



MLPERF TRAINING BENCHMARK

Peter Mattson¹ Christine Cheng² Cody Coleman³ Greg Diamos⁴ Paulius Micikevicius⁵ David Patterson¹⁶
Hanlin Tang² Gu-Yeon Wei⁷ Peter Bailis³ Victor Bittorf¹ David Brooks⁷ Dehao Chen¹ Debojyoti Dutta⁸
Udit Gupta⁷ Kim Hazelwood⁹ Andrew Hock¹⁰ Xinyuan Huang⁸ Atsushi Ike¹¹ Bill Jia⁹ Daniel Kang³
David Kanter¹² Naveen Kumar¹ Jeffery Liao¹³ Guokai Ma² Deepak Narayanan³ Tayo Oguntebi¹
Gennady Pekhimenko^{14,15} Lillian Pentecost⁷ Vijay Janapa Reddi⁷ Taylor Robie¹ Tom St. John¹⁶
Tsuguchika Tabaru¹¹ Carole-Jean Wu⁹ Lingjie Xu¹⁷ Masafumi Yamazaki¹¹ Cliff Young¹ Matei Zaharia³

ABSTRACT

Machine learning (ML) needs industry-standard performance benchmarks to support design and competitive evaluation of the many emerging software and hardware solutions for ML. But ML training presents three unique benchmarking challenges absent from other domains: optimizations that improve training throughput can increase the time to solution, training is stochastic and time to solution exhibits high variance, and software and hardware systems are so diverse that fair benchmarking with the same binary, code, and even hyperparameters is difficult. We therefore present MLPerf, an ML benchmark that overcomes these challenges. Our analysis quantitatively evaluates MLPerf’s efficacy at driving performance and scalability improvements across two rounds of results from multiple vendors.

1 INTRODUCTION

Machine learning (ML) has revolutionized numerous domains, including computer vision (Krizhevsky et al., 2012), language processing (Devlin et al., 2018; Radford et al., 2019), speech recognition (Hinton et al., 2012), and gaming (Silver et al., 2018; Mnih et al., 2013; Chan, 2018). Much of this progress owes to deep learning (DL), which involves training of large deep-neural-network (DNN) models on massive data sets. To keep up with this growing computational demand, hardware and software systems have garnered sizable investments (Amodei & Hernandez, 2018).

As the number of hardware and software systems for DL training increases (Paszke et al., 2017; Abadi et al., 2016; Chen et al., 2015; Jia et al., 2014; Jouppi et al., 2017; Chen et al., 2018; Markidis et al., 2018; Intel, 2019), so does the need for a comprehensive benchmark. History shows that benchmarks accelerate progress (Hennessy & Patterson, 2011); for example, breakthroughs in microprocessor and relational-database systems in the 1980s inspired industry consortiums to create *Standard Performance Evaluation*

Corporation (SPEC) for Unix servers (Dixit, 1991) and the *Transaction Processing Performance Council* (TPC) for transaction processing and databases (Council, 2005). These organizations helped develop and maintain benchmarks that their respective communities then embraced. Their success inspired the formation of MLPerf, a consortium of commercial and academic organizations, to design a comprehensive benchmark suite for DL.

Unlike other computational workloads, DL allows a range of statistical, hardware, and software optimizations that can change the mathematical semantics of the underlying operators. Although these optimizations can boost *performance* (i.e., training speed), some change the learning dynamics and affect the final model’s *quality* (i.e., accuracy). Even accommodating different system scales (e.g., varying the number of chips) requires changing hyperparameters, potentially affecting the amount of computation necessary to reach a particular quality target. By contrast, other compute benchmarks can evaluate systems through targeted microbenchmarks.

DL is also intrinsically approximate and stochastic, allowing multiple equally correct solutions—unlike conventional computing, which tends to allow just one correct solution. As a result, implementations and training times can vary while the final quality remains the same. Since it is approximate, DL requires careful definition of equally valid solution classes and the appropriate degrees of freedom.

Prior work has varied in granularity but has either left the

¹Google ²Intel ³Stanford University ⁴Landing AI ⁵NVIDIA
⁶University of California, Berkeley ⁷Harvard University ⁸Cisco
⁹Facebook ¹⁰Cerebras ¹¹Fujitsu ¹²Real World Technologies
¹³Synopsys ¹⁴University of Toronto ¹⁵Vector Institute ¹⁶Tesla
¹⁷Alibaba. Correspondence to: Peter Mattson <peremattson@google.com>.

above challenges unaddressed or lacked critical workloads representative of modern ML. Microbenchmarks such as DeepBench (Baidu, 2017) are affordable to run and enable a fair comparison of competing systems by isolating hardware and software from statistical optimizations, but they fail to reflect the complexity of real workloads and have limited utility. Although throughput benchmarks like Fathom and TBD (Adolf et al., 2016; Zhu et al., 2018; Google, 2017) evaluate full model architectures across a broad range of tasks to better reflect the diversity and complexity of real workloads, they limit model architecture and training innovations that advance the state-of-the-art. DAWNBench (Coleman et al., 2017) measures end-to-end training time, subject to a quality threshold (i.e., time to train), and it accommodates innovative solutions (i.e., new model architectures and training techniques, such as progressive resizing and cyclic learning rates). It additionally collects source code to promote reproducibility. DAWNBench’s flexibility, however, also made it difficult to draw fair comparisons between hardware and software platforms. MLPerf builds on the strengths of prior work; it combines a broad set of benchmarks like Fathom or TBD, an end-to-end training metric like DAWNBench, and the backing of a broad consortium like SPEC.

MLPerf aims to create a representative benchmark suite for ML that fairly evaluates system performance to meet five high-level goals:

- Enable fair comparison of competing systems while still encouraging ML innovation.
- Accelerate ML progress through fair and useful measurement.
- Enforce reproducibility to ensure reliable results.
- Serve both the commercial and research communities.
- Keep benchmarking effort affordable so all can participate.

This paper focuses on the design and rationale for the MLPerf Training benchmark (a related MLPerf Inference benchmark is beyond the present scope). Although prior ML benchmarking efforts (Coleman et al., 2017; Adolf et al., 2016; Google, 2017; Baidu, 2017; Zhu et al., 2018) each contributed to meeting one or more of the above goals, we created MLPerf to address *all* of them holistically, building on the lessons learned from these efforts. To this end, MLPerf Training does the following:

- Establish a comprehensive benchmark suite that covers diverse applications, DNN models, and optimizers.
- Create reference implementations of each benchmark to precisely define models and training procedures.

- Establish rules that ensure submissions are equivalent to these reference implementations and use equivalent hyperparameters.
- Establish timing rules to minimize the effects of stochasticity when comparing results.
- Make submission code open source so that the ML and systems communities can study and replicate the results.
- Form working groups to keep the benchmark suite up to date.

