Open Access

Journal of Artificial Intelligence & Robotics

Article type: Review Article

HPC-AI benchmarks - A comparative overview of high-performance computing hardware and AI benchmarks across domains

Narges Lux[1#]; Chirag Mandal[1#]; Jonathan Decker[2]; Jannik Luboeinski[3,4]; Julian Drewljau[5]; Tino Meisel[1]; Christian Tetzlaff [3,4]; Christian Boehme[1]; Julian Kunkel[1,2]*

[1]Society for Scientific Data Processing mbH Göttingen (GWDG), Göttingen, Germany.

[2]Georg-August-Universitat Göttingen, Institute for Computer Science, Göttingen, Germany.

[3]University Medical Center Göttingen, Neuro- and Sensory Physiology, Göttingen, Germany.

[4]Campus Institute Data Science (CIDAS), Göttingen, Germany.

[5]Leibniz University Hannover, Institute for Microelectronic Systems, Hannover, Germany.

[#]First Authors.

Corresponding Author: Julian Kunkel

Society for Scientific Data Processing mbH Göttingen (GWDG), Göttingen, Germany.

Email: julian.kunkel@gwdg.de

Received: Oct 21, 2024

Accepted: Dec 23, 2024

Published Online: Dec 30, 2024

Website: www.joaiar.org

Volume 2 [2025] Issue 1

Abstract

In the rapidly evolving fields of Artificial Intelligence (AI) and High-Performance Computing (HPC), benchmarking is a critical tool for optimizing system performance. It guides the selection of hardware architectures, software frameworks, and model configurations, as well as the parameterization of models for specific tasks. Despite considerable progress in HPC-AI benchmarking and hardware optimization for a range of AI algorithms, there is a pressing need for further research. Key areas include the interoperability of AI software and hardware, the standardization of AI benchmarks across various industries, and the comparison of domain-specific studies to fill existing knowledge gaps. To help address these needs, we assess the impact of different hardware architectures on the performance of various AI models and propose to establish a standardized benchmarking methodology for comparing these models across diverse hardware platforms.

Our survey provides a comprehensive summary of pertinent AI benchmarking tools, offering an analysis of their focus on training or inference, use cases, metrics, features, benefits, and limitations, particularly within the medical domain. We also explore hardware benchmarking for systems such as GPUs, neuromorphic devices, and FPGAs. This overview aims to guide researchers and practitioners in making informed choices for their AI applications, thereby contributing significantly to the field of HPC-AI benchmarking.

Keywords: Benchmarking; Domain exploration; Medical field; HPC-AI; Interoperability; Standardization.

Citation: Kunkel J, Lux N, Mandal C, Decker J, Luboeinski J, et al. HPC-AI benchmarks - A comparative overview of high-performance computing hardware and AI benchmarks across domains. J Artif Intell Robot. 2024; 1(2): 1017.

Introduction

Background: Artificial Intelligence (AI) is a rapidly growing field that is revolutionizing many aspects of our lives via technologies such as autonomous vehicles and personalized recommendations [1]. High-Performance Computing (HPC), which involves the use of supercomputers and computer clusters to deliver high performance, plays a crucial role in the advancement of AI [2]. In the context of AI, HPC is used to train complex machine learning models, process large volumes of data, and perform complex simulations [3]. This combination of HPC and AI, also called HPC-AI, is leading to exciting advancements in various fields, including healthcare, climate modeling, and autonomous vehicles [4]. However, the realm of HPC-AI presents several challenges, such as the management of vast computational resources, the design of an efficient data supply subsystem [5], and the complex, volatile, and unpredictable dynamics of AI training. Even minor alterations in models, hyperparameters, or optimization strategies can significantly impact the final accuracy or the rate of training convergence [6]. Additionally, the exorbitant cost associated with training a cutting-edge AI model indicates that a thorough benchmarking of all system components is required. When it comes to AI benchmarking, a plethora of benchmarks are available. However, identifying the most pertinent ones for a specific use case can be a daunting task [7]. Benchmarks often fall short in accurately capturing the true capabilities and limitations of AI systems, leading to potential misconceptions about their safety and reliability [8]. Furthermore, the rate at which benchmark saturation is achieved is escalating [9,10]. Taking into account all the points mentioned above, it becomes evident that the compatibility between AI software and hardware, the uniformity of AI benchmarks across diverse sectors, and the comparative analysis of domain-specific studies are of paramount importance. These aspects are critical for the effective evaluation and optimization of AI systems, and form the motivation for our survey.

Related works: The field of HPC-AI benchmarking has seen significant contributions in recent years [11]. Jiang et al. [11] have underscored the importance of HPC-AI benchmarks that are representative, repeatable, and simple, and proposed a benchmark suite to evaluate HPC systems for scientific deep learning (HPC AI500). Thiyagalingam J [12] introduced [12] the SciMLBench suite of scientific machine learning benchmarks. However, these suites do not fully capture the complexities and variability of real-world scenarios or emerging AI algorithms beyond machine learning, indicating a need for more comprehensive benchmarking tools. In the broader context of AI applications, reports such as “Gen AI use cases by type and industry” [13] and “Notes from the AI Frontier” [14] delve into how AI, including Generative AI, can tackle enterprise challenges and target numerous AI use cases, respectively. On the hardware side, a study by [15] offers a comparative analysis of various hardware accelerators, pointing out the need for more research on hardware optimization for diverse AI algorithms and use cases. In this context Shehzad et al. introduces a scalable System-on-Chip (SoC) solution for accelerating Deep Neural Networks (DNNs). While it improves computational time and power consumption, it lacks a comprehensive comparison with other hardware solutions and a detailed discussion on accuracy versus efficiency trade-offs [16]. Despite these extensive studies, there are knowledge gaps with respect to the interoperability of different AI software and hardware, the comprehensive comparison of domain-specific studies, and the standardization of AI benchmarks across various industries.

Organization of the paper: In this survey, we aim to address these gaps by providing a comprehensive overview of the AI benchmarking landscape, structured into two main parts, following our overview of used methods (Section 2). The first part, “Domain-specific cases”, delves into different tasks in AI, each with its own state-of-the-art models, commonly used frameworks, hardware, and specific metrics. This part aims to provide insights into the unique requirements and challenges of different tasks, contributing to a more holistic view of the AI landscape. The second part, “Hardware benchmarking”, brings in hardware aspects by comparing specific hardware architectures, the models used with them, and their efficiency as resulting from bench-marking solutions. Thereby, this part provides a deeper understanding of the impact of different hardware systems on AI performance, concluding its role as a critical factor in AI system design and deployment. The two parts are interconnected and each building upon the previous one to finally provide a comprehensive overview of AI and HPC benchmarking frameworks with their strengths, weaknesses, unique features, and relevance to real-world cases in various domains.

Methods

Research paper selection and survey structure

Benchmarking is a broad topic, encompassing various domains such as AI, HPC-AI, HPC, Edge/Internet of Things (IoT)/Mobile, and Database [17]. While a comprehensive examination of each benchmarking category might yield valuable insights, it would also result in an overwhelming amount of information, some of which may not be directly relevant to our focus. In this review, we have chosen to concentrate on common AI models that typically leverage HPC, thus positioning our work within the realm of HPC-AI benchmarking. This focus allows us to delve deeper into the specific challenges and opportunities within the intersection of AI and HPC.

Given the critical role that hardware plays in HPC performance, our survey also closely examines hardware benchmarking. We investigated different hardware configurations, such as GPUs and specialized accelerators, for their performance in handling large datasets, executing complex algorithms, and their impact on computational accuracy and speed.

We selected the reviewed studies based on their contribution to state-of-the-art HPC-AI, prioritizing articles that have provided innovative insights, substantial advancement, or in-depth analysis of the use of current HPC capability.

Process of benchmark evaluation

To be able to compare the selected benchmarks, we organize our analysis such that we separately consider the specific models, datasets, and metrics that are used. Furthermore, we categorize the considered benchmarks into three main types: time-series-, image-, and text-related (see Figure 1).

Models

The comparative benchmarks that we have selected employ a variety of models to tackle different AI tasks. Categories of these tasks are described in the following.

Time-series-related tasks

Language tasks: Language tasks include speech recognition, translation, and retrieval. Benchmarks have targeted multilingual pre-trained models [18] and transformer-based architectures [19]. Multilingual models are capable of understanding and generating multiple languages, rendering them critical in today's globalized world.

Image-related tasks

Object detection: Models such as ResNet-50, a 50-layer deep convolutional neural network [20], and Faster R-CNN, an object detection model that utilizes a region proposal network (RPN) with the CNN model [21], are frequently benchmarked due to their robust performance. With enhanced speed and accuracy, the most recent YOLO model, YOLOv7 [22], optimizes anchor-free detection and incorporates cutting-edge methods including self-adversarial training and cross-stage partial connections for superior performance in real-time object detection. On the other hand, Retina Net [23] achieves great accuracy and robustness by focusing on objects that are difficult to detect, employing a unique focal loss function to efficiently address class imbalance in single-stage object detection tasks.

