

# Benchmarking Neural Network Training Algorithms

<b>George E. Dahl</b> <sup>1*</sup>	GDAHL@GOOGLE.COM
<b>Frank Schneider</b> <sup>2*</sup>	F.SCHNEIDER@UNI-TUEBINGEN.DE
<b>Zachary Nado</b> <sup>1*</sup>	ZNADO@GOOGLE.COM
<b>Naman Agarwal</b> <sup>1*</sup>	NAMANAGARWAL@GOOGLE.COM
<b>Chandramouli Shama Sastry</b> <sup>3,4†</sup>	CHANDRAMOULI.SASTRY@GMAIL.COM
<b>Philipp Hennig</b> <sup>2†</sup>	PHILIPP.HENNIG@UNI-TUEBINGEN.DE
<b>Sourabh Medapati</b> <sup>1†</sup>	SMEDAPATI@GOOGLE.COM
<b>Runa Eschenhagen</b> <sup>2†</sup>	RE393@CAM.AC.UK
<b>Priya Kasimbeg</b> <sup>1†</sup>	KASIMBEG@GOOGLE.COM
<b>Daniel Suo</b> <sup>1†</sup>	DSUO@GOOGLE.COM
<b>Juhan Bae</b> <sup>3,5†</sup>	JBAE@CS.TORONTO.EDU
<b>Justin Gilmer</b> <sup>1†</sup>	GILMER@GOOGLE.COM
<b>Abel L. Peirson</b> <sup>6†</sup>	ALPV95@STANFORD.EDU
<b>Bilal Khan</b> <sup>1†</sup>	KBILAL@GOOGLE.COM
<b>Rohan Anil</b> <sup>1†</sup>	ROHANANIL@GOOGLE.COM
<b>Mike Rabbat</b> <sup>7†</sup>	MIKERABBAT@META.COM
<b>Shankar Krishnan</b> <sup>1†</sup>	SKRISHNAN@GOOGLE.COM
<b>Daniel Snider</b> <sup>3,5‡</sup>	DANS@CS.TORONTO.EDU
<b>Ehsan Amid</b> <sup>1†</sup>	EAMID@GOOGLE.COM
<b>Kongtao Chen</b> <sup>1†</sup>	KONGTAO@GOOGLE.COM
<b>Chris J. Maddison</b> <sup>3,5‡</sup>	CMADDIS@CS.TORONTO.EDU
<b>Rakshith Vasudev</b> <sup>8‡</sup>	RAKSHITH.VASUDEV@DELL.COM
<b>Michal Badura</b> <sup>1†</sup>	MBADURA@GOOGLE.COM
<b>Ankush Garg</b> <sup>1†</sup>	ANKUGARG@GOOGLE.COM
<b>Peter Mattson</b> <sup>1†</sup>	PETERMATTSON@GOOGLE.COM

<sup>1</sup>Google; <sup>2</sup>University of Tübingen; <sup>3</sup>Vector Institute; <sup>4</sup>Dalhousie University; <sup>5</sup>University of Toronto; <sup>6</sup>Stanford University; <sup>7</sup>Meta AI (FAIR); <sup>8</sup>Dell Technologies

---

\*. Corresponding authors.

†. These authors are listed in random order.

‡. These authors are listed in random order.

## Abstract

Training algorithms, broadly construed, are an essential part of every deep learning pipeline. Training algorithm improvements that speed up training across a wide variety of workloads (e.g., better update rules, tuning protocols, learning rate schedules, or data selection schemes) could save time, save computational resources, and lead to better, more accurate, models. Unfortunately, as a community, we are currently unable to reliably identify training algorithm improvements, or even determine the state-of-the-art training algorithm. In this work, using concrete experiments, we argue that real progress in speeding up training requires new benchmarks that resolve three basic challenges faced by empirical comparisons of training algorithms: (1) how to decide when training is complete and precisely measure training time, (2) how to handle the sensitivity of measurements to exact workload details, and (3) how to fairly compare algorithms that require hyperparameter tuning. In order to address these challenges, we introduce a new, competitive, time-to-result benchmark using multiple workloads running on fixed hardware, the ALGOPERF: TRAINING ALGORITHMS benchmark. Our benchmark includes a set of workload variants that make it possible to detect benchmark submissions that are more robust to workload changes than current widely-used methods. Finally, we evaluate baseline submissions constructed using various optimizers that represent current practice, as well as other optimizers that have recently received attention in the literature. These baseline results collectively demonstrate the feasibility of our benchmark, show that non-trivial gaps between methods exist, and set a provisional state-of-the-art for future benchmark submissions to try and surpass.

**Keywords:** benchmark, deep learning, neural network, training algorithms, optimizer

## Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
1.1	Contributions of This Work . . . . .	6
1.2	Training Algorithm Benchmark Goals and Scope . . . . .	7
<b>2</b>	<b>The Challenges of Empirical Comparisons of Training Algorithms</b>	<b>8</b>
2.1	Precisely Defining and Measuring Training Speed . . . . .	9
2.2	Dependence on the Workload . . . . .	10
2.2.1	Sensitivity of Optimizer Ranking to the Model Architecture . . . . .	11
2.2.2	Implications of Workload Sensitivity . . . . .	14
2.3	We Cannot Compare Families of Training Algorithms, Only Specific Instances	15
2.3.1	Training Algorithms With Different Hyperparameters . . . . .	15
2.3.2	Training Algorithms With Different Hyperparameter Search Spaces . . . . .	16
2.3.3	Training Algorithms With Different Tuning Goals . . . . .	18
2.4	Strong Baselines Are Far Too Hard to Obtain . . . . .	20
<b>3</b>	<b>Related Work</b>	<b>21</b>
3.1	Existing Benchmarks . . . . .	21
3.2	Methodological Critiques of Training Algorithm Comparisons . . . . .	23
3.3	Disagreement over Training Algorithms: The Case for Clear Benchmarks . . . . .	23
<b>4</b>	<b>Rules</b>	<b>25</b>
4.1	A Time-to-Result Benchmark . . . . .	25

4.1.1	Measuring Runtime	27
4.1.2	Standardizing Benchmarking Hardware	27
4.2	Specifying a Training Algorithm	28
4.2.1	Isolate the Training Algorithm	29
4.2.2	Incentivizing Generally Useful Submissions	30
4.3	Workloads	31
4.3.1	Fixed and Randomized Workloads	31
4.4	Tuning	33
4.4.1	External Tuning	34
4.4.2	Self-Tuning	34
4.5	Scoring and Reporting Results	35
4.5.1	Aggregation Using Performance Profiles	35
4.5.2	Integrating Performance Profiles for the Benchmark Score	36
4.5.3	Using Held-Out Workloads in Scoring	37
4.5.4	Measuring Year-Over-Year Benchmark Progress	37
<b>5</b>	<b>Target-Setting Experiments</b>	<b>38</b>
<b>6</b>	<b>Randomized Workloads Experiments</b>	<b>42</b>
6.1	Desiderata for Workload Variants	43
6.2	Creating and Testing Workload Variants	44
6.3	Workload Variants of the Benchmark	46
<b>7</b>	<b>Baseline Submissions</b>	<b>46</b>
7.1	Baseline Creation Procedure	48
7.2	Baseline Timing	49
7.3	Baseline Results	51
7.3.1	Baseline Results Comparing Search Spaces	53
<b>8</b>	<b>Discussion</b>	<b>56</b>
8.1	Target Setting	56
8.2	Randomized Workloads	58
8.3	Baselines	58
8.4	Benchmark Limitations	59
8.5	Future Work	62
<b>9</b>	<b>Conclusion</b>	<b>63</b>
	<b>Appendices</b>	<b>68</b>
	<b>Appendix A Experimental Details for Section 2</b>	<b>68</b>
A.1	Learning Rate Schedules	68
A.1.1	Warmup Cosine Decay	68
A.1.2	Warmup Linear Decay Constant	69
A.2	Details for Training Curves that Cross	69
A.3	Details for Sensitivity of Optimizer Ranking to the Model Architecture	70

A.3.1	WIDE RESNET with Stride Changes . . . . .	70
A.3.2	Architectural Modifications of Transformer Models . . . . .	70
A.4	Details for Comparing Instances of Training Algorithms . . . . .	70
A.4.1	Training Algorithms with Different Hyperparameters . . . . .	70
A.4.2	Training Algorithms with Different Hyperparameter Search Spaces . . . . .	74
A.4.3	Training Algorithms with Different Tuning Goals . . . . .	74
<b>Appendix B</b>	<b>Details for Target-Setting Experiments (Section 5)</b>	<b>74</b>
<b>Appendix C</b>	<b>Details for Baseline Experiments (Section 7)</b>	<b>77</b>
<b>Appendix D</b>	<b>Workload Details</b>	<b>77</b>
D.1	CRITEO 1TB . . . . .	77
D.1.1	DLRMSMALL Model . . . . .	82
D.1.2	CRITEO 1TB DLRMSMALL Workload Variants . . . . .	82
D.2	FASTMRI . . . . .	82
D.2.1	U-NET Model . . . . .	83
D.2.2	FASTMRI U-NET Workload Variants . . . . .	84
D.3	IMAGENET . . . . .	84
D.3.1	RESNET-50 Model . . . . .	84
D.3.2	IMAGENET RESNET-50 Workload Variants . . . . .	85
D.3.3	VISION TRANSFORMER Model . . . . .	85
D.3.4	IMAGENET VISION TRANSFORMER Workload Variants . . . . .	86
D.4	LIBRISPEECH . . . . .	86
D.4.1	CONFORMER Model . . . . .	87
D.4.2	LIBRISPEECH CONFORMER Workload Variants . . . . .	87
D.4.3	DEEPSPEECH Model . . . . .	87
D.4.4	LIBRISPEECH DEEPSPEECH Workload Variants . . . . .	88
D.5	OGBG . . . . .	88
D.5.1	GNN Model . . . . .	89
D.5.2	OGBG GNN Workload Variants . . . . .	89
D.6	WMT . . . . .	90
D.6.1	TRANSFORMER Model . . . . .	90
D.6.2	WMT TRANSFORMER Workload Variants . . . . .	91
<b>Appendix E</b>	<b>Preliminary Experiments for Randomized Workloads</b>	<b>91</b>
<b>References</b>		<b>94</b>

## 1. Introduction

Although artificial neural networks are extremely useful models, training them remains quite expensive. Moreover, investing more time and computational resources in training produces better, more accurate models. For example, training larger models, training longer, training on more data, or performing more exploratory experiments can all improve results. Training more efficiently would directly reduce costs and/or indirectly produce more accurate models. Although there are many ways to make training more efficient, in this work, we restrict our attention to improvements in training *algorithms*.

