, Cody Coleman, Greg Diamos, Paulius Micikevicius, David Patterson, Hanlin Tang, Gu-Yeon Wei, Peter Bailis, Victor Bittorf, et al. Mlperf training benchmark. *arXiv preprint arXiv:1910.01500*, 2019.
- [42] Hector Mendoza, Aaron Klein, Matthias Feurer, Jost Tobias Springenberg, and Frank Hutter. Towards automatically-tuned neural networks. In *Workshop on Automatic Machine Learning*, pages 58–65, 2016.
- [43] Giang Nguyen, Stefan Dlugolinsky, Martin Bobák, Viet Tran, Álvaro López García, Ignacio Heredia, Peter Malík, and Ladislav Hluchý. Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey. *Artificial Intelligence Review*, 52(1):77–124, 2019.
- [44] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.
- [45] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
- [46] Diego Peteiro-Barral and Bertha Guijarro-Berdiñas. A survey of methods for distributed machine learning. *Progress in Artificial Intelligence*, 2(1):1–11, 2013.
- [47] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.
- [48] Esteban Real, Sherry Moore, Andrew Selle, Saurabh Saxena, Yutaka Leon Suematsu, Jie Tan, Quoc V Le, and Alexey Kurakin. Large-scale evolution of image classifiers. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2902–2911. JMLR. org, 2017.
- [49] Esteban Real, Sherry Moore, Andrew Selle, Saurabh Saxena, Yutaka Leon Suematsu, Jie Tan, Quoc V Le, and Alexey Kurakin. Large-scale evolution of image classifiers. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2902–2911. JMLR. org, 2017.
- [50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [51] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [52] Jin-Hua Tao, Zi-Dong Du, Qi Guo, Hui-Ying Lan, Lei Zhang, Sheng-Yuan Zhou, Ling-Jie Xu, Cong Liu, Hai-Feng Liu, Shan Tang, et al. Bench ip: Benchmarking intelligence processors. *Journal of Computer Science and Technology*, 33(1):1–23, 2018.
- [53] Anh Truong, Austin Walters, Jeremy Goodsitt, Keegan Hines, Bayan Bruss, and Reza Farivar. Towards automated machine learning: Evaluation and comparison of automl approaches and tools. *arXiv preprint arXiv:1908.05557*, 2019.
- [54] Yuxin Wang, Qiang Wang, Shaohuai Shi, Xin He, Zhenheng Tang, Kaiyong Zhao, and Xiaowen Chu. Benchmarking the performance and power of ai accelerators for ai training. *arXiv preprint arXiv:1909.06842*, 2019.
- [55] Zhaobin Wang, Ke Liu, Jian Li, Ying Zhu, and Yaonan Zhang. Various frameworks and libraries of machine learning and deep learning: a survey. *Archives of Computational Methods in Engineering*, pages 1–24, 2019.
- [56] Tao Wei, Changhu Wang, Yong Rui, and Chang Wen Chen. Network morphism. In *International Conference on Machine Learning*, pages 564–572, 2016.
- [57] D E Womble, Mallikarjun Shankar, Wayne Joubert, J T Johnston, J C Wells, and J A Nichols. Early experiences on summit: Data analytics and ai applications. *Journal of Reproduction and Development*, 63(6), 2019.
- [58] Gangnam Yi and Vincenzo Loia. High-performance computing systems and applications for ai. *The Journal of Supercomputing*, 75(8):4248–4251, 2019.
- [59] Wei Zhang, Wei Wei, Lingjie Xu, Lingling Jin, and Cheng Li. Ai matrix: A deep learning benchmark for alibaba data centers. *arXiv preprint arXiv:1909.10562*, 2019.
- [60] Hongyu Zhu, Mohamed Akrou, Bojian Zheng, Andrew Pelegrinis, Amar Phanishayee, Bianca Schroeder, and Gennady Pekhimenko. Tbd: Benchmarking and analyzing deep neural network training. *arXiv preprint arXiv:1803.06905*, 2018.
- [61] Marc-André Zöllner and Marco F Huber. Survey on automated machine learning. *arXiv preprint arXiv:1904.12054*, 9, 2019.
- [62] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.