Image segmentation: Tasks often employ models like Faster/Mask-RCNN, an extension of the Faster R-CNN object detection algorithm used for both object detection and instance segmentation tasks [24], and SOLOv2, an anchor-free instance segmentation framework that achieves state-of-the-art performance on the COCO dataset [25]. Image classification: Benchmarks often use ImageNet-based models, which are models trained on the ImageNet dataset [26].

Text-related tasks: Natural Language Processing (NLP): Here, transformer-based models like BERT are utilized [27]. BERT, in particular, has appeared multiple times in the benchmarks, indicating its significance in NLP tasks.

Relational reasoning: Models like spiking/nonspiking RelNet can be used to draw logical conclusions about entity relationships [28]. Understanding and reasoning tasks frequently employ models like GPT-2 [29] and GPT-3 [30], which exhibit remarkable capabilities.

The multiple occurrence of certain models in the here considered benchmarks underscores their effectiveness and widespread use in their respective domains. By comparing a diverse set of models, these benchmarks aim to provide a comprehensive evaluation of various AI tasks.

Data

In AI benchmarking, datasets are essential because they provide the basis for assessing and contrasting the performance of different algorithms and models. They provide a standardized testing framework that makes it possible for practitioners and researchers to evaluate AI model's strengths and weaknesses in a variety of contexts and tasks. Some of the notable datasets that are used for performing the different tasks are mentioned below:

Image-related datasets

Prominent datasets including ImageNet, MNIST, CIFAR-10, and COCO are used to test and train different machine learning algorithms, especially in computer vision. ImageNet [31] is a large visual database with over 14 million images that have been tagged and categorized using the WordNet hierarchy. It is mostly used in studies related to visual object recognition. In contrast, MNIST [32] is a well-known dataset that was created especially for testing and training image processing algorithms. It consists of a significant collection of handwritten digits. CIFAR-10 [33] is a popular choice for training image recognition algorithms since it contains 60,000 32x32 color images divided into 10 classes. Common Objects in Context, or COCO [34], is a large dataset that aims to detect, segment, and caption objects in common scenes. It contains over 330,000 images, of which over 200,000 have object segmentations annotated on them, and 1.5 million object instances. As such, it is an essential tool for creating and evaluating object detection and image captioning algorithms.

Time-series & text-related datasets

Three important datasets - CoVoST2 [35], VATEX [36], and GLUE [37] - are intended to advance research in multilingual speech translation, multimodal video description, and natural language understanding, in that sequence.

CoVoST2 [35], which covers 2,880 hours of speech data with contributions from 78,000 people, builds on the original CoVoST by adding translations from 21 languages into English and from English into 15 languages. This large dataset promotes research in enormous multilingual speech-to-text translation, particularly for low-resource language pairs. VATEX [36] is a large multilingual video description dataset with more than 41,250 videos that have captions in both English and Chinese. It is an essential tool for developing multimodal research in single-language and bilingual contexts since it makes a variety of tasks easier, such as bilingual translation, cross-lingual vision-language processing, and video interpretation.

A standard for evaluating models' general language comprehension abilities is provided by GLUE (General Language Understanding Evaluation) [37], a benchmark made up of a collection of datasets that are used to measure model performance on a range of natural language understanding tasks.

Metrics

Evaluating the model performance via suitable metrics is the next step of the benchmarking procedure. Depending on the type of task to be performed, there are a lot of different metrics.

Time-series-related metrics

Word Error Rate (WER) and Character Level Error Rate (CER) metrics are used here to measure speech recognition accuracy, and BLEU [38], Meteor [39], Rouge-L [40], and CIDEr [41] metrics are used to measure text generation quality. Model efficiency is primarily determined by inference speed, dynamic power consumption, and energy cost per inference; latency, on the other hand, quantifies response time.

Image-related metrics

Object detection: TPAUC (True Positive Area Under Curve) [42] evaluates detection quality by combining precision and recall into a single metric over varying thresholds. On the other hand, another metric Multiscale IoU [43] measures the Intersection over Union (IoU) across different scales, ensuring robustness of object detection against changes in object size. Spatial Recall Index [44] is another effective metric which assesses the spatial coverage of detected objects compared to ground truth, emphasizing correct localization over multiple

detections. Bayesian methods in object detection [45] provide probabilistic estimates for uncertainties and variations, improving robustness and interpretability of detections.

Image segmentation: The Tversky Loss [46], which is especially helpful for unbalanced data, and Average Precision (AP) [47], which combines recall and precision, are essential measures for image segmentation. A common method for assessing overlap in segmentation, which may be used both generally and by class, is the Dice coefficient. Inference speed is a common metric used to assess model efficiency, and the sum of categorical cross-entropy loss is a useful tool for estimating prediction accuracy.

Image classification: The main statistic in this domain is Classification Accuracy, which measures the proportion of correctly predicted occurrences in a usually uncomplicated manner.

Text-related metrics

Natural Language Processing (NLP): A variety of measures are used in NLP benchmarks. The F1 Score combines precision and recall. A basic indicator of correctness is accuracy; calibration examines how predictions and results align with each other; and fairness/bias evaluates the predictive parity between groups. Human evaluation and answer quality both entail judging a model's results subjectively.

Relational reasoning: Energy per inference is used to examine the power efficiency. Latency is used to measure response time, and accuracy is the main metric for the outcome of the relational reasoning.

HPC hardware metrics

Two important measures for assessing the performance of HPC technology are floating-point operations per second (FLOPS) and latency/bandwidth. Measuring the FLOPS is useful for tasks that require heavy numerical computation, such as scientific simulations and machine learning, as it quantifies the number of floating-point calculations performed per second, an important aspect of computing capacity.

On the other hand, in HPC systems, latency and bandwidth evaluate the effectiveness of data transfer. While bandwidth measures the amount of data transmitted in a specific amount of time, latency refers to the amount of time it takes the data to move between points. High bandwidth and low latency are necessary to reduce bottlenecks and communication delays and improve system performance as a whole.

Thus, while FLOPS measures computational power, latency and bandwidth assess data transport capacities - enabling a comprehensive assessment of the best possible HPC performance.

Domain-specific cases

In this section, we compare several AI benchmarking initiatives, including the respective tasks, metrics, and models.

Many papers present new benchmarks to address domain-specific use cases such as Big Detection [48] or HELM [49]. These domain-specific benchmarks aim to verify the performance of AI models independently of their hardware, based on the prerequisite that the given model produces the same results on any hardware that is supported by the software implementation. However, this cannot avoid the issue of compatibility: Depending on the target hardware and the software which supports that hardware, only a limited selection of models may be usable [50]. Moreover, if a given use-case requires inference to run at a certain speed to match an input- or output-stream, for example, a camera feed [51] or live audio recording [52], then both software and hardware are highly relevant factors.

In this section, we consider four different categories of AI domains, with a particular emphasis on medical applications. The first category focuses on time-series benchmarks (Table 1), which encompass a variety of tasks including speech recognition, Automatic Speech Recognition (ASR), and multilingual video captioning. These results underscore the importance of comprehensive evaluation metrics [52] and the value of extensive, diverse datasets [36,35]. Furthermore, it is highlighted how neuromorphic hardware systems, such as SpiNNaker 2 and Loihi 1, significantly outperform conventional alternatives in terms of energy cost per inference, thereby emphasizing the potential for energy-efficient AI solutions in time-series applications [53,54]. As the second category, we consider imaging benchmarks (Table 2). These illuminate certain trends and considerations across image classification, segmentation, and object detection tasks. Importantly, it is emphasized that fine-tuning hyperparameters and leveraging pre-training and self-supervised learning are essential for improved robustness and accuracy [55,56]. Regarding image segmentation, we note that there is a clear trade-off between accuracy and speed. High-accuracy models like Dual-Swin-L have lower inference speeds, whereas faster models like YoloactEdge offer limited accuracy. This underscores the need to balance these metrics based on application requirements [51]. Additionally, we see that the lack of software support on embedded devices remains a significant challenge, affecting performance and limiting applicability in certain contexts [50]. The third category is dedicated to medical imaging benchmarks (Table 3). It underlines the necessity of adapting models for specific medical tasks. Foundation models require significant adaptation to effectively serve medical imaging tasks, highlighting the importance of task-specific customization [57,58]. For medical object detection, pre-trained models on fine-grained data excel in segmentation tasks, while those on coarse-grained data perform better in classification tasks. This distinction underscores the need for selecting appropriate training data granularity [59]. Finally, the fourth category encompasses text-related benchmarks (NLP and relational reasoning; Table 4), revealing significant advancements and ongoing challenges in various subfields. Specialized models like BioALBERT show superior performance, indicating the benefits of domain-specific adaptations [60]. In Large Language Models, models such as Med-PaLM are rapidly advancing, with some nearing human expert performance levels, demonstrating the potential of LLMs in specialized fields like medicine [61,62]. Significant performance disparities between open-source and proprietary models indicate substantial room for improvement, particularly in visual perception and multimodal understanding [28,63].