Unfortunately, as a community, we are currently unable to identify which existing training algorithms are best, let alone understand what novel methods are the most promising. A multitude of training algorithms have been proposed, and more are proposed every year. [Schmidt et al. \(2021\)](#) lists well over a hundred methods, mostly published in the last seven years. Naturally, each of these papers claims that their algorithm has significant benefits, but the vast majority of the deep learning community never uses any of these techniques. At present, most training algorithms are assumed to not be useful until they have been widely adopted, creating a chicken-and-egg problem. This state of affairs should be troubling for anyone trying to train neural networks or develop new training algorithms.

To provide actionable guidance to practitioners, we need to understand how existing and future methods perform in practice. Additionally, the community needs to incentivize research that makes actual progress on training algorithms. Currently, the paper that introduces a new technique is also the primary source of experimental evidence for its merits, which causes a conflict of interest if the researchers proposing the techniques also have control over the baselines included for comparison.

Although comparative studies and meta-analyses are effective ways to assess the merits of competing techniques in principle, so far, they have not been able to provide definitive answers. One reason may be that such studies are themselves often subject to criticism and accusations of bias: For any concrete empirical comparison, the proponents of any particular technique can always find fault with the details of the setup. Without broad agreement on delicate issues such as hyperparameter tuning, in many cases, such criticism may indeed be justified. Another issue is that the studies' retrospective nature shifts too much of the burden of proof away from the original inventors of training algorithms. Although a useful tool, empirical studies ultimately face an uphill battle to generate strong enough evidence to actually convince practitioners, especially since the truth can be quite complicated.

Instead of relying entirely on retrospective comparative studies, why not have researchers submit their methods to a common competitive evaluation? Such competitive benchmarks have historically worked well. For example, the deep learning community tried for years to convince the computer vision community that neural networks were better models for image classification by publishing papers, but deep learning was only widely adopted for computer vision tasks after neural networks showed success in high-profile competitive benchmarks (e.g., [Russakovsky et al., 2015](#)). Similarly, if we can construct benchmarks that capture useful notions of training speed in sufficiently realistic conditions, we could determine whether widely-used training algorithms are indeed the best currently available. If not, we could demonstrate that current incumbent methods should be replaced with better alternatives.

The mission of the MLCOMMONS<sup>®</sup><sup>1</sup> Algorithms Working Group is to create a set of rigorous and relevant benchmarks to measure neural network training speedups due to algorithmic improvements. The best benchmarks will capture what is needed to drive progress at a particular time in the field and continue evolving along with the needs of the community. This paper describes the working group’s first attempt to benchmark training algorithms for neural networks, and we intend to continue releasing improved versions periodically. We hope that anyone interested in improving upon the benchmark we propose here will consider joining the working group to help shape future versions.

### 1.1 Contributions of This Work

1. We precisely articulate the challenges of benchmarking training algorithms, provide concrete experiments demonstrating the methodological issues we identify, and explain how they hold back training algorithms research (Section 2).
2. We introduce the ALGOPERF: TRAINING ALGORITHMS benchmark — a competitive, time-to-result benchmark on multiple workloads running on fixed hardware for systematically comparing training algorithms (Section 4).
  - (a) Our benchmark defines a complete and workable procedure for setting (validation and test error) targets and measuring training time to reach them. Furthermore, this procedure produces targets that are reasonably competitive with results in the literature, given the resource constraints of the benchmark (Section 5).
  - (b) Our benchmark incentivizes generally useful training algorithms by computing a joint score across all workloads and by including randomized workloads to simulate novel problems.
  - (c) Our benchmark requires submissions to explicitly account for the hyperparameter tuning they need to achieve their results across workloads, giving submissions that are easier to tune an advantage. Submissions can either compete using limited, parallel tuning resources, or enter as a completely self-tuning and hyperparameter-free algorithm.
  - (d) We specify a set of benchmark workloads covering image classification, speech recognition, machine translation, MRI reconstruction, click-through rate prediction, and chemical property prediction tasks, and we design workload variants to challenge training algorithms to be more robust to natural workload modifications.
  - (e) We provide open-source JAX and PYTORCH implementations of all workloads, and a training algorithm API that supports submissions in both frameworks. By providing code along with details of the benchmark system, we make it easy to independently reproduce results on the benchmark.
3. We construct baselines by defining search spaces for eight popular optimizers (ADAMW, NADAMW, HEAVY BALL, NESTEROV, LAMB, ADAFACTOR, SAM(w. ADAM), DIS-

---

1. MLCOMMONS<sup>®</sup> and MLPERF<sup>™</sup> are registered and unregistered trademarks, respectively, of the ML-Commons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited.

TRIBUTED SHAMPOO) that includes both popular optimizers that represent current practice and methods that have received attention in the recent literature (Section 7).

- (a) Collectively, these baselines demonstrate that our benchmark is feasible and that there is a non-trivial gap between different methods.
  - (b) We show that our benchmark score successfully favors algorithms that are easier to tune and that successful training algorithms require good search spaces that are tailored to the tuning budget. Specifically, baselines using adaptive methods (ADAMW and NADAMW) score more highly than baselines using non-adaptive methods (HEAVY BALL and NESTEROV), in large part because of the difficulty of constructing good search spaces for the latter.
  - (c) We set a provisional state of the art on the benchmark under our external tuning ruleset based on extensive experiments including tens of thousands of tuning trials. NADAMW with our search space performs well on every (fixed) workload.
4. We construct a set of 24 workload variants (three for each fixed workload) that make it possible to detect improvements in robustness over current popular methods, and demonstrate the need for additional research on hyperparameter transfer (Section 6).

**How to read this paper.** This paper serves a dual purpose: it acts both as a research report on benchmarking neural network training algorithms, and as a form of technical documentation of the ALGOPERF: TRAINING ALGORITHMS benchmark and its rules. Readers who plan to participate in the competition may wish to start by reading the sections that describe the rules (Section 4), using them as a higher-level companion to the latest version of the complete rules that can be found online, and then move to the experiments performed on this benchmark (Sections 5 to 7). Readers who are primarily interested in the research aspects of the paper may wish to skip, or only briefly skim, the rules section (Section 4) during an initial read, with the exception of Section 4.5 which is essential to understanding the scores reported in Section 7.

## 1.2 Training Algorithm Benchmark Goals and Scope

In order to construct a benchmark for general neural network training algorithms that measures training speed, we need to decide what we mean by a *general training algorithm*, and how we measure training *speed*.

**Design goal.** We aim to encourage *general-purpose* training algorithms that are easy to apply across different data modalities and model architectures. It is legitimate to wonder what the best training algorithm is for a particular workload, e.g. “RESNET-50 on IMAGENET,” or even a single workload family, e.g. “convolutional neural net image classifiers.” But given the still rapid rate of change among architectures in the field, it is more promising for the community to develop relatively general methods first. A downside of this design choice is that it implicitly assumes a clean separation between model and algorithm.

**Separating “algorithm” and “model”.** We adopt the paradigm of machine learning as optimization; i.e., we view the optimizer and the model as two separate modules and focus on the first. Nevertheless, we must consider more pieces of the training pipeline

than just the optimizer because the separation can be subtle. For example, should batch normalization (Ioffe and Szegedy, 2015) be considered an aspect of the training algorithm or the modeling? An ideal benchmark should allow competitors a maximum of innovation and creativity while still providing enough constraints to make fair comparisons and glean useful insights from the results. Section 4.2 explains how we delineate what constitutes a training algorithm and, thus, the design domain of competition entries. In general, we aim to offer flexibility while enforcing generality: We allow competitors to choose optimization algorithms, update rules, and control regularization. We also allow algorithms that reorder or re-sample the training data. But we restrict or prohibit methods that only apply to specific model architectures or data modalities.

**Measuring training performance.** Although this work, like the majority of the community, uses the metaphor of optimization, optimization and *learning* are, of course, not technically the same. We ultimately care about how quickly a method reaches a satisfactory out-of-sample error, as estimated by a validation or test set error rate. To incentivize practically useful techniques, success in our competition must be based on this end goal, not on how quickly an algorithm can actually minimize the training objective. Unfortunately, this forces us to confront the complexities of regularization and overfitting in the benchmark (which are also present, of course, in any real application). Section 4.1 details how we define “time to result”. In short, we define a target out-of-sample error rate for every workload and score submissions based on how quickly they reach these targets.

**Internal and external hyperparameter tuning.** Comparing generally-applicable neural network training algorithms requires careful attention to the rules around hyperparameter tuning for benchmark submissions. In particular, should certain types of tuning be viewed as integral parts of the training algorithm or external to the training algorithm? At the moment, choosing hyperparameters is still largely the job of the user, so algorithms should perform well under an external tuning schedule. However, methods that do not require external tuning at all are, of course, desirable, and internal parameter tuning is arguably among the biggest opportunities for improvement in resource use. So we also need an opportunity for methods that effectively tune themselves to shine. Section 4.4 describes our approach to hyperparameter tuning in detail.