The rest of the paper is organized as follows. In § 2, we discuss the main challenges to benchmarks for DL training, as well as related prior work. In § 3, we review the benchmarks in our suite, the time-to-train metric, and quality thresholds. In § 4, we describe the submission, review, and reporting of results for the various categories. Finally, in § 5 and § 6, we review progress between the first two MLPerf benchmarking rounds, along with future work directions.

2 BACKGROUND

We begin by describing in § 2.1 the unique challenges of benchmarking ML relative to other compute tasks (Dongarra, 1988; Council, 2005) and then review prior ML-benchmarking efforts in § 2.2.

2.1 Unique Challenges of Benchmark Training

ML benchmarking faces unique challenges relative to other compute benchmarks, such as LINPACK (Dongarra, 1988) and SPEC (Dixit, 1991), that necessitate an end-to-end approach. After an ML practitioner selects a data set, optimizer, and DNN model, the system trains the model to its state-of-the-art *quality* (e.g., Top-1 accuracy for image classification). Provided the system meets this requirement, the practitioner can make different operation, implementation, and numerical-representation choices to maximize system *performance*—that is, how fast the training executes. Thus, an ML performance benchmark must ensure that systems under test achieve state-of-the-art quality while providing sufficient flexibility to accommodate different implementations. This tradeoff between quality and performance is challenging because multiple factors affect both the final quality and the time to achieve it.

2.1.1 Effect of Optimizations on Quality

Although many optimizations immediately improve traditional performance metrics such as throughput, some can decrease the final model quality, an effect that is only observable by running an entire training session. For example, the accuracy difference between single-precision training and

lower-precision training only emerges in later epochs (Zhu et al., 2016). Across several representation and training choices, the validation-error curves may only separate after tens of epochs, and some numerical representations never match the final validation error of full-precision training (lower validation error directly corresponds to higher accuracy: $\text{accuracy} = 1 - \text{error}_{\text{validation}}$). Thus, even though microbenchmarks (Baidu, 2017; Chetlur et al., 2014) can assess an optimization’s performance impact, a complete training session is necessary to determine the quality impact and whether the model achieves the desired accuracy. Owing to the introduction of systems with varying numerics (Abadi et al., 2016; Banner et al., 2018; Kster et al., 2017; Micikevicius et al., 2018) and performance optimizations, ML benchmarks must include accuracy metrics.

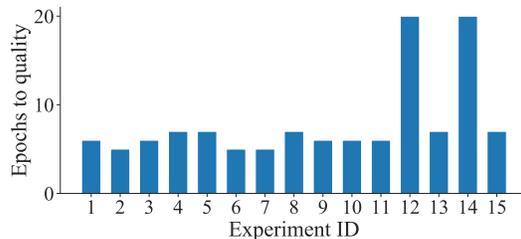
2.1.2 Effect of Scale on Time to Train

ML training on large distributed systems with many processors typically involves data parallelism and large minibatches to maximize system utilization and minimize training time. In turn, these large minibatches require adjustments to optimizer parameters, such as the learning rate (Krizhevsky, 2014; Goyal et al., 2017). Together, these changes affect the learning dynamics and can alter the number of iterations required to achieve the target accuracy. For example, MLPerf v0.5 ResNet-50 takes about 64 epochs to reach the target Top-1 accuracy of 74.9% at a minibatch size of 4K,¹ whereas a minibatch size of 16K can require more than 80 epochs to reach the same accuracy, increasing computation by 30%. Larger minibatches, however, permit efficient scaling to larger distributed systems, reducing the time to train the model. The tradeoffs between system size, minibatch size, and learning dynamics present another challenge for a DL-focused performance benchmark.

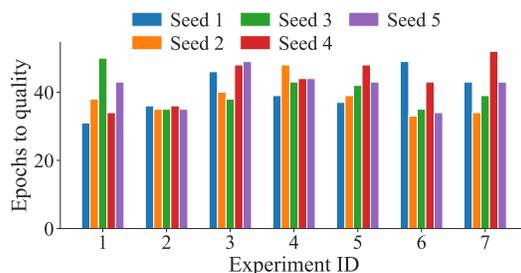
2.1.3 Run-to-Run Variation

DNN training involves many stochastic influences that manifest in substantial run-to-run variation (Choromanska et al., 2015; Gori & Tesi, 1992; Auer et al., 1996; Coleman et al., 2019). Different training sessions for the same model using the same hyperparameters can yield slightly different accuracies after a fixed number of epochs. Alternatively, different training sessions can take a different number of epochs to reach a given target accuracy. For example, Figure 1 shows the number of epochs needed to reach target accuracy for two MLPerf v0.5 benchmarks using reference implementations and default batch sizes. Several factors contribute to this variation, such as application behavior (e.g., random weight initialization and random data traversal) and system characteristics (e.g., profile-driven algorithm

¹Source: MLPerf v0.5 results (<https://mlperf.org/training-results-0-5>).



(a) NCF.



(b) MiniGo.

Figure 1. Training epochs to reach the target quality for the MLPerf v0.5 NCF (a) and MiniGo (b) benchmarks. Each experiment uses identical hyperparameters except for the random seed. For MiniGo, we observed considerable variability across runs even when fixing the random seed (same color).

selection and the non-commutative nature of floating-point addition). Large distributed-training tasks can involve asynchronous updates, altering the gradient-accumulation order. These variations make it hard to reliably compare system performance.

2.1.4 Diverse Software

Multiple ML software frameworks have emerged, each of which executes similar but distinct computations owing to various implementations and constraints (Abadi et al., 2016; Paszke et al., 2017; Chen et al., 2015; Jia et al., 2014). Software frameworks and the underlying math libraries employ different algorithms to implement the same operation. For example, convolutional and fully connected layers—two compute-intensive operators prevalent in modern DNN models—typically use cache blocking to exploit processor memory hierarchies. Different block sizes and processing orders (which optimize for different hardware), although algebraically equivalent, yield slightly divergent results. In addition, operators can execute using various algorithms. For example, convolution layers can be executed using a variety of algorithms, including GEMM-based and transform-based (e.g., FFT or Winograd) variants. In fact, the cuDNN v7.6 library provides roughly 10 algorithms for the forward pass of a convolutional layer,² some of which vary in tiling

²Source: cuDNN (<https://docs.nvidia.com/deeplearning/sdk/cudnn-developer-guide>).

or blocking choices depending on the hardware. Although mathematically equivalent, different implementations will produce different numerical results, as floating-point representations have finite precision.