Considering the papers presented in Table 1-4, only a few of them list the hardware or software that was used. Nevertheless, the ML frameworks that were used can often be identified by looking through the provided code samples connected to the papers. For hardware, since the majority of project relies on some type of NVIDIA GPUs [64], it can be assumed that most likely NVIDIA GPUs were used if no information is given. This does not seem surprising as support for NVIDIA GPU drivers has been closely integrated into machine learning libraries such as PyTorch. Furthermore, we assume that using specific hardware

also commonly implies the usage of the associated programming platform (in the case of an NVIDIA GPU, e.g., CUDA).

For software, the majority of the mentioned papers relies on PyTorch with only a few using Tensor Flow. For example, in the case of [50], the authors found that they could not use Py Torch on the embedded boards they were testing.

The provided tables shall serve as a handy reference for choosing the appropriate metrics, models, or datasets for specific tasks based on individual requirements. In this context, we have presented a short overview of the hardware and software currently in use. Yet, given the rapid pace of progress in AI, it is crucial to stay updated with the latest advancements in hardware development. Therefore, in the following section, we delve into recent and emerging HPC hardware systems, comparing different options and discussing their respective use cases in the context of AI applications.

Hardware benchmarking

The importance of hardware-software compatibility in AI model benchmarking leads us to considering hardware benchmarks (Table 5). The performance of AI models can be influenced by the hardware and software they run on, making it crucial to understand and compare different hardware systems. Hardware benchmarking serves to test and compare the performance of different hardware systems, providing insights into which hardware is best for a specific AI task. In general, we see that different accelerators have their strengths and weaknesses in specific applications. For example, Xilinx FPGA shows comparable or superior inference performance to NVIDIA's V100/A100 [65], while Intel's Habana Gaudi2 demonstrates high performance across tasks like Image Classification and NLP [66].

Esperanto.ai consumes one-tenth the energy of A100 while delivering better performance in LLMs and Recommendation Models [67]. Thus, the benchmarking process helps optimize the performance of AI models in real-world applications. In this section, we delve into hardware benchmarking specifics, considering methodologies, challenges, and potential solutions.

Neuromorphic systems

Neuromorphic hardware systems are becoming increasingly important for high-performance computing. In general, the approach of neuromorphic computing is to employ, for engineering purposes, structures and computational processes similar to those found in biological brains. This particularly involves highly parallelized computation with decentralized memory as well as spike signal transmission and sparse connectivity. The major goals of neuromorphic computing are to implement enhanced cognitive computation for neuroscience and AI, to achieve better scalability, and to minimize energy consumption [68-71]. Due to the numerous different approaches to implement neuromorphic systems, benchmarking is difficult, and no general overarching framework has been established so far [72-74]. Furthermore, to examine the benefits of neuromorphic hardware, measures of energy efficiency such as the energy-delay product (EDP) or energy per inference must be considered systematically, but have mostly been neglected by conventional benchmarking approaches.

Systems with particular relevance for high-performance computing are Intel's Loihi chips [75,69] and the SpiNNaker chips developed at University of Manchester and TU Dresden [76-78]. Both Loihi and SpiNNaker are now available in their second generation. SpiNNaker implementations can be customized in many different ways and thus, in principle, a wide range of different benchmark tests is possible. The existing studies so far include benchmarking of large-scale neural network simulations [79] and keyword spotting [54]. On Loihi 1, a number of benchmark tests have been performed as well, including sequential MNIST, CIFAR10, gesture recognition, relational reasoning, and keyword spotting [53,69,54,28]. It is further worth noting that the chip has not only been found suitable for approaches with bio-inspired neural networks but also for the efficient implementation of other algorithms, e.g., to solve graph search problems [69]. On Loihi 2, so far, (preliminary) results have been obtained for the following benchmarks: MNIST with Lava-DL Bootstrap [80], N-MNIST with Lava-DL Slayer [81], deep noise suppression [82], and COCO with TinyYOLOv3 [83]. Given the novelty of the Loihi 2 chip, these tests have not yet been embedded in a framework that allows a systematic comparison of the results. Note that at the moment, we can only reference benchmarking results comparing neuromorphic systems with other systems based on slightly different models (Table 1 and Table 4).

In the light of the mentioned issues, recently, the neuromorphic research community has kicked off the development of NeuroBench, which is a framework similar to MLPerf for benchmarking neuromorphic systems [74]. It is planned to develop two tracks, one "algorithmic" and one "systems" track, to disentangle software- and hardware-level characteristics. There is the hope that NeuroBench will eventually become a useful and established tool to compare neuromorphic platforms and implementations with respect to accuracy, average precision, runtime, as well as energy efficiency in a variety of established and novel tasks such as keyword spotting, gesture recognition, object detection, and prediction of chaotic functions. Such comparisons between neuromorphic systems will likely provide the basis for comparing and benchmarking neuromorphic hardware with other hardware such as GPU or FPGA systems.

Field-programmable gate arrays (FPGAs)

FPGAs as reconfigurable hardware platforms have a spot in the design space as flexible hardware accelerators, generally outperforming CPU- and GPU-based solutions in terms of energy efficiency, while being less complex to develop and less expensive than ASICs. In particular, FPGAs are a common implementation platform for traditional signal processing algorithms that require low latency, high throughput within energy constraints, as well as for prototyping architecture designs.

With the trend towards machine learning applications, FPGAs have generally been outperformed by GPU solutions in terms of raw processing power and the ease of programmability required by innovations in network design and network execution. As energy efficiency again becomes increasingly important alongside to flexibility, FPGAs exist as adaptable and reconfigurable energy-efficient accelerators [84,85]. Combined with the potentially high performance due to massive parallelism, the ability to implement data pre-/postprocessing on the same platform increases the relevance of FPGAs for AI applications.

For FPGA platforms, the implementation or choice of the accelerator architecture is critical to performance, as FPGAs are not native accelerators - meaning that benchmarks are highly dependent on the architecture design. With tightly integrated on-chip memory and a configurable on-chip network, high memory bandwidths can be achieved in addition to optimized parallel

computing. However, FPGAs are often resource-constrained, so that mapped neural network processing architectures are limited in size and therefore in absolute computational power, by the available memory and logic resources. Most research focuses on neural network inference, as training algorithms are not as easily and efficiently implemented due to the maximum size of memory-buffers and high-precision multiplication resources. Neural network accelerator designs for FPGAs tend to implement quantized computations and networks to improve the resource utilization and increase parallelism. While most accelerators use int8-quantization, the chosen quantization is flexible, and even binary networks can be implemented efficiently [86, 85]. In addition to the accelerator and platform design toolflows [87], FPGA-based accelerators require an accelerator-specific controller or library to implement and control neural network operations. Intel and Xilinx provide SDKs for their own inference engine IPs - i.e. accelerator design, system integration, runtime and software for quantizing, optimizing, and compiling neural networks implemented in TensorFlow, PyTorch, ONNX or others. This lack of common open source libraries increases the complexity of FPGA ML accelerator architecture implementation and integration. New FPGA solutions such as Xilinx's Versal and Intel's Stratix 10NX integrate AI-engines and tensor cores in addition to traditional DSP slices for optimized neural network processing, enabling higher performance for matrix- and vector computations, but trending more towards specialized hardware [88,89]. A notable example is the XVDPU, a high-performance CNN accelerator on the Versal platform, which demonstrates the potential of FPGAs in AI applications [90]. Intel's Stratix 10NX FPGA lists 143 int8 TOPS and an average computing speedup of 12x compared to NVIDIA's V100 [91].

For efficient computation, the architecture design is specialized and adapted to specific neural networks. It therefore does not natively support new primitives or operators, limiting the neural networks that can be compiled for specific accelerator designs. This and the focus on accelerating inference also increases the difficulty of comparing FPGA-based accelerators to CPU- and GPU-based implementations, as common benchmarks such as MLPerf cannot be used for FPGA designs, limiting comparability to specific networks or tasks of inference benchmarks.

Intel Habana Gaudi2 (ASIC)

Intel Habana Gaudi2, a second-generation deep learning accelerator, offers high-performance capabilities for scalable AI training and inference. It has participated in the MLPerf benchmarking, showcasing its proficiency in various AI tasks including image classification, segmentation, LLMs, and natural language processing (NLP) [92,93]. While its training times [92] were relatively higher than NVIDIA's DGX A100 for BERT and ResNet-50, its per-accelerator performance outperformed the A100 in ResNet-50 and was slightly lower in BERT [94]. The MLCommons benchmark results highlighted that both the Habana Gaudi2 and 4th Gen Intel Xeon Scalable processors deliver strong AI training performances. This challenges the prevailing notion that only NVIDIA GPUs are suitable for generative AI and large language models, positioning Intel's AI solutions as effective and scalable alternatives in the market [66].