## 2. The Challenges of Empirical Comparisons of Training Algorithms

An ideal empirical comparison of training algorithms should be (as much as possible) convincing, informative, and practically relevant. A convincing comparison generates high quality empirical evidence, makes “fair” comparisons to strong baselines, and is not misleading. An informative comparison disentangles the causes of any measured improvements and provides insight into the observed phenomena. A practically relevant comparison measures situations that are likely to arise in important applications and studies conditions as similar as possible to current practice in applied work. Unfortunately, several basic challenges make achieving these desiderata far from simple.



## 2.1 Precisely Defining and Measuring Training Speed

Currently, papers proposing training algorithms for deep neural networks tend to shy away from making quantitative, *empirical* claims, although some will make theoretical convergence rate claims that depend on assumptions that preclude them from being directly applicable to neural network training (e.g., assuming a convex loss function). For example, in abstracts we see phrases such as “frequently delivers faster convergence” (Lucas et al., 2019), “outperforms other methods with fast convergence and high accuracy” (Zhuang et al., 2020), or “works well in practice and compares favorably to other stochastic optimization methods” (Kingma and Ba, 2015). In contrast, deep learning modeling papers usually make precise quantitative claims. For example, in abstracts we see language such as “achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results” (Vaswani et al., 2017), “achieves state-of-the-art 84.3% top-1 accuracy on IMAGENET, while being  $8.4\times$  smaller and  $6.1\times$  faster on inference” (Tan and Le, 2019), or “achieves 86.2% top-1 IMAGENET accuracy, while being  $4.7\times$  faster” (Bello et al., 2021).<sup>2</sup> Some examples of recent optimizer papers that make quantitative claims are Chen et al. (2023), Xie et al. (2022), and Tian and Parikh (2022), but these claims are still using different metrics or are referring to different benchmarks or workload versions.

To the detriment of progress on training algorithms, a lack of quantitative claims means there is no clear notion of the “state of the art”, only what is popular or topical. Most training algorithm papers will display the training curve (loss vs time) for their method and some baseline, but without a shared understanding of how such curves should be converted into quantitative measurements of training speed, they will only provide an illusory sense of precision. Even valiant attempts to make quantitative sense of such a plot will produce claims tortured with caveats. For example, Tian and Parikh (2022) claims their method’s loss is “always significantly lower beyond [the first] 30% of the training procedure.”

The root cause of the preponderance of vague training speed claims in the literature is that a direct comparison of two training curves is ill-posed. Certainly, if one curve is strictly below the other, we can say it is “better”, but even then, it is not clear by how *much*. Real-world training curves are noisy and tend to cross each other, often multiple times. Figure 1 (left) shows two sample validation error curves for RESNET-50 trained on IMAGENET, achieving a final validation error of 24.0% (—) and 24.4% (—), respectively (see Appendix A.2 for exact experiment details). These curves intersect multiple times, even after ignoring the early part of training. Even more importantly, the curves of the best validation error seen so far (Figure 1, right) also intersect multiple times, showing that which run is leading the race swaps back and forth. The run that trains RESNET-50 the fastest depends on what it means for training to be complete.

Consequently, to measure the training time, we need a clear criterion that indicates when training is complete. There can be many possible criteria, and it is not obvious which is the correct one. As a result, many papers avoid the issue altogether, relying instead on more vague descriptions of training results. Straight-forward approaches, such as computing the area under the validation loss curve, fail to capture what is relevant in practical applications. For example, improvements that affect only the initial stage of training are irrelevant in

---

2. Ironically, Bello et al. (2021) showed that the gains observed in Tan and Le (2019) were largely due to an improved training procedure, even though Tan and Le (2019) was presented as a modeling paper.

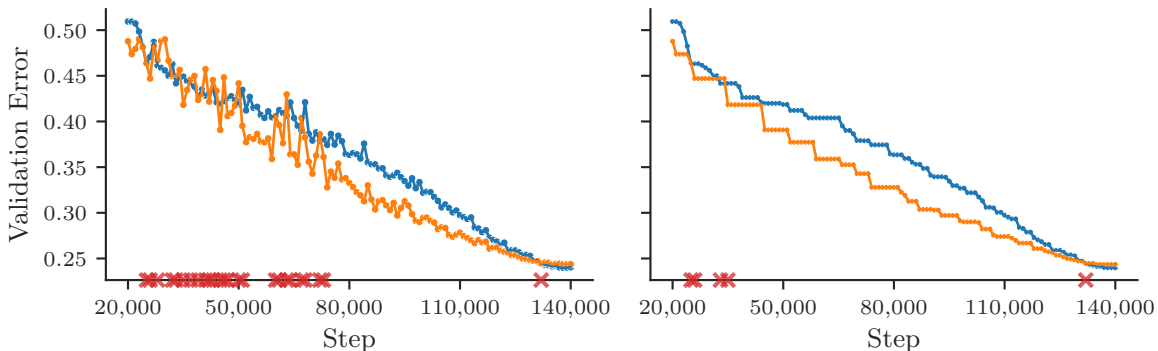


Figure 1: **A direct comparison of training curves is ill-posed if they intersect.** *Left:* The validation error for two different runs (—, —) of ADAM on RESNET-50 on IMAGENET. *Right:* The best validation error obtained so far by each curve (i.e. the running minimum of the left subplot). Even after removing the early part of training and looking at the best result so far (*right*), the curves intersect multiple times (highlighted by red crosses (X) at the bottom of each plot). Through much of the plot, the orange curve is winning, but the blue curve overtakes it near the end of the runs.

most cases. Despite these difficulties, a precise definition of training speed is essential. If the community can standardize how training time should be measured, we can shift the literature towards clear quantitative comparisons between training algorithms.

The simplest approach is to set a target error rate and measure how long it takes to first achieve this target. Specifically, we propose measuring training time based on how long it takes a run to reach a workload-specific, out-of-sample error target that is selected to be a competitive, near-state-of-the-art error rate. For any given comparison between algorithms, standardizing hardware is also a prerequisite for measuring wall-clock time, although perhaps it would be sufficient to standardize a cost model for primitive operations. Determining an appropriate target error rate for a particular workload is not trivial. [Section 4.1](#) describes our approach to setting error rate targets. Targets for a given workload may need to be revised as new results emerge, but they should be more stable than the state-of-the-art result on the corresponding dataset since a workload freezes a specific model as well. Ultimately, targets will always be somewhat arbitrary, but that is all the more reason they should be standardized *before* comparing training algorithms. Our general philosophy is to try and pick the best error rates we can consistently and reliably achieve with standard methods.

## 2.2 Dependence on the Workload

The performance of a training algorithm is necessarily a function of the workload it is timed on. Ultimately, in applications, we are interested in finding the method best suited for the particular workload in front of us. In general, this will be a novel workload without many previous results. The more we can generalize from previous results on related workloads, the better we can prioritize what to try. For generic methods research, on the other hand, we are interested in characterizing how methods perform across a diverse set of workloads representative of current important applications. Once again, we cannot measure every

possible method even on a single workload, let alone on all possible workloads, so we are forced to extrapolate. When building generic methods, we need to detect when a method provides a generally useful improvement and thus we also need to find a meaningful way to aggregate performance across multiple workloads.

### 2.2.1 SENSITIVITY OF OPTIMIZER RANKING TO THE MODEL ARCHITECTURE

Although it is obvious that the model and dataset affect how well different algorithms perform, the situation is worse than it might initially appear because even seemingly small details about the workload can have a large effect on our results. As a recent example, [Nado et al. \(2021\)](#) describes several small implementation details that make a large difference in matching state of the art MLPERF™ RESNET-50 IMAGENET training speed results. For example, they mention needing to set the initial value of the batch normalization scale parameters in the final layer of each residual block to a value less than one, match the exact v1.5 version of RESNET-50, match the virtual batch size of batch normalization, and (perhaps least surprisingly) avoid applying  $L^2$  regularization to bias variables.

To further illustrate how brittle training results can be, we ran three different experiments, each making a minor change to a different workload: changing the stride in the final residual block of a WIDE RESNET on CIFAR-10, adding extra batch normalization layers to RESNETS, and switching between two different published TRANSFORMER models that only differ by the placement of the layer normalization block.

**Wide ResNet with stride changes** In the first experiment, we start with a standard WIDE RESNET ([Zagoruyko and Komodakis, 2016](#)) architecture on CIFAR-10. For this workload, we compare the NESTEROV optimizer with ADAMW on a simple learning rate sweep with cosine decay (full experimental details provided in [Appendix A.3.1](#)). After selecting the best learning rate from the sweep, NESTEROV achieves a better test error than ADAMW, see [Table 1](#) and [Figure 2](#). The superiority of NESTEROV in this experiment is largely due to ADAMW overfitting in these conditions; i.e., ADAMW achieves a better training loss than NESTEROV. However, if we change the convolutional strides in the final residual block from the standard  $2 \times 2$  stride to a  $1 \times 1$  stride, the performance of NESTEROV shows a large 3.5% increase in test error. ADAMW, on the other hand, is largely unaffected by this architectural change. NESTEROV struggles to train the  $1 \times 1$  stride model due to an early training instability caused by the large loss curvature at initialization—the largest eigenvalue of the loss Hessian changes from 32 to 1052 with this architectural modification and initial weight distribution. This kind of instability can be dealt with using learning rate warmup ([Gilmer et al., 2021](#)). Adding 1000 steps of linear warmup makes NESTEROV stable at large learning rates despite the high initial loss curvature (see [Figure 8](#) in the Appendix).