Additionally, frameworks occasionally implement the same function in mathematically different ways. For example, modern training frameworks implement stochastic gradient descent with momentum in two ways:

$$\begin{aligned} \text{momentum} &= \alpha \cdot \text{momentum} + \eta \cdot \frac{\partial L}{\partial w} \\ w &= w - \text{momentum} \end{aligned} \quad (1)$$

$$\begin{aligned} \text{momentum} &= \alpha \cdot \text{momentum} + \frac{\partial L}{\partial w} \\ w &= w - \eta \cdot \text{momentum} \end{aligned} \quad (2)$$

The Caffe framework (Jia et al., 2014) implements the first approach, whereas PyTorch (Paszke et al., 2017) and TensorFlow (Abadi et al., 2016) implement the second. These approaches differ mathematically if the learning rate η changes during training—a common technique. Although this difference is tiny in many cases, it can hinder training convergence for larger minibatches.

Variations also arise owing to the frameworks’ programming interface. For example, PyTorch and TensorFlow interpret asymmetric padding differently, complicating the task of porting model weights between them. Data-augmentation pipelines across frameworks can also apply image augmentations (e.g., crop, zoom, and rotation) in different orders.

Although ONNX (Bai et al., 2019), TVM (Chen et al., 2018), and similar emerging tools enable interoperability of model architectures across frameworks, their support remains limited. Moreover, ML systems involve a range of optimizations that extend beyond the model architecture, such as preprocessing, precision, and communication methods. Benchmarks must accommodate the wide diversity of deployed systems despite this lack of a standard way to specify every training aspect.

2.2 Prior Work

Prior ML benchmarks vary in granularity and scope. Microbenchmarks such as DeepBench (Baidu, 2017) measure kernel-level operations that appear in commonly deployed models. Benchmarking such low-level operations fails to address the challenges associated with numerical precision, hyperparameter choices, and system scale, which we described in the previous section. Furthermore, it neither captures the end-to-end application, nor accounts for memory- and cache-hierarchy effects across layers and operations, nor measures the data preprocessing that deep learning commonly employs.

Several benchmarks are defined at the granularity of entire DNN models. Fathom and Google TF Benchmarks (Adolf et al., 2016; Google, 2017) provide a reference suite of DNN models that span a wide application space, but they specifically measure model throughput and fail to account for accuracy. Similarly, TBD (Training Benchmarks for DNNs) (Zhu et al., 2018) profiles training on GPUs (but not other architectures) across diverse workloads, measuring characteristics such as memory and hardware utilization. Our benchmark builds on the diversity of applications in these projects while also capturing the quality and performance tradeoffs.

DAWNBench (Coleman et al., 2017) was the first multi-entrant benchmark competition to use “time to train” (originally called time to accuracy) to measure the end-to-end performance of deep-learning systems; it allowed optimizations across model architectures, optimization procedures, software frameworks, and hardware platforms. Our benchmark follows a similar approach but handles more-diverse tasks (§ 3.1), and it uses important rules and mechanisms in the Closed division (§ 4.2.1) to enable fair comparisons of hardware and software systems.

Several other benchmarks are under development. AI Matrix measures workloads at different granularities (microbenchmarks, layer-wise benchmarks, end-to-end model benchmarks, and synthetic benchmarks) (aim). Deep500, although not a benchmark, provides a software framework for measuring DL-training performance (Ben-Nun et al., 2019).

3 MLPERF TRAINING BENCHMARK

We now present the MLPerf Training benchmark, detailing the workloads (§ 3.1), timing rules (§ 3.2), quality-threshold choices (§ 3.3), and reference implementations and hyperparameters (§ 3.4).

3.1 Benchmark Suite

To create a fair and useful benchmark suite for modern ML workloads, we curated a representative set of tasks from several major ML areas, including vision, language, recommendation, and reinforcement learning. Our selection of benchmarks was primarily based on commercial and research relevance, representing diverse compute motifs. To establish relevance, we relied on feedback from the tens of commercial and academic organizations that support MLPerf. To keep the suite affordable, we selected a compact but representative set of seven benchmarks, which we describe below and summarize in Table 1. Although these benchmarks already cover a wide range of research and industrial tasks, we are continuously exploring additional ones to keep the suite relevant to the ML community (§ 6).

Table 1. MLPerf Training v0.5 benchmarks.

Benchmark	Data set	Model	Quality Threshold
Image classification	ImageNet (Deng et al., 2009)	ResNet-50 v1.5 (MLPerf, 2019b)	74.9% Top-1 accuracy
Object detection (lightweight)	COCO 2017 (Lin et al., 2014)	SSD-ResNet-34 (Liu et al., 2016)	21.2 mAP
Instance segmentation and object detection (heavyweight)	COCO 2017 (Lin et al., 2014)	Mask R-CNN (He et al., 2017a)	37.7 Box min AP, 33.9 Mask min AP
Translation (recurrent)	WMT16 EN-DE (WMT, 2016)	GNMT (Wu et al., 2016)	21.8 Sacre BLEU
Translation (nonrecurrent)	WMT17 EN-DE (WMT, 2017)	Transformer (Vaswani et al., 2017)	25.0 BLEU
Recommendation	MovieLens-20M (GroupLens, 2016)	NCF (He et al., 2017b)	0.635 HR@10
Reinforcement learning	Go (9x9 Board)	MiniGo (MLPerf, 2019a)	40.0% Professional move prediction

3.1.1 Image Classification

Image classification is the most common task for evaluating ML-system performance (Coleman et al., 2017; Adolf et al., 2016; Zhu et al., 2018; Goyal et al., 2017; Jia et al., 2018; Mikami et al., 2018; Ying et al., 2018; Google, 2017; Narayanan et al., 2019). A classifier selects a class that best describes the contents of a given image. Classification model architectures also serve as feature extractors for many other computer-vision workloads, including object detection, captioning, and style transfer. We use the ILSVRC 2012 ImageNet classification data set, consisting of 1.28 million training images and 50,000 validation images (Deng et al., 2009). Our model-quality metric is the Top-1 accuracy on the validation set.

ResNet-50 is a residual network (He et al., 2016a;b); such networks and their derivatives remain the state of the art in image classification, and system studies commonly use them (Goyal et al., 2017; Jia et al., 2018; Mikami et al., 2018; Ying et al., 2018; Sun et al., 2019). Several slightly different ResNet-50 implementations appear in training-framework repositories, preventing comparison of earlier system-performance claims because of model differences. To ensure meaningful system comparison, MLPerf uses the ResNet-50 v1.5 model, which performs addition after batch normalization, omits 1×1 convolution from the skip connection of the first residual block, and applies downsampling by the 3×3 convolutions. MLPerf also specifies the appropriate parameter initialization, optimizer schedule, and data augmentation.

3.1.2 Object Detection and Segmentation

Object detection and segmentation are crucial components of many industrial systems for robotics, autonomous driving, video analytics, and social networks. Object detection is a regression task as opposed to a classification task: it returns bounding-box coordinates for objects in a given image. Segmentation assigns an object class to each input-image pixel. Although pretrained image-classification models commonly serve as the backbone (feature extractor) for DNN object detectors and segmenters, these DNN tasks differ from image classification in their compute characteristics. Examples include additional layer types (upsampling, ROIalign, NMS, and sorting); moreover, the inputs have greater resolution. MLPerf uses the 2017 COCO data set (Lin et al., 2014) consisting of 118,000 training images and 5,000 validation images. Model-quality measurement uses mAP for both detection and segmentation.