GPU and similar systems

SambaNova: AI hardware and software solutions. Their DataScale SN10 system, benchmarked using MLPerf, achieves over 20x higher throughput and lower latency compared to NVIDIA's A100 [95]. The SN40L system supports models with up to 5 trillion parameters on a single node, enhancing LLM training, inference efficiency, and multimodal capabilities. SambaNova's suite now includes new models like Llama2 variants and BLOOM 176B, advancing open-source language models and providing multilingual capabilities for businesses managing and developing LLMs.

Graphcore: Graphcore specializes in machine learning hardware and software platforms, utilizing its Intelligence Processing Unit (IPU) chips for AI workloads [96]. In the MLPerf Training benchmark v1.1, their IPU-POD16 completed ResNet-50 training in 28.3 minutes. Larger configurations like IPU-POD64, 128, and 256 also showed impressive results. Notably, Graphcore's BERT Large model achieved the fastest single-server training time of 10.6 minutes in MLPerf 1.1, with the lowest host processor to AI compute ratio.

Graphcore's development environment, Poplar SDK, simplifies the implementation of deep learning frameworks such as TensorFlow/Keras, PyTorch, and ONNX.

Esperanto.ai: Esperanto.ai [67] has created compute servers with thousands of 64-bit RISC-V cores. These servers deliver high performance through advanced on-chip memory and compute fabric, potentially reducing GPU dependence. Their 64-bit microprocessor handles large data sets, aiming for high performance while lowering memory bandwidth and power usage. The custom RISC-V core features quad-issue out-of-order execution and supports multiple operating systems, including Linux.

Esperanto has developed solutions for machine learning frameworks like PyTorch and TensorFlow using the RISC-V open-source ecosystem [97]. Their ET-SoC-1 includes over 1,000 64-bit RISC-V CPUs, each with a vector/tensor unit, designed for AI and non-AI parallel workloads. Their software development kit supports efficient LLM execution. Benchmarking against Intel's Xeon Platinum 8380H and NVIDIA's GPUs, Esperanto's technology shows competitive performance and energy efficiency in MLPerf and ResNet-50 benchmarks. The ET-SoC-1 used one-tenth the energy of NVIDIA A100 for recommendation models and claimed better relative performance. However, results vary with workloads and system configurations, and factors like cost, software ecosystem, and support are also crucial.

AMD (MI2xx, MI3xx): The AMD MI2xx series, specifically the MI250X, is an exascale-class GPU accelerator designed for HPC workloads, with key specifications including a CDNA2 architecture, TSMC 6 nm FinFET lithography, 14080 stream processors, 220 compute units, a peak engine clock of 1700 MHz, 128 GB of HBM2e memory, and a peak memory bandwidth of 3.2 TB/s [98]. The MI3xx series, particularly the MI300X, is a flagship GPU-only accelerator with 192 GB of HBM3 memory and a CDNA 3 architecture, aimed at the large language model market [99]. Benchmarking data shows that the MI250X provides up to 4.9X the performance of competitive accelerators for double precision (FP64) HPC applications [100]. While specific benchmarking data for the MI3xx series is not readily available, the system is designed for processing the currently largest and most complex LLMs.

NVIDIA: NVIDIA [101] has participated in all the MLPerf benchmarks for both training and inference, using their GPU-based systems. They have submitted results for the following tasks: speech recognition, NLP, recommender systems, object

detection, image classification and more. In the MLPerf training 2.0, NVIDIA AI was the only platform attending all 8 tasks, and was the fastest (lower time to train) in 4 out of 8 tests, DLRM, Mini GO, 3D U-net, and RNN-T. It showed the fastest (higher relative per accelerator) performance in 6 out of 8 tests, DLRM, Mini GO, 3D U-net, RNN-T, BERT, and Mask R-CNN [101].

Future directions and limitations

While this paper provides a comprehensive overview of the currently most important hardware systems and their role in AI benchmarking, it is essential to acknowledge further emerging technologies that are poised to revolutionize the field.

Integrated chips for AI workloads, such as NVIDIA's Grace - Blackwell, AMD's MI series or Intel's announced Falcon Shore, combine multiple processing units into one chip, promising significant performance gains for AI workloads while reducing energy consumption.

Evaluating their scalability and versatility across different AI workloads and applications is crucial for understanding their benchmarking potential. This exploration can uncover new insights into their performance capabilities and limitations, ultimately optimizing AI applications on integrated superchips.

The integration of accelerators like GPUs or specialized AI chips into edge devices can enhance AI application performance, improving efficiency and reducing latency. However, practical challenges like power consumption, thermal management, and integration complexity must be carefully evaluated to determine the feasibility of using accelerator devices at the edge for AI benchmarking.

In addition, quantum computing has the potential to impact AI benchmarking by unlocking exponential speedups in certain computations, fundamentally altering the landscape of AI bench-marking. Moreover, quantum machine learning approaches can recognize patterns in data that classical algorithms might overlook, potentially leading to improved performance on specific AI benchmarks [102,103]. Furthermore, the quantum advantage could be very sensitive to the given class of a dataset [104]. This could give rise to novel, custom-crafted benchmark datasets suited to material sciences and quantum chemistry applications. Further research is necessary to explore quantum computing's full implications for AI benchmarking and harness its potential.

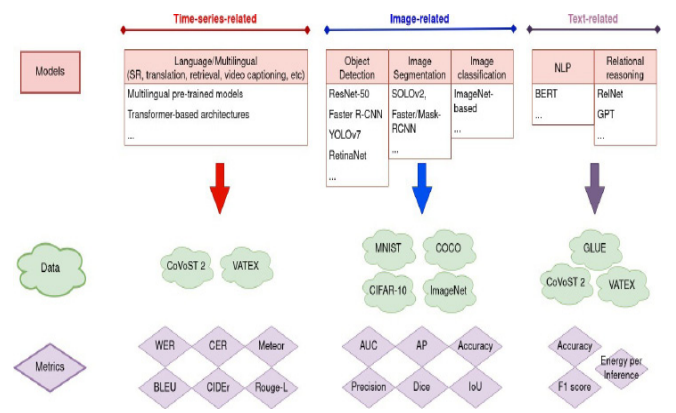

Figure 1: Different domains of benchmarking. We categorize the models considered here into three main types: time-series-, image-, and text-related. Under each category, certain relevant models are listed. In the data section, the relative datasets are listed. Lastly, the metrics section lists the diverse evaluation metrics that are used to assess the performance of the models on the given datasets.

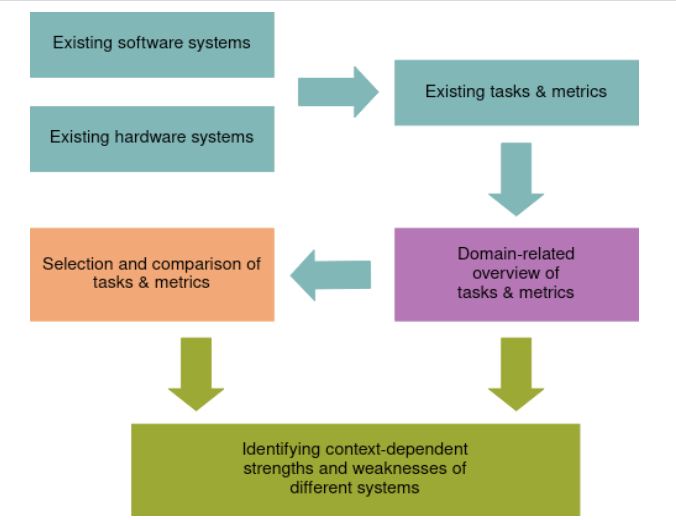

Figure 2: Overview of our proposed approach. We have collected information from existing tasks and metrics for different software and hardware systems. Based on this, we have drawn a selection of most representative findings, and present it in a domain-related overview. In the future, a similar selection of tasks and metrics shall be used to obtain a concise practical comparison of the particular strengths and weaknesses of hardware/software systems in different domains.

Table 1: Time-series benchmarks.

Task	Ref	Benchmark	Metric(s)	Model(s)	Key Findings
Speech Recognition	[52]	-	WER	-	Custom (based on neurons with after hyperpolarizing currents)
ASR, LangID, Translation, Retrieval	[18]	FLEURS	CER, Accuracy, % Precision at 1 (P@1)	Multilingual models including mSLAM	Provides baselines for the tasks based on multilingual pre-trained models, e.g., mSLAM. FLEURS includes parallel speech and text in 102 languages. Multimodal pre-training model shows promise for languages with ample unlabeled data.
ASR, Translation	[35]	CoVoST 2	WER, CER, BLEU	Transformer based architecture [19]	CoVoST 2 is the largest speech translation dataset with 2880 hours from 78K speakers.
Multilingual video captioning-Video - guided machine translation	[36]	VATEX	METEOR [38], ROGUEL [40], CIDEr [41], Accuracy	Attention based single or separate encoder decoder	VATEX dataset is larger and more diverse than MSRVTT [115]. Unified multilingual model outperforms monolingual models. Video context effectively aids in aligning languages for translation.
Keyword Spotting	[53]	Custom	Dynamic Power con-sumption, inference	Custom	Loihi 1 outperforms "conventional" (non-spiking) alternatives (Movidius NCS, Jetson TX1, Xeon E5-2630 CPU, Quadro K4000 GPU) regarding energy cost per inference at comparable accuracy.