**Additional batch normalization layers** A similar inversion of rank ordering between NESTEROV and ADAMW occurs when we add in additional batch normalization ([Ioffe and Szegedy, 2015](#)) layers to the 200-layer RESNETV2 architecture ([He et al., 2016b](#)). This batch normalization layer is added after every residual connection, similar to the structure of the original POST-LAYER NORM TRANSFORMER ([Vaswani et al., 2017](#)). This architectural change actually benefits NESTEROV with the 50 layer model, offering slight validation error and stability improvements. However, when this RESNET with extra batch normalization

Training Algorithm	Stride $2 \times 2$	Stride $1 \times 1$
	<i>(standard)</i>	
NESTEROV	<b>0.0376</b>	0.0726
ADAMW	0.0407	0.0420
NESTEROV + Warmup	0.0380	<b>0.0378</b>

Table 1: **Performance of different training algorithms on Wide ResNet with stride changes.** After changing from the standard  $2 \times 2$  stride to a  $1 \times 1$  stride, the performance of NESTEROV drops significantly, while ADAMW is largely unaffected by this architectural change. Adding a learning rate warmup to NESTEROV (i.e. NESTEROV + Warmup) allows it to recapture its original performance.

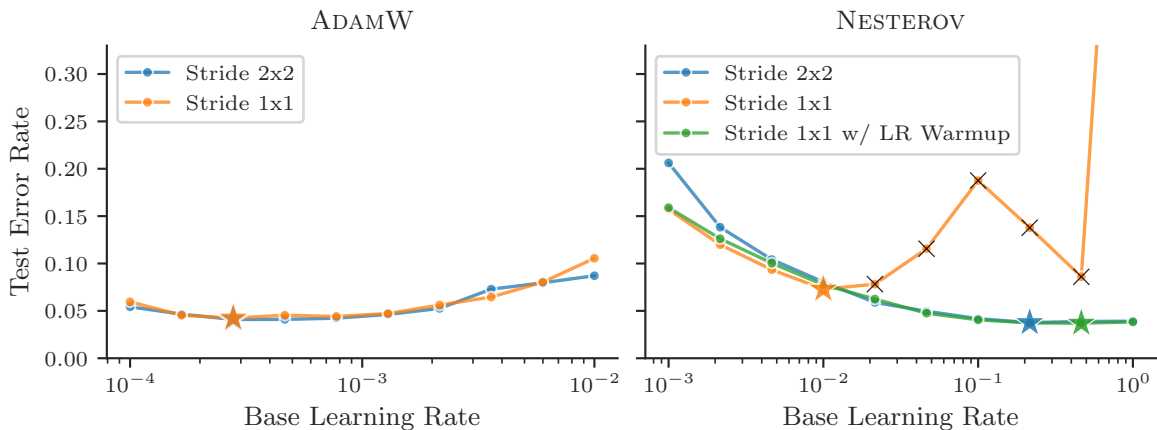


Figure 2: **Small architectural changes can result in different rankings between training algorithms.** In the above plot, we compare a learning rate sweep with ADAMW (*right*) and NESTEROV (*left*) on two variants of the CIFAR-10 WIDE RESNET. The first variant is the standard architecture, which uses a  $2 \times 2$  convolutional stride in each residual block, labeled “Stride  $2 \times 2$ ” (—). The second variant changes the strides in the final residual block to  $1 \times 1$ , labeled “Stride  $1 \times 1$ ” (—). On the standard architecture, NESTEROV outperforms ADAMW (the best-performing hyperparameter setting per architecture is highlighted with a colored star, e.g. (★)). However, the stride change results in significant training instability at high learning rates with NESTEROV (highlighted with a black cross (✕)). This instability causes NESTEROV to under perform ADAMW. If a learning rate warmup is applied (—) then we recover the performance of the original model. ADAMW on the other hand is unaffected by the training instability caused by this architectural change.

layers is scaled to 200 layers, it becomes completely untrainable with NESTEROV at any learning rate in the sweep (see Table 2). ADAMW is also affected by this architectural change, but is able to train successfully with a 3% drop in final performance. Similar to the WIDE RESNET stride change, this instability results from a dramatic increase in the initial loss curvature—with the extra batch normalization layers the initial loss curvature exceeds  $10^{10}$ ! Prepending a linear learning rate warmup in this case is insufficient to resolve the

training instability. However, with the addition of gradient clipping, we are able to recover the performance of the original architecture.

Training Algorithm	ResNet-200 ( <i>standard</i> )	ResNet-200 <i>Extra-BN</i>
NESTEROV	<b>0.2090</b>	No Feasible Trials
ADAMW	0.2626	0.2722
NESTEROV + Grad Clip	0.2091	<b>0.2094</b>

Table 2: **Performance of different training algorithms after adding an extra batch normalization layer to a ResNet-200.** While ADAMW is only slightly affected by adding extra batch normalization layers, vanilla NESTEROV becomes untrainable. Adding gradient clipping (i.e. NESTEROV + Grad Clip), however, allows it to recover the original architecture’s performance.

**Architectural modifications to Transformer models** The TRANSFORMER model is the most widely used architecture for natural language processing tasks. There are two popular versions of this model. First, the original version of the model from Vaswani et al. (2017), the POST-LAYER NORM (POST-LN) TRANSFORMER, which places the layer normalization between the residual blocks. Second, the PRE-LAYER NORM (PRE-LN) TRANSFORMER (Xiong et al., 2020), which places the layer normalization inside the residual blocks. Figure 3 shows the differences between the two architectures.

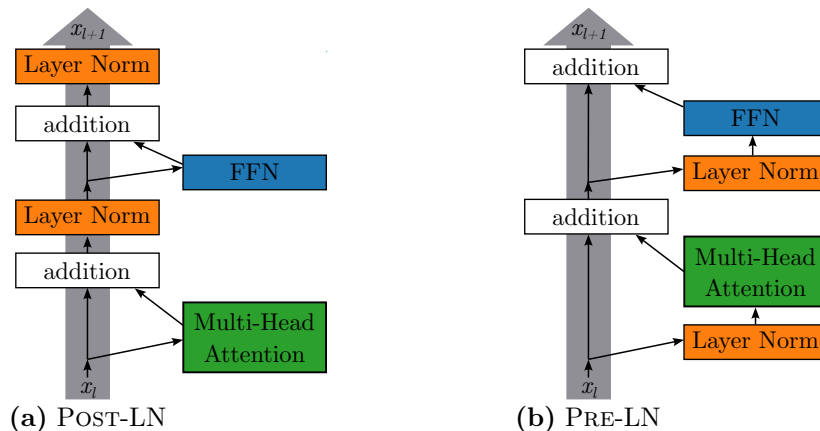


Figure 3: **Comparison between (a) Post-Layer Norm Transformer layer (Post-LN) and (b) Pre-Layer Norm Transformer layer (Pre-LN).** Figure recreated from Xiong et al. (2020) with permission.

This architectural modification affects training algorithms differently (Table 3). The experiment to produce Table 3 was conducted as follows: we designed a customized search space for each optimizer and ran 100 trials using random samples from this search space for each of the architectural modifications. Each of the trials were run for a fixed number of steps (in this case 133K steps). The best BLEU score from each trial was chosen for the

Training Algorithm	Pre-LN		Post-LN		Difference
	<i>Best</i>	<i>Confidence interval</i>	<i>Best</i>	<i>Confidence interval</i>	
ADAMW	31.3306	$31.0386 \pm 0.2811$	30.8098	$29.6819 \pm 1.1374$	0.5208
NADAMW	31.3381	$31.1153 \pm 0.2681$	30.8894	$29.6122 \pm 1.2608$	0.4487
NESTEROV	29.2951	$28.9091 \pm 0.8864$	26.3211	$20.0433 \pm 7.0693$	2.9740
SHAMPOO	30.2051	$29.9168 \pm 0.2816$	29.9482	$29.0940 \pm 1.1800$	0.2569

Table 3: **Test set BLEU score for Pre-Layer Norm (Pre-LN) and Post-Layer Norm (Post-LN) architectures for different training algorithms.** The architectural modification of PRE-LN vs. POST-LN affects ADAMW, NADAMW, NESTEROV, and SHAMPOO very differently as highlighted by the performance difference between the best-performing trials for PRE-LN and POST-LN. Table 16 in the Appendix shows the analogous table with the cross-entropy loss instead of the BLEU score.

analysis. We performed Bootstrap analysis (Efron and Tibshirani, 1993) on this data using a bootstrap sample size of 20 (corresponding to a plausible submission) and calculated the mean and standard deviation of the resulting distribution.

### 2.2.2 IMPLICATIONS OF WORKLOAD SENSITIVITY

As we have seen above, seemingly minor changes to the training pipeline, such as model architecture tweaks or the exact placement of normalization layers, can have a large effect on the results different training algorithms achieve. As a consequence, empirical comparisons can fall into several failure modes and fail to achieve our desiderata of being convincing, informative, and practically relevant. First, it is very tempting to claim that a particular method provides a general improvement even though it is only useful on a very specific workload. Second, comparisons to previously published results can easily fail to exactly match the training pipeline details and end up with better or worse results for reasons unrelated to the methods being evaluated. Achieving a true apples-to-apples comparison usually requires reproducing earlier results in the same codebase and matching other details of the implementation and setup, an extremely labor-intensive prospect. Sometimes this issue might manifest as improvements improperly attributed to the training algorithm. Third, common benchmark workloads may not be very representative of the applications that are most important to the community, especially since the importance of various applications naturally waxes and wanes. Popular benchmark tasks also suffer from a selection bias that leaves us with unusually well-behaved training problems (e.g., unusually stable) or with training problems that co-evolved with the training algorithms used by whoever created them. For example, it shouldn’t surprise us that ADAM works especially well for model architectures built to train using ADAM.