Mask R-CNN (He et al., 2017a) is a popular object-detection and instance-segmentation model for images. It has two stages: the first proposes regions of interest, and the second processes them to compute bounding boxes and segmentation masks. Mask R-CNN provides high-accuracy results for these tasks, but at the cost of higher latency as well as greater compute and memory requirements. The benchmark training uses images resized to 800 pixels on the shorter side and employs ResNet-50 as the backbone.

Single Shot Detection (SSD) (Liu et al., 2016) serves in real-time applications that require low-latency solutions. These applications include autonomous driving, robotics, and video analytics. Compared with Mask R-CNN (Huang et al., 2016) and other two-stage solutions, SSD trades speed for accuracy. Instead of full images, training uses 300×300

crops. We chose a ResNet-34 backbone to represent current real-time applications. ResNet-34 has a different residual-block structure than ResNet-50, increasing the diversity of computational motifs that MLPerf covers.

3.1.3 Translation

Neural machine translation converts a sequence of words from the source language to a target language; many industrial applications employ this technology. As is common in translation research, we use the WMT English-to-German (EN-DE) data set (WMT, 2017), which contains about 4.5 million sentence pairs. Our model-quality metric is the Bilingual Evaluation Understudy Score (Bleu) score on the Newstest2014 test set. We include two translation benchmarks to account for the two model architectures that translation and other sequence-data tasks often employ.

Transformer (Vaswani et al., 2017) is an attention-based model that achieves state-of-the-art language-translation quality. It consists of an encoder and decoder, each being a stack of six blocks. Every block comprises a multihead attention layer and point-wise fully connected layers.

GNMT (Wu et al., 2016) is a recurrent neural network (RNN) for language translation. Even though it achieves lower accuracy than Transformer on the WMT English-to-German data set, it appears in the suite to represent RNN applications. These applications span numerous tasks, but language-translation data sets and publications are more common, enabling clearer system comparison. GNMT is the suite’s only RNN. It consists of an eight-layer encoder and an eight-layer decoder, each using 1,024 LSTM cells with skip connections.

3.1.4 Reinforcement Learning

Reinforcement learning (RL) is responsible for the recent dramatic increase in compute demand (Amodei & Hernandez, 2018), and it serves in control systems. RL algorithms can train agents (which includes neural networks) that rival humans at video games, go, and chess—major milestones in machine learning (Silver et al., 2018; Mnih et al., 2013; Chan, 2018). RL has a different computational profile than the other ML benchmarks: it generates training data through exploration instead of relying on a predetermined data set.

MiniGo (MLPerf, 2019a), inspired by AlphaGo (Silver et al., 2016; 2017; 2018), trains a single model that represents both value and policy functions for a 9×9 game board. Training uses self-play (simulated games) between agents to generate data; rather than using a simulator, it performs many forward passes through the model to generate actions. We chose MiniGo to keep MLPerf more ML oriented, since many other RL problems employ simulators (physics, video-game environments, etc.) to generate data,

spending most of their time in computations unrelated to ML. To measure quality, we calculate the percentage of predicted moves that match human reference games.

3.1.5 Recommendation

Recommendation systems are a major commercial workload for Internet companies (Naumov et al., 2019; Zhou et al., 2018; Cheng et al., 2016). These workloads are characterized by large embedding tables followed by linear layers.

Neural collaborative filtering (NCF) (He et al., 2017b) was our choice for the benchmark. It is trained to predict user-item interactions. More so than for other tasks, this recommender’s compute characteristics depend on the data set. For example, the data set defines the embedding-table size as well as the memory-access patterns. Thus, a representative data set is crucial to a representative benchmark. Unfortunately, however, public data sets tend to be orders of magnitude smaller than industrial data sets. Although MLPerf v0.5 adopted the MovieLens-20M data set (GroupLens, 2016) for its NCF benchmark, v0.7 will employ a synthetically generated data set and benchmark while retaining the characteristics of the original data (Belletti et al., 2019)

3.2 Time-to-Train Performance Metric

To address the ML-benchmarking challenges of system optimization and scale that we outlined in § 2.1.1 and § 2.1.2, MLPerf’s performance metric is the time to train to a defined quality target. It incorporates both system speed and accuracy and is most relevant to ML practitioners. As an end-to-end metric, it also captures the auxiliary operations necessary for training such models, including data-pipeline and accuracy calculations. The metric’s generality enables application to reinforcement learning, unsupervised learning, generative adversarial networks, and other training schemes. Time to train overcomes the challenges in § 2.1.1 and § 2.1.2 by preventing submissions from using quality-reducing optimizations while still allowing for extensive system-scale and software-environment flexibility.

3.2.1 Timing Rules

We chose the timing requirements to ensure fair system comparisons and to represent various training use cases. Timing begins when the system touches any training or validation data, and it stops when the system achieves the defined quality target on the validation data set.

We exclude from timing several components that can carry substantial overhead and that are unrepresentative of real-world differences.

System initialization. Initialization, especially at large

scales, varies on the basis of cluster-administrator choices and system-queue load. For example, it may involve running diagnostics on each node before starting the training job. Such overheads are unindicative of a system’s training capability, so we exclude them from timing.

Model creation and initialization. Some frameworks can compile the model graph to optimize subsequent execution. This compilation time is insignificant for the longer training sessions when using industry-scale data sets. MLPerf, however, uses public data sets that are usually much smaller than industry ones. Therefore, large distributed systems can train some MLPerf benchmarks in minutes, making compilation times a substantial portion of the total time. To make benchmarks representative of training on the largest industrial data sets, we allow exclusion of up to 20 minutes of model-creation time. This limit ensures that MLPerf captures smaller training jobs, and it discourages submissions with compilation approaches that are too computationally and operationally expensive to use in practice.

Data reformatting. The raw input data commonly undergoes reformatting once and then serves in many subsequent training sessions. Reformatting examples include changing image-file formats and creating a database (e.g., LMDB, TFRecords, or RecordIO) for more-efficient access. Because these operations execute once for many training sessions, MLPerf timing excludes reformatting. But it prohibits any data processing or augmentation that occurs in training from moving to the reformatting stage (e.g., it prevents different crops of each image from being created and saved before the timed training stage).

3.2.2 Number of Timing Runs

To address the stochastic nature and resulting run-to-run variance of modern deep-learning methods described in § 2.1.3, MLPerf requires that submissions provide several runs of each benchmark to stabilize timing. We determined the number of runs, which varies among benchmarks, by studying the behavior of reference implementations. Vision tasks require 5 runs to ensure 90% of entries from the same system are within 5%; all other tasks require 10 runs to ensure 90% of entries from the same system are within 10%. MLPerf drops the fastest and slowest times, reporting the arithmetic mean of the remaining runs as the result.