	[54]	Power Benchmarks same as [53]	Inference speed, energy cost per inference	Power Benchmarks model same as in [53]	SpiNNaker 2 performs better than Loihi 1 in both metrics. It is important to note that on SpiNNaker, the rate-coded DNN is implemented directly, while on Loihi, it has to be converted to a spiking DNN.
Sequential Image Classification	[28]	(Sequential) MNIST	Classification accuracy, latency, energy per inference	Custom (based on neurons with after hyperpolarizing currents)	Better latency, better energy per inference, thus also better EDP with spiking NN on neuromorphic hardware (Loihi 1) compared to non-spiking CPU and GPU solutions.

Abbreviations: WER: Word Error Rate; ASR: Automatic Speech Recognition; Speech LangID: Speech Language Identification; CER: Character level Error Rate; METEOR: Metric for Evaluation of Translation with Explicit ORdering; BLEU: Bilingual Evaluation Understudy; ROUGE: Recall-Oriented Understudy for Gisting Evaluation; EDP: Energy Delay Product.

Table 2: Imaging benchmarks.

Task	Ref	Benchmark	Metric(s)	Model(s)	Key Findings
Image	[55]	DEIC	Balanced Classification Accuracy	Various	Always fine tune hyperparameters and Harmonic Net- works [116] and Cross-entropy were the best methods.
Classification	[56]	ARES bench	Classification Accuracy	55 ImageNet models	Pre-training and self-supervised learning improves natural robustness.
Image Segmentation	[51]	COCO 2017	AP, inference speed in FPS	Baseline: Mask R-CNN, Accuracy: SOTR, Dual-Swin-L, Speed: Cen- terMask, YOLACT, YolactEdge, BlendMask, SOLOv2	There is a trade off between AP and FPS, Dual-Swin-L has the best accuracy but low speed, YoloactEdge was the fastest with limited accuracy.
Object Detection	[48]	BigDetection	AP	Various ResNet,Swin-B models	BigDetection provides a valid alternative to Microsoft COCO.
	[50]	20 images from COCO	Inference time, pre processing, warm up	CenterNet, Single Shot Multibox Detection, EfficientDet, Faster R-CNN, Mask R-CNN	Limited software support for embedded devices, NXP i-MX8M- PLUS performed better but only 2 boards were tested.

Abbreviations: AP: Average Precision.

Table 3: Medical imaging benchmarks.

Task	Ref	Benchmark	Metric(s)	Model(s)	Key findings
Image	[55]	DEIC	Balanced Classification Accuracy	Various	Always fine tune hyperpa- rameters and Harmonic Net- works [116] and Cross-entropy were the best methods.
Classification	[56]	ARES-bench	Classification Accuracy	55 ImageNet models	Pre-training and self-supervised learning improves natural robust-ness.
Image Segmentation	[51]	COCO 2017	AP, inference speed in FPS	Baseline: Mask R-CNN, Accuracy: SOTR, Dual-Swin-L, Speed: CenterMask, YOLACT, YolactEdge, BlendMask, SOLOv2	There is a trade off between AP and FPS, Dual-Swin-L has the best accuracy but low speed, YoloactEdge was the fastest with limited accuracy.
Object Detection	[48]	BigDetection	AP	Various ResNet, Swin-B models	BigDetection provides a valid al-ternative to Microsoft COCO.
	[50]	20 images from COCO	Inference time, pre processing, warm up	CenterNet, Single Shot Multibox Detection, EfficientDet, Faster R-CNN, Mask R-CNN	Limited software support for em- bedded devices, NXP i-MX8M- PLUS performed better but only 2 boards were tested.

Abbreviations: AUC: Area Under the ROC Curve.

Table 4: Text-related (NLP and Relational Reasoning) benchmarks.

Task	Ref	Benchmark	Metric(s)	Model(s)	Key findings
Biomedical NLP	[60]	BLURB	F1-Score	BioALBERT	BioALBERT outperforms state- of-the-art in all but inference.
LLM	[49]	HELM	Accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency	30 closed, limited-access and open models	Text-davinci-002 performs the best for a accuracy, robustness and fairness, benefitting from instruction-tuning.
Medical LLM	[62]	MultiMedQA	Answer quality, human evaluation	Flan-PaLM, Med-PaLM	Med-PaLM exceeds state-of-the- art but not yet at human expert level.
	[61]	MultiMedQA	Answer quality, human evaluation	PaLM 2, Med-PaLM 2	Med-PaLM 2 further exceeds Med-PaLM and reaches human expert level.

Relational Reasoning	[28]	bAbI tasks (17 of 20 tasks)	Accuracy, latency, energy per inference	Spiking RelNet for Loihi 1, non-spiking RelNet from [119] for GPU	Worse latency, better energy per inference, and better EDP on Loihi 1 compared to non-spiking GPU solution.
Understand-ing and Reasoning	[63]	MMMU	Micro- averaged accuracy	4 open-source Large Multi- modal Models (LMMs) and GPT-4V(ision)	MMMU presents significant challenges, with even advanced models like GPT-4V achieving only 56% accuracy, indicating substantial room for improvement. Performance disparity between opensource models and proprietary ones like GPT-4V. Underscores the need for more research in vi- sual perception, knowledge representation, reasoning abilities, and multimodal joint understanding.

Table 5: Hardware benchmarking. This table compares various hardware accelerators used in AI and ML applications, including FPGAs and GPUs, from leading manufacturers.

Type	Hardware	Benchmark	Ref	Performance	Notes
FPGA	Xilinx VCK5000, Alveo U280	MLPerf (partial)	[120], [65]	Comparable to V100/A100 or outperforming in inference	Inference only – ResNet-50, SSD-ResNet34 (Image classification, object detection)
AI Chip	Intel Habana Gaudi2	MLPerf	[92], [94], [66]	Higher relative per accelerator performance in ResNet-50 than A100	(Image classification, Image segmentation, LLM, NLP) and inference (LLM)
GPU	SambaNova	MLPerf	[95]	20x faster than A100	DLRM model, Terabyte Click- through dataset
	Graphcore	MLPerf	[96]	Results close to A100	ResNet-50, BERT Large
	Esperanto.ai	MLPerf	[67], [97]	One-tenth energy usage compared to A100 and better perfor-mance compared to A100	LLMs (Generative AI), Recommendation models
	AMD (MI250)	-	[98]	Up to 4x faster compared to A100	Specs Comparisons
	NVIDIA (Grace Hop-per, A100, H100)	MLPerf	[101], [121], [122]	H100 were 6.7x faster than A100 (MLPerf training V2.1) in BERT model, Gracehopper showed 4.5x more performance over A100 (offline MLPerf inference) in BERT 99.9%	All

Discussion

Our review of benchmarking methods highlights the diverse landscape of AI systems. Key findings emphasize the effectiveness of pre-training and fine-tuning [56,55], the necessity of considering trade-offs (e.g. between accuracy and speed) [51,6], and the significant potential for enhancing AI performance by adapting models to specific domains. In this context, it is important to note the role of different hardware platforms. For instance, the inference performance of Xilinx FPGA is on par with or even surpasses that of NVIDIA's V100/A100 [65]. Meanwhile, Intel's Habana Gaudi2 shows high performance in areas such as Image Classification and NLP [66]. Several GPU accelerators, such as SambaNova, Graphcore, Esperanto.ai, and AMD's MI250, show significant advantages over NVIDIA A100 in various benchmarks [95,67,98]. The diversity in hardware platforms underscores the importance of selecting the right accelerator based on specific AI or ML workload requirements.

Furthermore, this review points to the importance of holistic evaluations that consider not just accuracy, but also robustness, fairness, bias [105], and efficiency [49].

Despite the valuable insights offered by AI benchmarks, they are often limited by a lack of inference performance [106,11,107], absence of representative workloads or real-world scenarios [108,109], or insufficient coverage of tasks, datasets, and metrics [110,6,111]. The work of Huerta et al. has underscored the significance of HPC in AI, with a particular emphasis on image-related data. Complementarily, our review paper broadens this scope by exploring a more diverse range of domains [112].

Also other limitations of the current state of AI benchmarking need to be noted, including the generalizability of benchmarks, the need for large datasets and expensive training, and the lack of clarity in defining the state-of-the-art [113]. However, there is a clear trend towards energy-efficient designs and competitive performance enhancements across different AI and ML workloads. This reflects ongoing innovation in the hardware sector aimed at improving efficiency and performance [67].

Looking ahead, we propose conducting a practical study that employs a specific set of designated tasks for each domain to obtain a fair comparison of the performance of various software and hardware systems (Figure 2). This comparative analysis should enable researchers and practitioners to precisely discern the strengths and weaknesses of different systems in a context-dependent manner.