Our benchmark takes several steps to handle workload sensitivity. First and foremost, we fix a set of workloads (see Section 4.3) and prevent pipeline changes that are not part of the training algorithm (see Section 4.2). By separating changes to the training algorithm and other parts of the pipeline, we can avoid the case where training speedups are incorrectly attributed, at the cost of having to occasionally split philosophical hairs on what parts of



the training pipeline are off limits for submissions to alter. We also try to include a set of workloads that is relatively diverse, while still being representative of the most popular deep learning applications. However, workload diversity and relevance could always be better. Also, adding workloads comes at a cost both in engineering labor and in terms of our ability to run comprehensive experiments, sharply limiting what was feasible for the first version of our benchmark. Long term, we will need an even larger community effort to standardize workloads for training algorithm research and contribute measurements of new methods in a standardized way. Workloads should be implemented in open source code (ideally in multiple frameworks) and any variations or deviations should come with new names, in order to avoid confusion and misleading comparisons.

### 2.3 We Cannot Compare Families of Training Algorithms, Only Specific Instances

Most training algorithms in deep learning are not actually procedures we can run so much as algorithm templates we can instantiate with particular hyperparameters. Unfortunately, we cannot eliminate these hyperparameters because they exist for a reason: using different hyperparameter values for different workloads can lead to much better results (Section 2.3.1). To compare training algorithms with free hyperparameters, we must either set the hyperparameters to specific values or choose a specific procedure to tune them. If we choose a tuning protocol, we should view the protocol, including the hyperparameter search space, as part of the training algorithm definition, since the same algorithm template combined with a different hyperparameter tuning protocol (or even just a different search space Section 2.3.2) can produce significantly different results. For a fair comparison, it is necessary to ensure that all training algorithms are tuned for the same tuning *goal* (Section 2.3.3).

#### 2.3.1 TRAINING ALGORITHMS WITH DIFFERENT HYPERPARAMETERS

To illustrate the issues with using a single hyperparameter setting for all workloads, we fixed a training algorithm and studied the impact of per-workload hyperparameter tuning on the validation performance after a fixed number of training steps. Assuming the training algorithm definition specifies a hyperparameter search space  $\mathcal{H}$ , we can use quasirandom search to sample a finite set  $H \subseteq \mathcal{H}$  of candidate hyperparameters. Let  $\text{val}(w, h)$  be the validation metric value achieved on workload  $w$  by running the training algorithm with hyperparameter setting  $h$  for a fixed number of steps. Furthermore, let  $\text{val}_H(w) = \min_{h \in H} \text{val}(w, h)$  be the best validation score achieved among all hyperparameter settings  $h \in H$ .<sup>3</sup>

Suppose now that we used the same hyperparameter setting  $h$  for every workload instead of tuning the hyperparameters separately for each workload. Then we can consider the following measure of the worst-case relative performance degradation over all workloads

$$\varphi(h, H) = \max_w \left| \frac{\text{val}(w, h) - \text{val}_H(w)}{\text{val}_H(w)} \right|.$$

---

3. To simplify the notation, we assume that the validation metric should be minimized, e.g., as is the case for an error rate. For cases where a higher validation metric is better, e.g. the BLEU score, we would instead simply define  $\text{val}_H(w) = \max_{h \in H} \text{val}(w, h)$ .

Note that if  $\text{val}(w, h) = \text{val}_H(w)$  for all workloads (i.e., if the same  $h$  achieves the best validation score across all workloads), then  $\varphi(h, H)$  will be zero. Generally the higher the value of  $\varphi(h, H)$ , the larger the relative performance gap between employing  $h$  versus tuning over  $H$ . We can now find the  $h \in \mathcal{H}$  with the lowest such degradation, i.e.

$$\Phi(H) = \min_{h \in H} (\varphi(h, H)) = \min_{h \in H} \left( \max_w \left| \frac{\text{val}(w, h) - \text{val}_H(w)}{\text{val}_H(w)} \right| \right).$$

$\Phi(H)$  denotes the *best* worst-case relative performance degradation achieved among the hyperparameters  $h \in H$ . For a set of hyperparameters  $H$ , the quantity  $\Phi(H)$  thus measures how much validation performance degrades from using any single setting of the hyperparameters in  $H$  that is shared across all workloads instead of tuning within  $H$  to find the best per-workload setting. Thus, we can view it as a measure of how much the training algorithm benefits from per-workload hyperparameter tuning on a particular set of workloads, tuning over a particular set of hyperparameter search points  $H$ . By definition,  $\Phi(H) \geq 0$  and if it is zero then there is a hyperparameter setting  $h^* \in H$  that achieves the best performance on every workload among all settings in  $H$ . A larger value of  $\Phi(H)$  suggests that the training algorithm benefits more from workload-specific hyperparameter tuning.

We computed  $\Phi(H)$  for four training algorithms on eight workloads (for the details of these workloads, see [Table 7](#) and [Appendix D](#)). We constructed  $H$  by sampling 100 hyperparameter settings with quasirandom search from the tuning search spaces defined in [Table 8](#), for each training algorithm. [Table 4](#) shows the  $\Phi$ -values achieved by different methods using these specific search spaces. Note that  $\Phi(H)$  is a random variable that depends on the random seeds in training and is a function of the set of hyperparameters  $H$  which were sampled from the search space  $\mathcal{H}$ . In these experiments, we used a single training run per hyperparameter, i.e., we used one sample to estimate the mean of  $\text{val}(w, h)$ .

The results illustrate the performance gains that can be achieved by using workload-specific tuning of hyperparameters. We see that for every algorithm the  $\Phi$ -value is least 0.169 which implies that for every hyperparameter setting there is a workload for which the performance of that setting is at least 16.9% worse than the performance of the optimal hyperparameter setting for that particular workload. [Tables 17 to 20](#) in the appendix show the performance of the *optimal per-workload* hyperparameters (i.e.,  $\text{val}_H(w)$ ), the performance of the *optimal overall* hyperparameters per workload (i.e.,  $\text{val}(w, h^*)$  where  $h^*$  is the hyperparameter that minimizes  $\Phi(H)$ ), and their associated relative performance degradation for each training algorithm.

### 2.3.2 TRAINING ALGORITHMS WITH DIFFERENT HYPERPARAMETER SEARCH SPACES

Differences in tuning protocols can wreak havoc on empirical work. [Choi et al. \(2019b\)](#) argue that they are the single most important factor explaining contradicting results from empirical optimizer comparisons. Learning rate schedules are an especially pernicious free parameter to tune because different algorithms result in different implicit schedules. Ideally, these implicit schedules would be separated from the other properties of the algorithm ([Agarwal et al., 2020](#)). There is an entire literature on semi-automated tuning algorithms for hyperparameter tuning that includes Bayesian optimization and other black-box optimization techniques. However, even if a particular Bayesian optimization tool became



Training Algorithm	$\Phi(H)$
ADAMW	0.195425
NADAMW	0.169197
NESTEROV	0.230001
HEAVY BALL	0.239372

Table 4: **Using workload-specific hyperparameters can significantly improve the performance of training algorithms.** The  $\Phi$ -values obtained by different training algorithms with search spaces as defined in Table 8. Higher values indicate the necessity of tuning to achieve good performance across multiple workloads.

standard, such tools still require search spaces as input and any procedure for constructing search spaces would need to be aware of the tuning budget (Ariafar et al., 2022).

Even if we fix everything about the tuning protocol except the search space, changes to the search space alone suffice to change training results. To illustrate this, we defined two search spaces for ADAMW (Table 5) and compared the best hyperparameter points found from each search space for each workload in our benchmark. The first search space (ADAMW NARROW) is completely contained within the second search space (ADAMW BROAD). In principle, with a large enough tuning budget, the broader search space can be no worse than the narrower one. However, at any particular tuning budget, either search space might end up performing better. Both search spaces seem reasonable *a priori*. Indeed, they both contain good points, although without the benefit of hindsight one might be concerned that the narrow search space does not allow small enough values of weight decay.

Search Space	Learning Rate	Weight Decay	$1 - \beta_1$	$\beta_2$
ADAMW NARROW	[2e-4, 5e-3]	[2e-2, 0.5]	0.1	0.999
ADAMW BROAD	[5e-6, 2e-2]	[5e-6, 2.0]	[1e-3, 1.0]	0.999

Table 5: **Hyperparameter search spaces for two training algorithms using AdamW.** All hyperparameters are sampled using a log-uniform distribution with the lower and upper bounds as shown in the table.

From each of the search spaces, we sampled 100 points using quasirandom search. Using these 100 points, we simulated tuning with a budget of  $T$  trials by repeatedly sampling groups of  $T$  trials, with replacement, from the 100 real results and taking the best trial within the group based on its performance on the validation set. Table 6 shows the results of simulated tuning with a budget of 20 trials for the two search spaces. The validation performance is the median over 1000 simulations. At this budget, the narrow search space achieves markedly better validation metrics across all workloads in our benchmark.

Although it might seem obvious that the results would depend on the search space, it is still not standard within the literature to report exact tuning protocols or even supply search space details. The narrow and broad search spaces we considered here could have instead had less—or even no—overlap and produced even more dramatic differences in results. It is extremely easy to select a particular search space for tuning a baseline and then claim that

Workload		AdamW Narrow			AdamW Broad		
		Median	$Q_1$	$Q_3$	Median	$Q_1$	$Q_3$
CRITEO 1TB	DLRMSMALL	<b>0.12401</b>	0.123967	0.124025	0.124087	0.12396	0.124025
FASTMRI	U-NET	<b>0.734746</b>	0.734590	0.734936	0.734311	0.734054	0.734522
IMAGENET	RESNET-50	<b>0.23256</b>	0.23094	0.23330	0.24334	0.23904	0.24708
	ViT	<b>0.21992</b>	0.22118	0.22118	0.23616	0.22694	0.24038
LIBRISPEECH	CONFORMER	<b>0.075989</b>	0.075962	0.076817	0.080673	0.078963	0.087340
	DEEPSPEECH	<b>0.112706</b>	0.112353	0.113485	0.120674	0.116902	0.127974
OGBG	GNN	<b>0.28214</b>	0.281595	0.284034	0.276307	0.275642	0.279285
WMT	TRANSFORMER	<b>31.3523</b>	31.2824	31.3946	30.9950	30.9129	31.1748

Table 6: **Performance across multiple workloads for AdamW with two different hyperparameter search spaces.** Shown are the median, as well as the lower and upper quartiles ( $Q_1$  and  $Q_3$ ) of the best observed validation metric. The results are for a budget of  $T=20$  trials (see Table 21 for results with  $T=5$ ) across 1000 simulations. At this budget, ADAMW NARROW performs significantly better across all test workloads.

another algorithm outperforms it without including appropriate caveats about the search space. Our experiment here shows that even if we consider two very reasonable search spaces, merely neglecting to select a search space that matches the tuning budget could subtly weaken a baseline and befuddle a comparison between training algorithms.