3.3 Choice of Quality Thresholds

For each benchmark, we chose quality metrics near the state of the art for the corresponding model and data set (Table 1), basing our choice on experiments with the reference implementations. Some of these thresholds are slightly lower than results in the literature, enabling us to benchmark across software frameworks and to ensure that training sessions

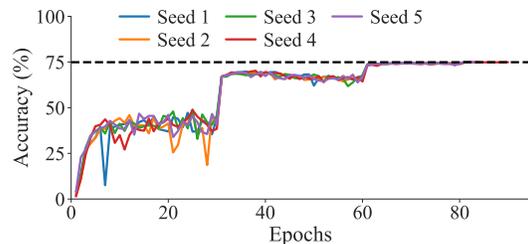


Figure 2. Top-1 accuracy of MLPerf v0.5 ResNet-50 benchmark over 100 epochs for five runs (denoted by color) with identical hyperparameters but different random seeds. The dashed line indicates the quality target of 74.9% Top-1 accuracy. The early training phase exhibits much more variability than later phases.

consistently achieve the quality metric. Although selecting a lower threshold that is achievable earlier in a training session reduces submission resources, we chose higher thresholds that require longer training sessions for two reasons: First, we must prevent optimizations from adversely affecting the final results (challenges described in § 2.1.1 and § 2.1.2). Second, we must minimize run-to-run variation, which tends to be much higher early in training. For example, Figure 2 shows accuracy for five training sessions of MLPerf v0.5’s ResNet-50 v1.5 reference implementation, where the first 30 epochs exhibit considerably more noise.

3.4 References and Hyperparameters

MLPerf provides a reference implementation for each benchmark, using either the PyTorch or TensorFlow framework. References also include scripts or directions to download and preprocess public data sets. References are not optimized for performance (meaning they should not be used for performance assessment or comparison), as their main purpose is to define a concrete implementation of a benchmark model and training procedure. All submitters must follow these references—they may reimplement a benchmark in their framework of choice as long as the DNN model and training operations are mathematically equivalent to the reference. Furthermore, MLPerf uses reference implementations to establish the required quality thresholds.

MLPerf rules specify the modifiable hyperparameters (Table 2) as well as restrictions on their modification. These restrictions are intended to balance the need to tune for different systems with limiting the size of the hyperparameter search space to be fair to submitters with smaller compute resources. For example, to accommodate a wide range of training-system scales, submissions must be able to adjust the minibatch size used by SGD in order to showcase maximum system efficiency (this approach is similar in concept to the Top500 LINPACK benchmark, which allows systems to choose the problem size). To ensure that training still

Table 2. MLPerf modifiable hyperparameters.

Model	Modifiable Hyperparameters
All that use SGD	Batch size, Learning-rate schedule parameters
ResNet-50 v1.5	
SSD-ResNet-34	Maximum samples per training patch
Mask R-CNN	Number of image candidates
GNMT	Learning-rate decay function, Learning rate, Decay start, Decay interval, Warmup function, Warmup steps
Transformer	Optimizer: Adam (Kingma & Ba, 2015) or Lazy Adam, Learning rate, Warmup steps
NCF	Optimizer: Adam or Lazy Adam, Learning rate, β_1 , β_2
Go (9x9 board)	

converges to the required threshold, other hyperparameters—such as the learning rate schedule—may need adjustment to match. For example, a common ResNet training practice is to increase the learning rate linearly with the minibatch size (Goyal et al., 2017). Although these hyperparameter searches are a common ML task, MLPerf’s focus is on system optimization rather than hyperparameter exploration and we do not want to penalize submitters who are unable to do extensive searches. Therefore we restrict hyperparameter tuning to subset of all possible parameters and values.

Further, we allow “hyperparameter borrowing” during the post-submission review process in which one submitter may adopt another submitter’s hyperparameters for a specific benchmark and resubmit their result (with no other hardware or software changes allowed). In the first two rounds, hyperparameter borrowing was used successfully to improve several submissions indicating hyperparameters are somewhat portable. Typically borrowing occurred across systems of similar scale, but did result in convergence across different numerics (FP16, bfloat16, and FP32), architectures (CPU, GPU, and TPU), and software implementations (TF, cuDNN, and MKL-DNN). MLPerf working groups review the hyperparameter choices and requirements for each benchmark round to account for advances in training ML models at scale.

4 BENCHMARKING PROCESS

Next, we outline the process for submission and review (§ 4.1) and for reporting results (§ 4.2) to account for innovative solutions, availability, and scale. We have run two

rounds of the MLPerf benchmark: v0.5 and v0.6. The time between rounds is about a few months, allowing us to update the suite after each one. Every round has a submission and review period followed by publication of results.

4.1 Submission and Review

An MLPerf submission consists of a system description, training-session log files, and all code and libraries required to reproduce the training sessions. All of this information is publicly available on the MLPerf GitHub site, along with the MLPerf results, allowing for reproducibility and enabling the community to improve the results in subsequent rounds. A system description includes both the hardware (number of nodes, processor and accelerator counts and types, storage per node, and network interconnect) and the software (operating system as well as libraries and their versions). A training-session log file contains a variety of structured information including time stamps for important workload stages, quality-metric evaluations at prescribed intervals, and hyperparameter choices. These logs are the foundation for analyzing results.

Before publishing results, submissions are peer-reviewed for compliance with MLPerf rules. Submitters receive notification of noncompliance, where applicable, and they may resubmit after addressing any such problems. Additionally, we permit some hyperparameter borrowing as described earlier during this period.

4.2 Reporting Results

Each MLPerf submission has several labels: division (open or closed), category (available, preview, or research), and system type (on-Premises or cloud).

4.2.1 Submission Divisions

MLPerf has two submission divisions: closed and open. Both require that submissions employ the same data set and quality metric as the corresponding reference implementation.

The *closed* division is intended for direct system comparison, so it strives to ensure workload equivalence by requiring that submissions be equivalent to reference implementations. Equivalence includes mathematically identical model implementations, parameter initialization, optimizer and training schedules, and data processing and traversal. To ensure fairness, this division also restricts hyperparameter modification.

The *open* division is intended to encourage innovative solutions of important practical problems and to encourage hardware/software co-design. It allows submissions to employ model architectures, optimization procedures, and data augmentations that differ from the reference implementa-

tions.

4.2.2 System Categories

To allow for a broad range of research and industry systems, we defined three submission categories: available, preview, and research. These categories encourage novel techniques and systems (e.g., from academic researchers), but they also distinguish between shipping products and proof-of-concept or early engineering samples.