By benchmarking these systems against one another, meaningful conclusions regarding their efficacy and suitability for specific tasks within each domain can be drawn. The comparative study will thus allow to evaluate the theoretical expectations discussed in our present work, pinpointing which system works best for specific tasks in different domains. Ultimately, this will facilitate informed decision-making by identifying which hardware/software system excels in particular domains, thereby enhancing the efficiency of applying different computational systems for diverse fields.

The KISSKI project, a research initiative focused on AI meth-

ods, is working towards such a comparison. The project, spearheaded by the Universities of Hannover and Göttingen, currently establishes a highly available AI service center for critical and sensitive infrastructures, particularly in the fields of medicine and energy. The next step in our KISSKI project is to systematically benchmark different hardware systems to identify which system is optimal for each domain. Furthermore, KISSKI will also focus on developing a standardized protocol for benchmarking, ensuring consistency and comparability across different systems. This will not only help in identifying the best hardware for each domain but also contribute to the broader goal of enhancing the efficiency and effectiveness of AI applications in high-performance computing.

Conclusion

Understanding the different existing benchmarks for AI is crucial for several reasons. For researchers and developers, benchmarks provide a standardized way to evaluate and compare the performance of different hardware and software configurations. For organizations investing in AI and HPC technologies, understanding benchmarks can help assess the value and potential return on investment of different solutions.

The landscape of AI benchmarking is vast and dynamic, with advancements across various domains such as speech recognition, image classification, and medical image classification. Models such as BioALBERT and ResNet have already surpassed previous state-of-the-art performances, demonstrating the potential of AI when coupled with HPC. Yet, there is a clear need for continued development and refinement of benchmarks to keep pace with the evolving capabilities of AI models [114]. In the light of many newly emerging hardware systems, more research on specialized hardware for specific AI tasks is also urgently needed [74]. And finally, given their crucial role in assessing system performance, the importance of targeting as many critical metrics as possible has become evident.

Through its examination of various benchmarking methods and a proposed pragmatic approach, this survey aims to provide a comprehensive framework for evaluating and optimizing AI applications across diverse computational systems and domains.

Declarations

Conflict of interest statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author contributions: Narges Lux - Planning, Researching, Joint primary author, Editing. Chirag Mandal - Planning, Researching, Joint primary author, Editing. Jonathan Decker - Planning, Researching, Writing, Editing. Jannik Luboeinski - Planning, Researching, Writing, Editing. Julian Drewljau - Researching, Writing. Tino Meisel - Researching, Writing. Christian Tetzlaff - Planning, Reviewing, Editing. Julian Kunkel - Planning, Reviewing, Editing. Christian Boehme - Planning, Reviewing, Editing.

Funding: Bundesministerium für Bildung und Forschung (BMBF), grant number FKZ 01 IS 22 093 A-E.

References

1. Michael R. King. The Future of AI in Medicine: A Perspective from a Chatbot. Annals of Biomedical Engineering. 2023; 51: 291-295.
2. Gangman Yi, Vincenzo Loia. High-performance computing systems and applications for AI. The Journal of Supercomputing. 2019; 75: 4248-4251.
3. Run: AI. HPC and AI: Better Together. 2023. https://www.run.ai/guides/hpc-clusters/hpc-and-ai, n.d.
4. Shuroug A Alowais, Sahar S Alghamdi, Nada Alsuhebany, Tariq Alqahtani, Abdulrahman I Alshaya, et al. Revolutionizing healthcare: The role of artificial intelligence in clinical practice. BMC Medical Education. 2023; 23: 689.
5. Daniel Reed, Dennis Gannon, and Jack Dongarra. Reinventing High Performance Computing: Challenges and Opportunities. arXiv preprint arXiv.2203.02544. 2022.
6. BenchCouncil. AIBench Training: Balanced AI Benchmarking, Bench Council. 2020. https://www.benchcouncil.org/aibench/training/index.html,n.d.
7. M´onica V Martins, Daniel Tolledo, Jorge Machado, Lu´ıs M T Baptista, Valentim Realinho. Early Prediction of student's Performance in Higher Education: A Case Study. In Trends and Applications in Information Systems and Technologies. Springer. 2021.
8. Xianhai Zhao, Yunjun Zhao, Mingyue Gou, Chang-Jun Liu. Tissue-preferential recruitment of electron transfer chains for cytochrome P450-catalyzed phenolic biosynthesis. Science. 2023; 380(6641): 136-138.
9. The Decoder. Are we running out of AI benchmarks? The Decoder. 2022.
10. Simon Ott, Adriano Barbosa-Silva, Kathrin Blagec, Jan Brauner, Matthias Samwald. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. arXiv: 2203.04592. 2022.
11. Zihan Jiang, Wanling Gao, Fei Tang, Xingwang Xiong, Lei Wang, et al. Hpc AI500: Representative, repeatable and simple HPC AI benchmarking. arXiv preprint arXiv:2102.12848. 2021.
12. Jeyan Thiyagalingam, Mallikarjun Shankar, Geoffrey Fox, Tony Hey. Scientific machine learning benchmarks. Nature Reviews Physics. 2022; 4(6): 413-420.
13. Deloitte. Gen AI use cases by type and industry. 2023. https://www2.deloitte.com/us/en/pages/consulting/articles/gen-ai-use-cases.html, n.d.
14. McKinsey. Notes from the AI Frontier: Insights from Hundreds of Use Cases. 2023. https://www.mckinsey.com/~/media/mckinsey/featured%20insights/artificial%20intelligence/notes%20from%20the%20ai%20frontier%20applications%20and%20value%20of%20deep%20learning/notes-from-the-ai-frontier-insights-from-hundreds-of-use-cases-discussion-paper.ashx.
15. Manar Abu Talib, Sohaib Majzoub, Qassim Nasir, Dina Jamal. A systematic literature review on hardware implementation of artificial intelligence algorithms. The Journal of Supercomputing. 2021; 77: 1897-1938.
16. Faisal Shehzad, Muhammad Rashid, Mohammed H Sinky, Saud S Alotaibi, Muhammad Yousuf Irfan Zia. A scalable system-on-chip acceleration for deep neural networks. IEEE Access. 2021; 9: 95412-95426.
17. BenchCouncil. HPC AI500: A Benchmark Suite for HPC AI Systems. 2023. https://www.benchcouncil.org/aibench/hpcai500/specification.html,n.d.
18. Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, et al. Fleurs: Few-shot learning evaluation of universal representations of speech. In 2022 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2023; 798-805.

19. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L- ukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

20. 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.

21. Ross Girshick. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 1440-1448, 2015.

22. Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 7464-7475, 2023.

23. Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dolla´r. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision,pages 2980-2988, 2017.

24. M Emin Sahin, Hasan Ulutas, Esra Yuce, and Mustafa Fatih Erkoc. Detection and classification of COVID-19 by using faster R-CNN and mask R-CNN on CT images. Neural Computing and Applications, 35(18):13597-13611, 2023.

25. Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. Advances in Neural information processing systems, 33:17721-17732, 2020.

26. P Sharma. popular image classification models in ImageNet challenge (ILSVRC) competition history. https://machinelearningknowledge.ai/popular-image-classificationmodels-in-imagenetchallenge-ilsvrc competition-history, n.d. Accessed: 2024-01-07.

27. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

28. Arjun Rao, Philipp Plank, Andreas Wild, and Wolfgang Maass. A long short-term memory for AI applications in spike-based neuromorphic hardware. Nature Machine Intelligence,4(5):467-479, 2022.

29. Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

30. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877-1901,2020.

31. 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.

32. Yann LeCun. The MNIST database of handwritten digits. http://yann. lecun. com/exdb/mnist/,1998.

33. Alex Krizhevsky and Geoffrey Hinton. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009.

34. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll´ar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740-755. Springer, 2014.

35. Changhan Wang, Anne Wu, and Juan Pino. Covost 2 and massively multilingual speech-to-text translation. arXiv preprint arXiv:2007.10310, 2020.

36. Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4581-4591, 2019.

37. Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 60-65, 2018.

38. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311-318, 2002.

39. Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the ninth workshop on statistical machine translation, 376-380, 2014.

40. Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74-81, 2004.

41. Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4566-4575, 2015.

42. Hanfang Yang, Kun Lu, Xiang Lyu, and Feifang Hu. Two-way partial auc and its properties. Statistical methods in medical research, 28(1):184-195, 2019.

43. Azim Ahmadzadeh, Dustin J Kempton, Yang Chen, and Rafal A Angryk. Multiscale iou: A metric for evaluation of salient object detection with fine structures. In 2021 IEEE International Conference on Image Processing (ICIP), pages 684-688. IEEE, 2021.

44. Patrick Mu¨ller, Mattis Brummel, and Alexander Braun. Spatial recall index for machine learning algorithms. In London Imaging Meeting, volume 2, pages 58-62. Society for Imaging Science and Technology, 2021.

45. Ali Harakeh, Michael Smart, and Steven L Waslander. Bayesod: A bayesian approach for uncertainty estimation in deep object detectors. In 2020 IEEE International Conference on Robotics and Automation (ICRA), pages 87-93. IEEE, 2020.

46. Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3D fully convolutional deep networks. In International workshop on machine learning in medical imaging, pages 379-387. Springer, 2017.