### 2.3.3 TRAINING ALGORITHMS WITH DIFFERENT TUNING GOALS

Even if we fix everything about the tuning procedure *including* the search space, differences in tuning *goals* can lead to unfair comparisons. Specifically, tuning to achieve the best validation error within a fixed training time budget is *not* the same as tuning to achieve a fixed validation error as fast as possible. Suppose some previously published result achieves a particular validation error rate on a particular workload after training for a certain amount of time using Algorithm X. Now suppose that we demonstrate that Algorithm Y can achieve the same validation error in substantially less time (on the same hardware). In many cases, this kind of comparison, although seemingly innocuous, will be unfair to Algorithm X because the result with Algorithm X was from a paper that was not trying to minimize training time, but was instead engaged in an implicit competition to get the best possible validation error, within their available budget.

We tuned the hyperparameters for ADAMW for RESNET-50 on IMAGENET for two training step budgets, 186,666 and 233,333. Both studies use the same search space and, by using the same seed, the same 100 hyperparameter samples. They also used the same cosine decay learning rate schedule with a linear learning rate warmup. However, the cosine decay schedule depends on the maximum number of training steps and therefore differs between the two studies (for the complete search space see Table 8, ADAMW). In both studies, the hyperparameter setting achieving the best validation error happened to be the same (see Table 22 in Appendix A.4.3). We then retrained this best trial for both studies using 20 different random seeds. Figure 4 shows the best validation error achieved so far versus the training step for Trial A (—, step budget of 186,666 steps), and Trial B

(—, step budget of 233,333 steps). Both trials achieve nearly identical validation errors (22.4% – 23.0% depending on the seed), but Trial A is clearly (by design) much faster. Training longer tends to very slightly improve the validation error (median of 22.6% vs a median of 22.7%), but this improvement is hard to detect with the variance across seeds. Furthermore, if we compare larger and larger pairs of training step budgets, any difference between the budgets will be swamped by the variance across different runs. This experiment shows that we can get (roughly) the same validation error result faster, *simply by setting a lower training step budget*.

Researchers trying to achieve the best headline validation error number to publish will naturally give their experiments a generous step budget to make their lives easier, and not try to find the minimum step budget that can still reproduce their result. Thus we might imagine published results more like Trial B when they are being tuned for the best validation error using a somewhat arbitrary, but generous, step budget. In general, learning rate decay schedules are necessary for the very best results, and we should expect the best validation error rates to be achieved near the end of training when the learning rate has already been reduced. However, common decay schedules make it extremely dangerous to assume that some state-of-the-art result, in terms of out-of-sample error, was achieved with a near-minimal number of training steps. Unfortunately, that is precisely what we are implicitly assuming when we compare results tuned to minimize training time to reach a particular error rate with results tuned to achieve the best possible error rate given a particular time (or step) budget.

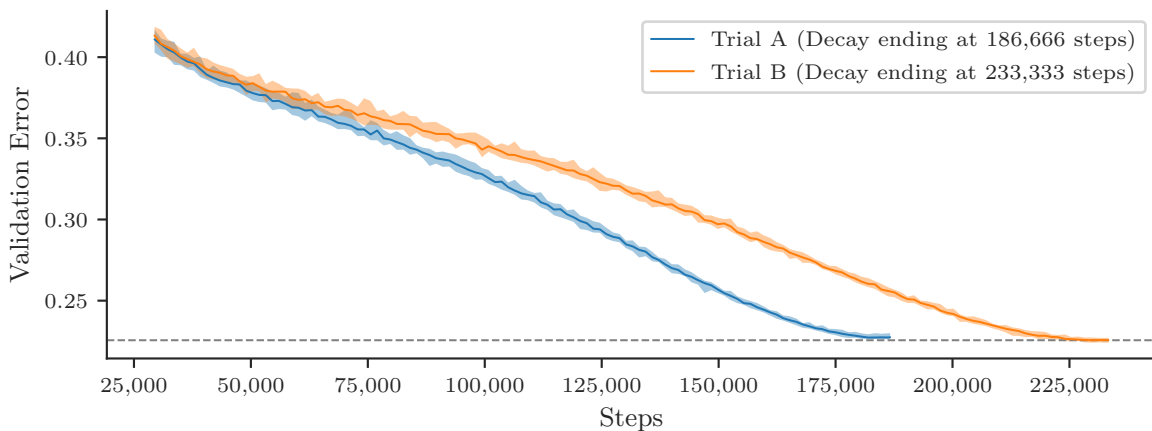


Figure 4: **For a fair comparison, training curves need to be tuned for the same criterion.** Two ADAMW training runs (—, —) for RESNET-50 on IMAGENET using hyperparameters tuned within the same search space, but using different step budgets. Since the cosine learning rate decay schedule stretches with a longer step budget, we see “slower” training caused by the larger step budget (—). For each of the hyperparameter settings, we ran 20 different random seeds to create min/max error bounds around a median trajectory (■, ■). The dashed gray line (--) denotes the best median validation error achieved by both training runs. See [Appendix A.4.3](#) for experimental details.

In light of the strong effect tuning protocol details will have on the results of any empirical comparison between training algorithms, we have no choice but to view the tuning

protocol as part of the algorithm. Additionally, we can only compare two results that are tuned to optimize the same criterion, i.e., two results that are both tuned to minimize time-to-result or are both tuned to minimize validation error at a fixed time budget. Once we fold the tuning protocol into the definition of the method, we can study one algorithm with two different tuning protocols as if it was two separate methods. Algorithm designers could then also provide guidance on how the methods they introduce should be tuned. Currently, new training algorithms rarely come with sufficient guidance on how they should be tuned in various budget scenarios. Despite not needing much (or any) tuning being a common selling point, we are not aware of any popular training algorithm for neural networks where this property has been precisely defined and convincingly demonstrated, although this would certainly be a valuable contribution. Unfortunately, the tuning protocol is necessarily a function of the tuning budget, which makes it difficult for algorithm designers to specify how their method is best tuned in all possible budget scenarios. Nevertheless, even providing a short list of budget scenarios along with tuning guidance would improve the situation dramatically, especially since the largest tuning budget scenarios are probably less critical to address for practitioners. Ultimately, methods that are easier to tune—or even fully self-tuning—would be extremely valuable if they could be supported by compelling experimental evidence. [Section 4.4](#) describes our approach to tuning in detail.

Even if we view results as conditional on a tuning protocol, publication bias and other similar selection bias effects can, in effect, amount to implicit tuning. Ultimately, the root of the problem is that quantifying tuning effort when developing a training algorithm is nearly impossible. If we try 1000 different methods on a small set of workloads and find one that seems to work across this small set “without much tuning”, we might just have obfuscated the tuning process. However, “off the books” implicit tuning should not bother us if the resulting training algorithms generalize to new workloads. Therefore, the solution to this issue is a combination of various generalization incentives and safeguards: preventing workload cherry-picking by standardizing sets of benchmark workloads, sharply limiting per-workload adaptation and tuning in empirical comparisons, using larger sets of more diverse benchmark workloads while measuring performance aggregated across workloads, and perhaps even generating novel workload variations randomly to create “held-out” workloads.

## 2.4 Strong Baselines Are Far Too Hard to Obtain

Researchers proposing new training algorithms have total control over the baselines they compare with their algorithms. Even with everyone acting in good faith, it is very easy to accidentally make an experimental design choice that gives a new algorithm an unfair advantage. For example, we could select a learning rate schedule family that works well with a new technique and not even know how appropriate it is for our baselines. Or we might select workloads where controlling overfitting is essential for good results, but not apply sufficient regularization to a baseline that minimizes training loss *faster* than the novel method we are studying. There are essentially limitless possible choices we can make when designing experiments that might inadvertently hamstring our baselines and show overoptimistic improvements for some novel method. Researchers just have far too many choices to make when comparing training algorithms. Even worse, the least careful

experiments not only take the least effort to conduct, but also tend to produce the most impressive-seeming results.

Although the problem of weak baselines exists in other parts of deep learning (and in machine learning as a whole), training algorithm comparisons seem particularly fraught compared to, say, model comparisons on a single task (e.g., what is the best neural network architecture for large vocabulary speech recognition). There are many factors at play, including the challenges described above, but experimental methodologies for studying training algorithms seem less mature as well. There is a long tradition of competitions surrounding particular datasets, e.g., the IMAGENET competition (Russakovsky et al., 2015), that have at least produced a level playing field within the narrow boundaries of the original rules. We can apply the same approach to training algorithm comparisons. Direct competition under a shared set of rules can cut through the Gordian knot of tangled incentives and inconsistent experimental protocols.

### 3. Related Work

Our goal is to create a benchmark for neural network training algorithms. Related work can be grouped into three broad categories: Earlier efforts to benchmark training algorithms for deep learning (Section 3.1); previous work identifying issues with existing approaches for evaluating deep learning training algorithms (Section 3.2); and, finally, relevant prior work that makes a case for, or against, particular training algorithms (Section 3.3), with a special focus on disagreements in the literature.