The *available* category imposes requirements on both hardware and software availability. Hardware must be either available for third-party rental on a cloud service or, in the case of on-premises equipment, available for purchase. Supply and lead times for renting or purchasing should benefit the system scale and company size. To ensure that benchmark submissions are widely consumable and to discourage benchmark-specific engineering, we also require that software in this category be versioned and supported for general use.

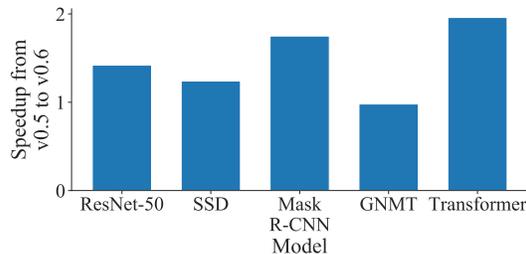
Preview systems contain components that meet the available-category criteria within 60 days of the submission date or by the next submission cycle, whichever is later. Any preview system must also be submitted to the available category by that time.

Research submissions contain components unintended for production. An example is an academic-research prototype designed as a proof of concept rather than a robust product. This category also includes systems that are built from production hardware and software but are larger in scale than available-category configurations.

4.2.3 Reporting Scale

Modern ML training spans multiple orders of magnitude in system power draw and cost. Therefore, comparisons are more useful if the reported performance includes the scale. A common scale metric, such as cost or power, is not definable across a wide range of systems (cloud, on-premises, and preproduction), so it requires differentiation by system type.

In the first two MLPerf rounds, we included the system configuration (number of processors and/or accelerators) alongside the performance scores. For on-premises examples, future versions will include a power-measurement specification. For cloud systems, we derived a “cloud-scale” metric from the number of host processors, amount of host memory, and number and type of accelerators. We empirically verified that cloud scale correlates closely with cost across three major cloud providers. Reporting of these scale metrics was optional in MLPerf v0.5 and v0.6.



(a) Speedup.

Model	Metric	v0.5	v0.6
ResNet-50	Top-1 accuracy	74.9	75.9
SSD	mAP	21.2	23
Mask R-CNN	Box / Mask min AP	37.7 / 39.9	Same
GNMT	Sacre BLEU	21.8	24
Transformer	BLEU	25	Same

(b) Quality targets.

Figure 3. Speedup in the fastest 16-chip entry from MLPerf version v0.5 to v0.6 for various benchmarks common to both (Figure 3a), along with quality-target increases (Figure 3b).

4.2.4 Reporting Scores

An MLPerf results report provides the time to train for each benchmark. Although a single summary score that spans the entire suite may be desirable for system comparisons, it is unsuited to MLPerf for two main reasons. First, a summary score implies some weighting of individual benchmark scores. Given the diversity of system users and the wide range of applications that MLPerf covers, no weighting scheme is universally representative. Second, a summary score becomes less meaningful if a submitter declines to report results on all benchmarks. Submitters can have multiple reasons for omitting some benchmarks—not all are practical at every system scale (for example, some models are untrainable at the minibatch sizes that the largest systems require for data-parallel training). Additionally, some processors may target only certain applications.

5 RESULTS

MLPerf, like all benchmarks, aims to encourage innovation through constructive competition; we measure progress by comparing results across submission rounds. We have conducted two MLPerf Training rounds thus far: v0.5 and v0.6. They were six months apart, and the underlying hardware systems were unchanged. The results that were either unmodified or underwent minor modifications between rounds show that MLPerf is driving rapid performance and scaling improvement in both the implementations and software stacks. Figure 3 shows that between the two submission rounds, the best performance results for a 16-chip system increased by an average of $1.3\times$ despite the higher

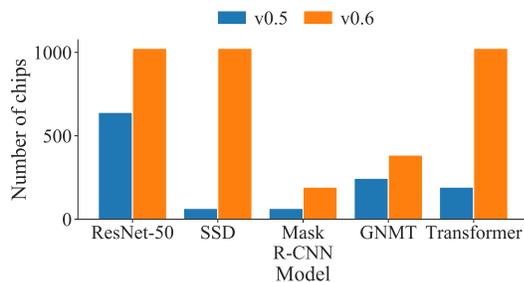


Figure 4. Number of chips necessary to produce the fastest time to solution for MLPerf versions v0.5 to v0.6. This number increased by as much as 5.5 \times .

quality targets. Figure 4 reveals that the number of chips necessary to produce the best overall performance result increased by an average of 5.5 \times . Some of this improvement owes to better benchmark implementations and some to rule changes, such as allowing the LARS (You et al., 2017) optimizer for large ResNet batches. But we believe submitters incorporated much of the performance and scaling improvements into the underlying software infrastructure and passed them on to users. We expect MLPerf to drive similar improvements through focused hardware innovation.

6 CONCLUSIONS

MLPerf Training is a suite of ML benchmarks that represent both industrial and academic use cases. In addition to being the only widely used ML-training benchmark suite boasting such coverage, it has made the following contributions:

- Precise definition of model architectures and training procedures for each benchmark. This feature enables system comparisons for equivalent workloads, whereas previous results often involved substantially different variants of a given model (for example, ResNet-50 has at least five variants).
- Reference implementations and rule definitions to address the challenges unique to benchmarking ML training. These challenges include the stochastic nature of training processes, the necessity of training to completion to determine the quality impact of performance optimizations, and the need for workload variation at different system scales (§ 2.1).

Although MLPerf focuses on relative system performance, as the online results demonstrate, it also offers general lessons about ML and benchmarking:

- Realistic data-set size is critical to ensuring realistic memory-system behavior—for example, the initial NCF data set was too small and could reside entirely in memory. Furthermore, when benchmarking data

sets that are smaller than industrial scale, training time should exclude the startup time, which would be proportionally less in actual use.

- Small hyperparameter changes can produce considerable performance changes. But, based on our experience with hyperparameter borrowing, hyperparameters are relatively portable at similar system scales, even across architectures, numerics, or software stacks.
- Frameworks exhibit subtle optimizer-algorithm variations that affect convergence.

ML is an evolving field, however, and we have much more to learn. To keep pace, MLPerf establishes a process to maintain and update the suite. For example, MLPerf v0.6 includes several updates: the ResNet-50 benchmark added LARS (You et al., 2017), GNMT’s model architecture improved to increase translation quality, and the MiniGo reference switched from Python to C++ to increase performance. The MLPerf organization welcomes input and contributions: <https://mlperf.org/get-involved>

ACKNOWLEDGEMENTS

In this section, we acknowledge all those who helped produce the first set of results or supported the overall benchmark development.

Intel: Cong Xu, Deng Xu, Feng Tian, Haihao Shen, Mingxiao Huang, Rachita Prem Seelin, Teng Lu, Xin Qiu, and Zhongyuan Wu.

Facebook: Maxim Naumov, Dheevatsa Mudigere, Mustafa Ozdal, Misha Smelyanskiy, Joe Spisak, Sy Choudhury, and Brian Gamidos.