47. Cornelis Joost van Rijsbergen. Information Retrieval, 2nd Edition. 1979.

48. Likun Cai, Zhi Zhang, Yi Zhu, Li Zhang, Mu Li, and Xiangyang Xue. BigDetection: A Large-scale Benchmark for Improved Object Detector Pre-training. March 2022.

49. Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher R´e, Diana Acosta-Navas, Drew A. Hudson,

Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic Evaluation of Language Models. October 2023.

50. David Cantero, Iker Esnaola-Gonzalez, Jose Miguel-Alonso, and Ekaitz Jauregi. Benchmarking Object Detection Deep Learning Models in Embedded Devices. Sensors, 22(11):4205, January 2022.

51. Sunguk Jung, Hyeonbeom Heo, Sangheon Park, Sung-Uk Jung, and Kyungjae Lee. Bench-marking Deep Learning Models for Instance Segmentation. Applied Sciences, 12(17):8856,January 2022.

52. Aks¨enova, D. van Esch, J. Flynn, and P. Golik. How might we create better benchmarks for speech recognition? In Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future, pages 22-34. Association for Computational Linguistics, August 2021.

53. Peter Blouw, Xuan Choo, Eric Hunsberger, and Chris Eliasmith. Benchmarking keyword spotting efficiency on neuromorphic hardware. In Proceedings of the 7th annual neuro-inspired computational elements workshop, pages 1-8, 2019.

54. Yexin Yan, Terrence C Stewart, Xuan Choo, Bernhard Vogginger, Johannes Partzsch, Sebastian H¨oppner, Florian Kelber, Chris Eliasmith, Steve Furber, and Christian Mayr. Comparing Loihi with a SpiNNaker 2 prototype on low-latency keyword spotting and adaptive robotic control. Neuromorphic Computing and Engineering, 1(1):014002, 2021.

55. Lorenzo Brigato, Bj¨orn Barz, Luca Iocchi, and Joachim Denzler. Tune It or Don't Use It: Benchmarking Data-Efficient Image Classification. August 2021.

56. Chang Liu, Yinpeng Dong, Wenzhao Xiang, Xiao Yang, Hang Su, Jun Zhu, Yuefeng Chen, Yuan He, Hui Xue, and Shibao Zheng. A Comprehensive Study on Robustness of Image Classification Models: Benchmarking and Rethinking. February 2023.

57. Dequan Wang, Xiaosong Wang, Lilong Wang, Mengzhang Li, Qian Da, Xiaoqiang Liu, Xiangyu Gao, Jun Shen, Junjun He, Tian Shen, Qi Duan, Jie Zhao, Kang Li, Yu Qiao, and Shaoting Zhang. MedFMC: A Real-world Dataset and Benchmark For Foundation Model Adaptation in Medical Image Classification. arXiv:2306.09579, June 2023.

58. Dominik Mu¨ller and Frank Kramer. MIScnn: A framework for medical image segmentation with convolutional neural networks and deep learning. BMC Medical Imaging, 21(1):12, January 2021.

59. Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Ruibin Feng, Michael B. Gotway, and Jianming Liang. A Systematic Benchmarking Analysis of Transfer Learning for Medical Image Analysis. arXiv:2108.05930, August 2021.

60. Usman Naseem, Adam G. Dunn, Matloob Khushi, and Jinman Kim. Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT. BMC Bioinformatics. 23(1):144, April 2022.

61. Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards Expert-Level Medical Question Answering with Large Language Models. arXiv:2305.09617, May 2023.

62. Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large Language Models Encode Clinical Knowledge. arXiv:2212.13138, December 2022.

63. Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. arXiv preprint arXiv:2311.16502, 2023.

64. S. Suganyadevi, V. Seethalakshmi, and K. Balasamy. A Review on Deep Learning in Medical Image Analysis. International Journal of Multimedia Information Retrieval, 11(1):19-38,March 2022.

65. Xilinx. DPUv3E for Alveo Accelerator Card with HBM. https://www.xilinx.com/developer/articles/dpuv3e-for-alveo-accelerator-card-with-hbm.html, 2020. Accessed: 2024-01-08.

66. Samir Bashir. MLCommons stellt die Ergebnisse des Intel Habana Gaudi2 und des Intel Xeon Scalable AI Benchmark der 4. Generation zur Verfu¨gung. https://www.igorslab.de/en/mlcommons-stellt-die-ergebnisse-des-intel-habana-gaudi2-und-des-intel-xeon-scalable-ai-benchmark-der-4-generation-zur-verfuegung/, 2023. Accessed: 2023-11-10.

67. Esperanto.ai. Esperanto Technologies. https://www.esperanto.ai/technology/, n.d. Accessed: 2023-11-10.

68. Catherine D Schuman, Thomas E Potok, Robert M Patton, J Douglas Birdwell, Mark E Dean, Garrett S Rose, and James S Plank. A survey of neuromorphic computing and neural networks in hardware. arXiv preprint arXiv:1705.06963, 2017.

69. Mike Davies, Andreas Wild, Garrick Orchard, Yulia Sandamirskaya, Gabriel A Fonseca Guerra, Prasad Joshi, Philipp Plank, and Sumedh R Risbud. Advancing neuromorphic computing with Loihi: A survey of results and outlook. Proceedings of the IEEE, 109(5):911-934, 2021.

70. Shruti R Kulkarni, Maryam Parsa, J Parker Mitchell, and Catherine D Schuman. Benchmarking the performance of neuromorphic and spiking neural network simulators. Neurocomputing, 447:145-160, 2021.

71. Catherine D Schuman, Shruti R Kulkarni, Maryam Parsa, J Parker Mitchell, Prasanna Date, and Bill Kay. Opportunities for neuromorphic computing algorithms and applications. Nature Computational Science, 2(1):10-19, 2022.

72. Mike Davies. Benchmarks for progress in neuromorphic computing. Nature Machine Intelligence, 1(9):386-388, 2019.

73. Craig M Vineyard, Sam Green, William M Severa, and C¸ etin Kaya Ko¸c. Benchmarking event-driven neuromorphic architectures. In Proceedings of the International Conference on Neuromorphic Systems, pages 1-5, 2019.

74. Jason Yik, Soikat Hasan Ahmed, Zergham Ahmed, Brian Anderson, Andreas G Andreou, Chiara Bartolozzi, Arindam Basu, Douwe den Blanken, Petrut Bogdan, Sander Bohte, et al. NeuroBench: Advancing neuromorphic computing through collab-

orative, fair and representative benchmarking. arXiv preprint arXiv:2304.04640, 2023.

75. Mike Davies, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Yongqiang Cao, Sri Harsha Choday, Georgios Dimou, Prasad Joshi, Nabil Imam, Shweta Jain, et al. Loihi: A neuromorphic manycore processor with on-chip learning. IEEE Micro, 38(1):82-99, 2018.

76. Eustace Painkras, Luis A Plana, Jim Garside, Steve Temple, Francesco Galluppi, Cameron Patterson, David R Lester, Andrew D Brown, and Steve B Furber. SpiNNaker: A 1-W 18-core system-on-chip for massively-parallel neural network simulation. IEEE Journal of Solid-State Circuits, 48(8):1943-1953, 2013.

77. Steve B Furber, Francesco Galluppi, Steve Temple, and Luis A Plana. The SpiNNaker project. Proceedings of the IEEE, 102(5):652-665, 2014.

78. Yexin Yan, David Kappel, Felix Neum¨arker, Johannes Partzsch, Bernhard Vogginger, Sebastian H¨oppner, Steve Furber, Wolfgang Maass, Robert Legenstein, and Christian Mayr. Efficient reward-based structural plasticity on a SpiNNaker 2 prototype. IEEE transactions on biomedical circuits and systems, 13(3):579-591, 2019.

79. Sacha J Van Albada, Andrew G Rowley, Johanna Senk, Michael Hopkins, Maximilian Schmidt, Alan B Stokes, David R Lester, Markus Diesmann, and Steve B Furber. Performance comparison of the digital neuromorphic hardware SpiNNaker and the neural network simulation software NEST for a full-scale cortical microcircuit model. Frontiers in Neuroscience, 12:291,2018.

80. Intel Neuromorphic Computing Lab. Lava-DL: Bootstrap SNN Training. https://github.com/lava-nc/lava-dl/blob/main/tutorials/lava/lib/dl/bootstrap/mnist/train.ipynb,n.d. Accessed: 7-11-2023.

81. Intel Neuromorphic Computing Lab. Lava-DL: N-MNIST Classification. https://github.com/lava-nc/lava-dl/blob/main/tutorials/lava/lib/dl/slayer/nmnist/train.ipynb,n.d. Accessed: 7-11-2023.

82. Jonathan Timcheck, Sumit Bam Shrestha, Daniel Ben Dayan Rubin, Adam Kupryjanow, Garrick Orchard, Lukasz Pindor, Timothy Shea, and Mike Davies. The Intel neuromorphic DNS challenge. Neuromorphic Computing and Engineering, 3(3):034005, 2023.