#### 3.1 Existing Benchmarks

Domain-specific benchmarks and challenges have driven advances in machine learning for decades (Garofolo et al., 1993; Krizhevsky and Hinton, 2009; Deng et al., 2009; Lin et al., 2014; Russakovsky et al., 2015; Panayotov et al., 2015; Bojar et al., 2015; Zbontar et al., 2018; Wu et al., 2018; Hu et al., 2020; Dwivedi et al., 2023; Dong and Yang, 2020). Such benchmarks generally provide a curated dataset—usually with pre-defined training, validation, and test splits—and specify what performance metrics should be used to measure progress. The research community uses benchmarks like these to demonstrate progress on specific tasks (e.g., image classification, speech recognition, machine translation). Ultimately, domain-specific benchmark datasets are a component of almost all training algorithms benchmarks, including ours. Progress in domain-specific benchmarks often involves a combination of new models, regularization techniques, and training algorithms. In many cases, however, it is difficult to disentangle how much improvement is due to, e.g., changes in the model versus the training procedure or tuning effort (Bello et al., 2021).

A separate set of benchmarks thus aims to measure specifically the performance of training algorithms on different systems, or in different frameworks. The MLPERF™ TRAINING benchmark (Mattson et al., 2020),<sup>4</sup> which grew out of the DAWN BENCH benchmark (Coleman et al., 2017), has the goal of evaluating system performance for neural network training. The *Closed Division* of the benchmark aims to measure the performance of machine learning hardware and systems by requiring mathematical equivalence to a reference implementation,

---

4. <https://mlperf.org/training-overview>

while still allowing submissions on different hardware. Similar to the MLPERF™ TRAINING benchmark, the DEEP LEARNING BENCHMARK SUITE,<sup>5</sup> DEEPBENCH,<sup>6</sup> and the TRAINING BENCHMARK FOR DNNs (Zhy et al., 2018) also focus on evaluating machine learning *systems* (accelerators, operating systems, and frameworks) using fixed, well-established workloads and training algorithms.

Unlike the Closed Division of the MLPERF™ TRAINING benchmark and other similar systems benchmarks, the *Open Division* of MLPERF™ TRAINING, has slightly more overlap with our goals. In this division, in addition to running on different systems, submissions are allowed to modify the training algorithm and model. However, allowing arbitrary hardware, potentially at radically different scales, makes it impossible to distinguish improvements due to algorithms from those due to extra computation; in fact, an algorithm may not perform the same at scales. Additionally, by allowing model changes, there is also no way to isolate improvements due to training algorithm modifications. Generally, since submissions are viewed as independent per-workload entries, there is no incentive to avoid hyper-specific changes that only help one particular benchmark workload. Even if a participant provides submissions for all workloads, they are free to use completely different training algorithms or models for each workload and perform unlimited workload-specific tuning.

DEEPOBS (Schneider et al., 2019), a software package and set of test problems for benchmarking deep learning optimization methods, aims to simplify research and development of training algorithms. DEEPOBS recommends reporting three key performance indicators: the final predictive error after a fixed number of steps, the wall-clock runtime per optimization step, and the ease of hyperparameter tuning. Since it positions itself as a research tool, the package does not provide guidance on how these three performance indicators should be weighted against each other or aggregated across the provided test problems. While DEEPOBS fixes the workloads (addressing the challenges mentioned in Section 2.2), it contains only one large-scale deep learning workload (training RESNET-50 on IMAGENET). The other workloads in DEEPOBS are mainly smaller-scale deep learning tasks (e.g., image classification on MNIST, CIFAR, and SVHN). Synthetic problems and smaller workloads are useful for rapid experimentation, and thus helpful to researchers. However, conclusions drawn from these problems do not always generalize to larger workloads (e.g., larger models and/or datasets). Furthermore, DEEPOBS leaves the user in control of many parts of the training protocol, such as the tuning budget, and thus it is ultimately up to the user to ensure a fair comparison.

In follow-up work, Schmidt et al. (2021) leverage the DEEPOBS framework to perform an empirical comparison of fifteen popular training algorithms. For this purpose, they fix the benchmarking protocol by comparing the methods at four different tuning budgets. This empirical comparison is not a competitive benchmark. Instead, Schmidt et al. (2021) determined the hyperparameter search spaces of the training algorithms themselves, aiming for a “fair representation” of each method based on statements made in the original publication, and not necessarily trying to make each method perform at its best (as the original authors may have done in a competitive setting). In addition, some of the issues of DEEPOBS described above carry over to the study by Schmidt et al. (2021). Notably,

---

5. <https://github.com/HewlettPackard/dlcookbook-dlbs>

6. <https://github.com/baidu-research/DeepBench>



their empirical comparison uses mainly small-scale deep learning problems, and there is no aggregated score across test problems.

[Moreau et al. \(2022\)](#) introduce the BENCHOPT software suite for evaluating machine learning optimization methods. They focus on features that a software platform should support (e.g., modularity, extensibility) to enable fair and reproducible comparisons. However, at the time of writing, although BENCHOPT includes 16 different optimization problems, only one is a deep neural network training workload.

### 3.2 Methodological Critiques of Training Algorithm Comparisons

[Bartz-Beielstein et al. \(2020\)](#) present a series of recommendations for defining benchmarks for optimization methods for several families of optimization problems, including deep learning optimizers. Although these recommendations generally align with the benchmark described in this paper, as we discussed in [Section 2](#) and again in [Section 4](#), below, a deep neural network training algorithm benchmark should not be restricted solely to optimization methods, and should also encompass aspects like regularization and data sampling.

Several recent works emphasize the importance of hyperparameter tuning in making fair comparisons of different training algorithms. [Schneider et al. \(2019\)](#) recommend finding and reporting the best-performing hyperparameters for each test problem in DEEPOBS. [Sivaprasad et al. \(2020\)](#) advocate that researchers introducing a new training algorithm should always provide a workload-agnostic prior distribution over hyperparameters that can be used in conjunction with random search. They illustrate how the relative ranking of different optimization methods may change on a given workload depending on the hyperparameter tuning budget. [Choi et al. \(2019b\)](#) illustrate how rankings can be sensitive to which hyperparameters are tuned, e.g., showing that tuning ADAM’s  $\beta_2$  and  $\epsilon$  can significantly improve performance compared to simply using their default values. They also illustrate that the choice of search space parameterization can influence the effectiveness of hyperparameter tuning, for example, explaining why tuning the relative initial learning rate  $\alpha_0/\epsilon$  of ADAM is more efficient than tuning  $\alpha_0$  and  $\epsilon$  independently.

### 3.3 Disagreement over Training Algorithms: The Case for Clear Benchmarks

Several of the studies discussed in [Section 3.2](#) are motivated by earlier work making contentious claims about certain training algorithms. Below, we review several examples of disagreements about training algorithms in the literature, since they demonstrate the opportunity for trusted, competitive benchmarks to promote joint progress.

As one prominent example, [Wilson et al. \(2017\)](#) showed that SGD converges to the maximum margin solution and provide a toy example where ADAGRAD finds a solution which generalizes poorly while SGD finds a solution that generalizes well. The same work also presented empirical results where adaptive methods such as ADAGRAD, RMSPROP, and ADAM generalized much worse than SGD with MOMENTUM. [Choi et al. \(2019b\)](#) subsequently argued that [Wilson et al.](#)’s empirical conclusion is only valid under what they consider restrictive conditions, namely fixing ADAM’s  $\beta_2$  and  $\epsilon$  parameters to their default values, and only tuning a constant learning rate. In contrast, tuning all of ADAM’s hyperparameters and incorporating a learning rate decay schedule allows ADAM to perform comparably or even slightly better than SGD with MOMENTUM. Additionally, [Agarwal](#)

et al. (2020) argued for the importance of separating the effects of search direction and per-layer learning rate scaling when comparing optimizers, showing that conflating the two also accounts for some of the findings of Wilson et al. (2017).

As another example, Liu et al. (2020) observed that the effective per-coordinate scaling factors can have high variance early in training, which can lead to unstable training dynamics. They proposed the rectified ADAM optimizer (RADAM) to compensate for this, and reported several experiments showing that RADAM outperforms standard ADAM with learning rate warmup on image classification and machine translation tasks. Subsequently, Ma and Yarats (2021) showed that using ADAM with an appropriately tuned learning rate warmup performs comparably or slightly better to RADAM on similar tasks.

There are also disagreements about whether optimizers that use off-diagonal curvature information are useful when training neural networks. Hinton and Salakhutdinov (2006) proposed a deep autoencoder model for reducing the dimensionality of the input data. Martens (2010) showed that the Non-Linear Conjugate Gradient method (Nocedal and Wright, 1999) is quite ineffective in minimizing the training loss in this problem. In contrast, non-diagonal methods such as truncated-Newton (Nocedal and Wright, 1999) provide a significant advantage. Since Martens (2010)’s work, the deep autoencoder problem has served as a standard benchmark for comparing diagonal and non-diagonal optimization techniques where the gap between the two methods is expected to be fairly substantial (Martens and Grosse, 2015; Goldfarb et al., 2020; Anil et al., 2020; Ren and Goldfarb, 2021; Ren et al., 2020; Zhao et al., 2022; Bae et al., 2022). However, Amid et al. (2022) found that with sufficient hyperparameter tuning, diagonal preconditioning based methods such as RMSPROP worked reasonably well on this problem compared to non-diagonal preconditioning based methods such as K-FAC (Martens and Grosse, 2015) and SHAMPOO (Gupta et al., 2018; Anil et al., 2020). Amid et al. (2022) thus argued that algorithmic improvements to diagonal methods can practically eliminate the gap between diagonal and non-diagonal methods.