Stanford: Work at Stanford received support in part from affiliate members and other Stanford DAWN project participants—Ant Financial, Facebook, Google, Infosys, NEC, and VMware—as well as Toyota Research Institute, Northrop Grumman, Cisco, SAP, NSF CAREER grant CNS-1651570, and NSF Graduate Research Fellowship grant DGE-1656518. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

Harvard: Work at Harvard received partial support from the Applications Driving Architectures (ADA) Research Center, a JUMP Center cosponsored by the SRC and DARPA, NSF CCF#1704834, and Intel Corporation. We would also like to thank Brandon Reagen.

University of Toronto: Work at the University of Toronto received partial support from an NSERC Discovery grant, the Canada Foundation for Innovation JELF grant, the Connaught Fund, and Huawei grants.

REFERENCES

- AI Matrix. URL <https://aimatrix.ai>.
- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. TensorFlow: A System for Large-Scale Machine Learning. In *OSDI*, volume 16, pp. 265–283, 2016.
- Adolf, R., Rama, S., Reagen, B., Wei, G.-Y., and Brooks, D. Fathom: Reference Workloads for Modern Deep Learning Methods. In *Workload Characterization (IISWC), 2016 IEEE International Symposium on*, pp. 1–10. IEEE, 2016.
- Amodei, D. and Hernandez, D. AI and Compute, 2018. URL <https://blog.openai.com/ai-and-compute/>.
- Auer, P., Herbster, M., and Warmuth, M. K. Exponentially Many Local Minima for Single Neurons. In *Advances in neural information processing systems*, pp. 316–322, 1996.
- Bai, J., Lu, F., Zhang, K., et al. ONNX: Open Neural Network Exchange. <https://github.com/onnx/onnx>, 2019.
- Baidu. DeepBench: Benchmarking Deep Learning Operations on Different Hardware. <https://github.com/baidu-research/DeepBench>, 2017.
- Banner, R., Hubara, I., Hoffer, E., and Soudry, D. Scalable Methods for 8-bit Training of Neural Networks. In *Advances in Neural Information Processing Systems*, pp. 5145–5153, 2018.
- Belletti, F., Lakshmanan, K., Krichene, W., Chen, Y.-F., and Anderson, J. Scalable Realistic Recommendation Datasets through Fractal Expansions. *arXiv preprint arXiv:1901.08910*, 2019.
- Ben-Nun, T., Besta, M., Huber, S., Ziogas, A. N., Peter, D., and Hoefler, T. A Modular Benchmarking Infrastructure for High-Performance and Reproducible Deep Learning. *arXiv preprint arXiv:1901.10183*, 2019.
- Chan, B. OpenAI Five, Jun 2018. URL <https://openai.com/blog/openai-five/>.
- Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C., and Zhang, Z. MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems. *arXiv preprint arXiv:1512.01274*, 2015.
- Chen, T., Moreau, T., Jiang, Z., Zheng, L., Yan, E., Shen, H., Cowan, M., Wang, L., Hu, Y., Ceze, L., et al. {TVM}: An Automated End-to-End Optimizing Compiler for Deep Learning. In *13th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 18)*, pp. 578–594, 2018.
- Cheng, H.-T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W., Ispir, M., et al. Wide & Deep Learning for Recommender Systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pp. 7–10. ACM, 2016.
- Chetlur, S., Woolley, C., Vandermersch, P., Cohen, J., Tran, J., Catanzaro, B., and Shelhamer, E. CuDNN: Efficient Primitives for Deep Learning. *arXiv preprint arXiv:1410.0759*, 2014.
- Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., and LeCun, Y. The Loss Surfaces of Multilayer Networks. In *Artificial Intelligence and Statistics*, pp. 192–204, 2015.
- Coleman, C., Narayanan, D., Kang, D., Zhao, T., Zhang, J., Nardi, L., Bailis, P., Olukotun, K., Ré, C., and Zaharia, M. DAWN Bench: An End-to-End Deep Learning Benchmark and Competition. *NIPS ML Systems Workshop*, 2017.
- Coleman, C., Kang, D., Narayanan, D., Nardi, L., Zhao, T., Zhang, J., Bailis, P., Olukotun, K., Ré, C., and Zaharia, M. Analysis of DAWN Bench, a Time-to-Accuracy Machine Learning Performance Benchmark. *ACM SIGOPS Operating Systems Review*, 53(1):14–25, 2019.
- Council, T. P. P. Transaction Processing Performance Council. *Web Site*, <http://www.tpc.org>, 2005.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-scale Hierarchical Image Database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Dixit, K. M. The SPEC Benchmarks. *Parallel computing*, 17(10-11):1195–1209, 1991.
- Dongarra, J. The LINPACK Benchmark: An Explanation. In *Proceedings of the 1st International Conference on Supercomputing*, pp. 456–474, London, UK, UK, 1988. Springer-Verlag. ISBN 3-540-18991-2. URL <http://dl.acm.org/citation.cfm?id=647970.742568>.
- Google. TensorFlow Benchmarks. <https://www.tensorflow.org/performance/benchmarks>, 2017.

- Gori, M. and Tesi, A. On the Problem of Local Minima in Backpropagation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (1):76–86, 1992.
- Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. *arXiv preprint arXiv:1706.02677*, 2017.
- GroupLens. MovieLens 20M Dataset, Oct 2016. URL <https://grouplens.org/datasets/movielens/20m/>.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016a.
- He, K., Zhang, X., Ren, S., and Sun, J. Identity Mappings in Deep Residual Networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016b.
- He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017a.
- He, X., Liao, L., Zhang, H., Nie, L., Hu, X., and Chua, T.-S. Neural Collaborative Filtering. In *Proceedings of the 26th international conference on world wide web*, pp. 173–182. International World Wide Web Conferences Steering Committee, 2017b.
- Hennessy, J. L. and Patterson, D. A. *Computer Architecture: A Quantitative Approach*. Elsevier, 2011.
- Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Kingsbury, B., et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal processing magazine*, 29, 2012.
- Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., and Murphy, K. Speed/Accuracy Trade-offs for Modern Convolutional Object Detectors, 2016.
- Intel. BigDL: Distributed Deep Learning Library for Apache Spark, 2019. URL <https://github.com/intel-analytics/BigDL>.
- Jia, X., Song, S., He, W., Wang, Y., Rong, H., Zhou, F., Xie, L., Guo, Z., Yang, Y., Yu, L., et al. Highly Scalable Deep Learning Training System with Mixed-Precision: Training ImageNet in Four Minutes. *arXiv preprint arXiv:1807.11205*, 2018.
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. Caffe: Convolutional Architecture for Fast Feature Embedding. In *ACM International Conference on Multimedia*, pp. 675–678. ACM, 2014.
- Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., Bates, S., Bhatia, S., Boden, N., Borchers, A., et al. In-Datcenter Performance Analysis of a Tensor Processing Unit. In *2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)*, pp. 1–12. IEEE, 2017.
- Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. *ICLR*, 2015.
- Krizhevsky, A. One Weird Trick for Parallelizing Convolutional Neural Networks, 2014.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.
- Kster, U., Webb, T. J., Wang, X., Nassar, M., Bansal, A. K., Constable, W. H., Elibol, O. H., Gray, S., Hall, S., Hornof, L., Khosrowshahi, A., Kloss, C., Pai, R. J., and Rao, N. Flexpoint: An Adaptive Numerical Format for Efficient Training of Deep Neural Networks. *NIPS*, 2017.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. SSD: Single Shot Multibox Detector. In *European conference on computer vision*, pp. 21–37. Springer, 2016.
- Markidis, S., Der Chien, S. W., Laure, E., Peng, I. B., and Vetter, J. S. NVIDIA Tensor Core Programmability, Performance & Precision. *arXiv preprint arXiv:1803.04014*, 2018.
- Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., and Wu, H. Mixed Precision Training. In *Proceedings of the International Conference on Learning Representations*, 2018.
- Mikami, H., Suganuma, H., U-chupala, P., Tanaka, Y., and Kageyama, Y. Massively Distributed SGD: ImageNet/ResNet-50 Training in a Flash. *arXiv preprint arXiv:1811.05233*, 2018.
- MLPerf. MLPerf Reference: MiniGo. <https://github.com/mlperf/training/tree/master/reinforcement>, 2019a.