83. Sumit Bam Shrestha. Deep Learning, 2023. Presentation at Intel Neuromorphic Research Community (INRC) Summer Workshop 2023.

84. Griffin Lacey, Graham W Taylor, and Shawki Areibi. Deep learning on fpgas: Past, present, and future. arXiv preprint arXiv:1602.04283, 2016.

85. Ahmed Ghazi Blaiech, Khaled Ben Khalifa, Carlos Valderrama, Marcelo AC Fernandes, and Mohamed Hedi Bedoui. A survey and taxonomy of FPGA-based deep learning accelerators. Journal of Systems Architecture, 98:331-345, 2019.

86. Kaiyuan Guo, Shulin Zeng, Jincheng Yu, Yu Wang, and Huazhong Yang. [DL] A survey of FPGA-based neural network inference accelerators. ACM Transactions on Reconfigurable Technology and Systems (TRETS), 12(1):1-26, 2019.

87. Stylianos I Venieris, Alexandros Kouris, and Christos-Savvas Bouganis. Toolflows for mapping convolutional neural networks on FPGAs: A survey and future directions. ACM Computing Surveys (CSUR), 51(3):1-39, 2018.

88. Brian Gaide, Dinesh Gaitonde, Chirag Ravishankar, and Trevor Bauer. Xilinx adaptive compute acceleration platform: VersalTM architecture. In Proceedings of the 2019 ACM/SIGDA. International Symposium on Field-Programmable Gate Arrays, pages 84-93, 2019.

89. Martin Langhammer, Eriko Nurvitadhi, Bogdan Pasca, and Sergey Gribok. Stratix 10 NX architecture and applications. In The 2021 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, pages 57-67, 2021.

90. Xijie Jia, Yu Zhang, Guangdong Liu, Xinlin Yang, Tianyu Zhang, Jia Zheng, Dongdong Xu, Zhuohuan Liu, Mengke Liu, Xiaoyang Yan, et al. Xvdpu: A high performance cnn accelerator on versal platform powered by ai engine. ACM Transactions on Reconfigurable Technology and Systems, 2022.

91. Andrew Boutros, Eriko Nurvitadhi, Rui Ma, Sergey Gribok, Zhipeng Zhao, James C Hoe, Vaughn Betz, and Martin Langhammer. Beyond peak performance: Comparing the real performance of AI-optimized FPGAs and GPUs. In 2020 International Conference on Field-Programmable Technology (ICFPT), pages 10-19. IEEE, 2020.

92. MLCommons. MLCommons Training Benchmarks. https://mlcommons.org/benchmarks/training/, n.d. Accessed: 2023-11-10.

93. MLCommons. MLPerf Inference v3.1 Results - MLCommons. https://mlcommons.org/en/inference-datacenter-31/, n.d. Accessed: 2023-11-07.

94. MLCommons. MLCommons Inference Datacenter Benchmark. https://mlcommons.org/benchmarks/inference-datacenter/, n.d. Accessed: 2023-11-10.

95. Sharon Goldman. Sambanova unveils new ai chip to power full-stack ai platform. https://venturebeat.com/ai/sambanova-unveils-new-ai-chip-to-power-full-stack-ai-platform/, n.d. Accessed: 2023-11-10.

96. Mrinal Iyer Matt Fyles. Performance at scale: Graphcore's latest mlperf training results. https://www.graphcore.ai/posts/performance-at-scale-graphcores-latest-mlperf-training-results, 2021. Accessed: 2023-11-10.

97. Jeffrey Burt. Esperanto chip drives ML inference performance and power efficiency. https://www.nextplatform.com/2021/09/20/esperanto-chip-drives-ml-inference-performance-and-power-efficiency/, n.d. Accessed: 2023-11-10.

98. AMD. AMD Instinct MI200 Series. https://www.amd.com/en/products/accelerators/instinct/mi200.html, 2021. Accessed: 2024-01-13.

99. AMD. AMD Instinct MI300X. https://www.amd.com/en/products/accelerators/instinct/mi300/mi300x.html, 2023. Accessed: 2024-01-13.

100. AMD. AMD MI200 Benchmarking. https://www.nextplatform.com/2021/12/06/stacking-up-amd-mi200-versus-nvidia-a100-compute-engines/, 2021. Accessed: 2024-01-13.

101. Shar Narasimhan. NVIDIA Partners AI MLPerf. https://blogs.nvidia.com/blog/nvidia-partners-ai-mlperf/, 2022. Accessed: 2023-11-10.

102. Leonardo Alchieri, Davide Badalotti, Pietro Bonardi, and Simone Bianco. An introduction to quantum machine learning: from quantum logic to quantum deep learning. Quantum Machine Intelligence, 3(2):28, November 2021.

103. Mo Kordzanganeh, Daria Kosichkina, and Alexey Melnikov. Parallel Hybrid Networks: An Interplay between Quantum and Classical Neural Networks. Intelligent Computing, 2:0028, October 2023.

104. Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R. McClean. Power of data in quantum machine learning. Nature Communi-

cations, 12(1):2631, May 2021.

105. J¨org Freiling and Sybille Huth. Limitations and challenges of benchmarking-a competence based perspective. In Competence Perspectives on Managing Interfirm Interactions, volume 8, pages 3-25. Emerald Group Publishing Limited, 2005.

106. Zhixiang Ren, Yongheng Liu, Tianhui Shi, Lei Xie, Yue Zhou, Jidong Zhai, Youhui Zhang, Yunquan Zhang, and Wenguang Chen. AIPerf: Automated machine learning as an AI-HPC benchmark. Big Data Mining and Analytics, 4(3):208-220, 2021.

107. MLCommons. MLPerf HPC v1.0 results. https://mlcommons.org/en/news/mlperf-hpc-v1/, 2021. Accessed: 2023-09-07.

108. Baidu-Research. DeepBench. http://research.baidu.com/Blog/index-view?id=100, 2018. Accessed: 2023-11-07.

109. Tal Ben-Nun, Maciej Besta, Simon Huber, Alexandros Nikolaos Ziogas, Daniel Peter, and Torsten Hoefler. 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.

110. Shenggui Li, Jiarui Fang, Zhengda Bian, Hongxin Liu, Yuliang Liu, Haichen Huang, Boxiang Wang, and Yang You. Colossal-AI: A unified deep learning system for large-scale parallel training. arXiv preprint arXiv:2110.14883, 2021.

111. Stanford DAWN. Stanford DAWN Deep Learning Benchmark (DAWNBench). https://dawn.cs.stanford.edu/benchmark/, n.d. Accessed: 2023-11-07.

112. Eliu A 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. Journal of Big Data, 7:1-12, 2020.

113. Ben Dickson. Why we must rethink AI benchmarks. https://bdtechtalks.com/2021/12/06/ai-benchmarks-limitations/, n.d. Accessed: 2023-11-28.

114. Usman Naseem, Adam G. Dunn, Matloob Khushi, and Jinman Kim. Benchmarking for Biomedical Natural Language Processing Tasks with a Domain Specific ALBERT. arXivpreprint arXiv:2107.04374, 2021.

115. Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5288-5296, 2016.

116. Matej Ulicny, Vladimir A. Krylov, and Rozenn Dahyot. Harmonic Networks for Image Classification. British Machine Vision Conference (BMVC), 2019.

117. Sheng He, Rina Bao, Jingpeng Li, Jeffrey Stout, Atle Bjornerud, P. Ellen Grant, and Yangming Ou. Computer-Vision Benchmark Segment-Anything Model (SAM) in Medical Images: Accuracy in 12 Datasets. May 2023.

118. Kathryn Wantlin, Chenwei Wu, Shih-Cheng Huang, Oishi Banerjee, Farah Dadabhoy, Veeral Vipin Mehta, Ryan Wonhee Han, Fang Cao, Raja R. Narayan, Errol Colak, Adewole Adamson, Laura Heacock, Geoffrey H. Tison, Alex Tamkin, and Pranav Rajpurkar. BenchMD: A Benchmark for Unified Learning on Medical Images and Sensors. arXiv:2304.08486, June 2023.

119. Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. Advances in neural information processing systems, 30, 2017.

120. Xilinx. Vitis AI Library User Guide (UG1354) v3.0. https://docs.xilinx.com/r/3.0-English/ug1354-xilinx-ai-sdk/VCK5000-Performance-with-an-8PE-DPUCVDX8H-350MHz,2023. Accessed: 2024-01-08.

121. NVIDIA. Hopper Supercomputer Accelerates MLPerf AI Training and HPC Workloads. https://blogs.nvidia.com/blog/mlperf-ai-training-hpc-hopper/, 2022. Accessed: 2024-01-07.

122. NVIDIA. NVIDIA Hopper H100 & L4 Ada GPUs Achieve Record-Breaking Performance in MLPerf AI Benchmarks. https://wccftech.com/nvidia-hopper-h100-l4-ada-gpus-achieve-record-breaking-performance-in-mlperf-ai-benchmarks/, 2023. Accessed: 2024-01-07.