- MLPerf. MLPerf Reference: ResNet in TensorFlow. https://github.com/mlperf/training/tree/master/image_classification/tensorflow/official, 2019b.
- Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing Atari with Deep Reinforcement Learning. *arXiv preprint arXiv:1312.5602*, 2013.
- Narayanan, D., Harlap, A., Phanishayee, A., Seshadri, V., Devanur, N. R., Ganger, G. R., Gibbons, P. B., and Zaharia, M. PipeDream: Generalized Pipeline Parallelism for DNN Training. In *Proceedings of the 27th ACM Symposium on Operating Systems Principles*, pp. 1–15, 2019.
- Naumov, M., Mudigere, D., Shi, H.-J. M., Huang, J., Sundaraman, N., Park, J., Wang, X., Gupta, U., Wu, C.-J., Azzolini, A. G., et al. Deep Learning Recommendation Model for Personalization and Recommendation Systems. *arXiv preprint arXiv:1906.00091*, 2019.
- Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic Differentiation in PyTorch. 2017.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language Models are Unsupervised Multi-task Learners. *OpenAI Blog*, 1(8), 2019.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the Game of Go with Deep Neural Networks and Tree Search. *nature*, 529(7587):484, 2016.
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the Game of Go without Human Knowledge. *Nature*, 550(7676):354, 2017.
- Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. A General Reinforcement Learning Algorithm that masters Chess, Shogi, and Go through Self-Play. *Science*, 362(6419):1140–1144, 2018.
- Sun, P., Feng, W., Han, R., Yan, S., and Wen, Y. Optimizing Network Performance for Distributed DNN Training on GPU Clusters: ImageNet/AlexNet Training in 1.5 Minutes. *arXiv preprint arXiv:1902.06855*, 2019.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is All You Need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.
- WMT. First Conference on Machine Translation, 2016. URL <http://www.statmt.org/wmt16/>.
- WMT. Second Conference on Machine Translation, 2017. URL <http://www.statmt.org/wmt17/>.
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144*, 2016.
- Ying, C., Kumar, S., Chen, D., Wang, T., and Cheng, Y. Image Classification at Supercomputer Scale. *arXiv preprint arXiv:1811.06992*, 2018.
- You, Y., Gitman, I., and Ginsburg, B. Large Batch Training of Convolutional Networks. *arXiv preprint arXiv:1708.03888*, 2017.
- Zhou, G., Zhu, X., Song, C., Fan, Y., Zhu, H., Ma, X., Yan, Y., Jin, J., Li, H., and Gai, K. Deep Interest Network for Click-through Rate Prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1059–1068. ACM, 2018.
- Zhu, C., Han, S., Mao, H., and Dally, W. J. Trained Ternary Quantization. *arXiv preprint arXiv:1612.01064*, 2016.
- Zhu, H., Akrouf, M., Zheng, B., Pelegris, A., Jayarajan, A., Phanishayee, A., Schroeder, B., and Pekhimenko, G. Benchmarking and Analyzing Deep Neural Network Training. In *2018 IEEE International Symposium on Workload Characterization (IISWC)*, pp. 88–100. IEEE, 2018.

A ARTIFACT APPENDIX

A.1 Abstract

This artifact description contains information about the complete workflow to reproduce Nvidia's v0.5 image classification submissions to MLPerf. We describe how to run this submission on a single-node DGX-1 system. More details for DGX-2 and multi-node systems are provided in the official MLPerf results repositories:

- Nvidia's v0.5 ResNet-50 submissions

Results from other tasks and submitters are also available:

- MLPerf v0.5 training results
- MLPerf v0.6 training results

However, these results have not been independently verified for reproducibility. Please see the MLPerf website (<https://mlperf.org/>) for the most up-to-date information and feel free to report issues on Github.

A.2 Artifact check-list (meta-information)

- **Algorithm:** Image classification ResNet-50 CNN
- **Program:** MLPerf (<https://mlperf.org/>)
- **Compilation:** nvidia-docker
- **Model:** ResNet-50 v1.5 ³
- **Data set:** ImageNet (<http://image-net.org/>)
- **Hardware:** NVIDIA DGX-1 or DGX-2
- **Metrics:** Time-to-Train: minutes to reach accuracy threshold (74.9% Top-1 for v0.5)
- **Output:** MLPerf compliant log file with timestamps and evaluation accuracy. Execution ends once the accuracy threshold is reached.
- **Experiments:** shell script included with the code (./run.sub)
- **How much disk space required (approximately)?:** 300 GB
- **How much time is needed to prepare workflow (approximately)?:** 2 hours
- **How much time is needed to complete experiments (approximately)?:** 8 hours
- **Publicly available:** Yes
- **Code licenses:** Apache License 2.0
- **Workflow framework used?:** MXNet
- **Archived (provide DOI)?:**
<http://doi.org/10.5281/zenodo.3610717>

³https://github.com/mlperf/training/tree/master/image_classification/tensorflow/official

A.3 Description

A.3.1 How to access

MLPerf v0.5 training results on Github:
https://github.com/mlperf/training_results_v0.5.

A.4 Installation

See the README.md for Nvidia's v0.5 ResNet-50 submission:
https://github.com/mlperf/training_results_v0.5/tree/master/v0.5.0/nvidia/submission/code/image_classification/mxnet/README.md.

A.5 Evaluation and expected result

Time-to-Train: 134.6 minutes.