Finally, there has been substantial disagreement in the literature about whether new training algorithms are necessary as the batch size increases. Although some of this disagreement is due to imprecise claims and experimental setups that implicitly demand perfect batch size scaling (Shallue et al., 2019), various training algorithms have been proposed to handle the supposed problem of “large batch training.” For example, You et al. (2017) and You et al. (2020) introduced the LARS and LAMB optimizers, respectively, and argued they were necessary at larger batch sizes. Subsequently, MLPERF™ TRAINING adopted LARS for its RESNET-50 on IMAGENET workload and various submitters have demonstrated impressive training speed results. Nevertheless, Nado et al. (2021) have since shown that, with appropriate hyperparameter tuning, it is possible to obtain similar training speed results using SGD with NESTEROV MOMENTUM with batch sizes up to 32,768. They also showed that stronger ADAM baselines could outperform LAMB results on BERT (Devlin et al., 2019) pre-training, and argued that there was no convincing evidence that LARS and LAMB should be used over standard techniques.

The examples above illustrate how the challenges in benchmarking training algorithms discussed in Section 2 directly affect the training algorithms community. Perhaps most critically, they emphasize the importance of tuning hyperparameters in a fair and consistent way to give each algorithm the best chance to perform well (Choi et al., 2019b; Sivaprasad



et al., 2020). Although this may sound straightforward, substantial care must be taken when defining the hyperparameter search space for each algorithm. Framing training algorithm comparisons as a competition has the crucial advantage that each participant individually strives to make their method work best under the constraints of the contest, with one participant’s method becoming another’s baseline. In contrast, the status quo in the literature is for researchers to make uncontrolled changes and depend on the vicissitudes of the (noisy) peer review process to enforce some notion of a “fair” comparison with previous work, resulting in confusing comparisons with baselines that tend to be much too weak. Working from a common, open codebase enables researchers to independently reproduce and verify the claims of others, and also makes it easier for entrants to the competitive benchmark to share their submissions with the community for future comparisons and studies.

## 4. Rules

The goal of our benchmark is to identify **general-purpose neural network training algorithms** that can speed up training on new workloads. Our benchmark measures time-to-result (Section 4.1) on a fixed hardware configuration. To ensure that the benchmark results have real-world relevance, we define goal error rates based on the validation and test sets instead of the training loss. In order to isolate the effect of the training algorithm, submissions must adhere to a specific API (Section 4.2) and cannot make alterations outside a limited number of functions. To incentivize generally useful algorithms, we require that a single training algorithm simultaneously performs well across multiple workloads (Section 4.3) without manual workload-specific adaptation. Instead, any adaptation to the workloads should either be possible with generic tuning methods (Section 4.4.1) or be performed as part of the timed training process (Section 4.4.2). All workloads are considered when calculating an aggregate benchmark score of the training algorithm (Section 4.5). The resulting benchmark score is intended to serve as an estimate of the performance of a training algorithm on unknown workloads.

In the following sections, we describe the essential elements of our benchmark rules and explain the reasoning behind them. This section is based on the rules at the time of writing. As the benchmark evolves, the rules may also change, and the most up-to-date, complete rules can be found at [github.com/mlcommons/algorithmic-efficiency/blob/main/RULES.md](https://github.com/mlcommons/algorithmic-efficiency/blob/main/RULES.md).

### 4.1 A Time-to-Result Benchmark

A submission’s score is based on the time to reach the target validation and test scores on each workload. Training is considered complete, and the clock stops, as soon as a submission achieves the validation and test targets. For practical reasons, submissions are limited by a maximum allowed runtime on each workload. If a submission fails to achieve the targets within this maximum runtime it will receive an infinite training time on this trial.<sup>7</sup> Although setting these targets will always be contentious to some degree, we need

---

7. Depending on the tuning ruleset, a submission may get several trials or only one trial per workload; see Section 4.4 for more details.

a systematic procedure to determine target validation and test scores that are competitive (ideally near the state of the art), while being achievable within a reasonable time budget.

For a given target-setting runtime budget, we defined the validation and test targets of a workload based on what can be reliably achieved using standard methods. Specifically, we used four popular training algorithms (HEAVY BALL, NESTEROV, ADAMW, and NADAMW) tuned with quasirandom search (Bousquet et al., 2017) on hand-engineered, workload-agnostic search spaces. For each training algorithm and workload, we ran 200 trials for the full target-setting runtime budget<sup>8</sup> and determined the best combination of training algorithm and hyperparameter settings by finding the trial that achieved the best validation error. We re-ran this combination of training algorithm and hyperparameters 20 times with different random seeds, taking the median of the best achieved validation errors across seeds to obtain a *validation* target. Out of the 10 repeated runs that achieved this validation target, we took the worst achieved test error across seeds as our *test* target. Taking the median validation performance after rerunning the best hyperparameter point prevents our procedure from selecting a lucky outlier. Our protocol defines both *validation* set targets and *test* set targets in order to implement the external tuning ruleset, as described in Section 4.4.1. Exact tuning details and results can be found in Section 5. It is important to note that the target-setting procedure does not constitute a valid submission since it uses 200 trials instead of the 20 trials available in the external tuning ruleset.

The procedure outlined above requires us to determine two runtime budgets for each workload. First, we need to set the *maximum allowed runtime* for the submissions. Ideally, this runtime budget would be infinite, as it would allow us to accurately gauge the time required for a submission to successfully train each workload. However, for practical reasons, we must limit this budget. Second, we need to set a *target-setting runtime budget* for the target-setting procedure described in the previous paragraph. We decided to use different runtime budgets for the submissions and target-setting. Specifically, the target-setting runtime budget is set to  $0.75 \times$  the maximum allowed runtime for submissions. By allowing submissions a more generous runtime budget, they have some leeway to spend extra time on certain workloads, if they can compensate for it on other workloads.

When setting maximum runtimes for a workload, it is important to find a balance between challenging and achievable targets. Increasing the maximum runtime may improve the performance but it also increases the overall runtime of the benchmark. To balance the dual objectives of stringent targets, near the state of the art in the literature, and making benchmark submissions practical to evaluate on the full suite of workloads, we aimed to limit the combined runtime of all fixed workloads to 100 hours on the benchmarking hardware. We used both preliminary experiments and published results on the datasets and models of our workloads as a guide in allocating this combined runtime budget to the individual workloads. The allowed maximum runtimes for the submissions on each workload are shown in Table 7. Section 5 discusses how the benchmark targets (which were set using  $0.75 \times$  the runtimes presented in Table 7) compare to results from the literature.

---

8. For simplicity, we converted these target-setting runtime budgets into step budgets since all four training algorithms used for target-setting happen to have nearly identical average step times.

#### 4.1.1 MEASURING RUNTIME

We selected elapsed real time, or wall-clock time, as our measure of runtime for training. We made this choice to maximize the practical relevance of our timing measurements and to avoid imposing new restrictions on how training algorithms could operate. To make meaningful comparisons of different algorithms in terms of wall-clock time, all algorithms must be run on a standardized hardware system (discussed further in [Section 4.1.2](#)).

The research literature contains examples of several alternatives to directly measuring wall-clock time. In some cases, researchers count training steps, number of forward passes, gradients, or some other abstract notion of iterations instead of an all-encompassing time measurement. Counting steps is convenient when iterations have the same, consistent cost during a single run and we can measure the average time-per-step for different algorithms easily. More generally, abstracting away the hardware and system conditions is appealing when it is possible, but such abstract runtime proxies are sadly not an option for a general training algorithms benchmark. Counting iterations is meaningless since submissions are free to redefine what a single step means. Counting gradient computations is similarly meaningless since submissions can vary in what derivatives (if any) they compute, or even use radically different batching schemes that include intelligent data selection.

Abstract notions of steps completed or examples processed do not necessarily reward algorithms that are the most useful in practice. Some algorithms might cleverly reclaim idle accelerator time while waiting for new data (e.g., by applying Data Echoing ([Choi et al., 2019a](#))). Some algorithms might be especially memory efficient and thus able to use larger batch sizes and better exploit hardware parallelism. Conversely, some algorithms might be impractical because they require too much extra memory.

Measuring wall-clock time has some disadvantages. Although we care about implementation quality to some extent, our intention is not to make a software benchmark since that goal is better served by comparing mathematically equivalent programs. Furthermore, variation in network congestion when reading data or writing results is of little interest, nor do we care about benchmarking unrelated processes running on the system and how they interfere with the training program. We expect that by using a standardized hardware system described below (namely, a single server with multiple GPUs), and by having the training program be the only significant (i.e., computationally intensive) program running on the system, that these additional factors will not substantially affect the measured runtimes.

#### 4.1.2 STANDARDIZING BENCHMARKING HARDWARE

To fairly compare wall-clock runtimes of different algorithms, the times should be measured on a standardized training system (e.g., hardware accelerators, memory, CPUs, interconnect) using a standardized execution environment, including consistent software versions. For the initial version of the benchmark, we selected a system with  $8 \times$  NVIDIA V100 GPUs with 16GB of VRAM per card since it is widely available on major cloud computing providers and offers a good compromise between performance and affordability. This official benchmarking system only needs to be used for final timing runs. Tuning a submission only requires a comparison between different hyperparameter settings of a single training algorithm, so it is fine to use a different but consistent system for tuning experiments.

Inevitably, accelerators and the systems available on the market will change in the future, so future iterations of the benchmark may adopt new benchmarking systems or even support multiple “weight classes” of systems. As long as future benchmarking systems are strictly more capable, especially in terms of accelerator memory capacity, it should be relatively straightforward to rerun the baselines and the top-performing previous submissions on new systems, at least using the old batch sizes to provide a pessimistic bound.

## 4.2 Specifying a Training Algorithm

The rules define a precise API to be used when specifying a training algorithm submission. A submission to the competitive benchmark must define four *submission functions*. The submission functions allow submitters to define how the submitted training algorithm updates the model parameters (`update_params`) and how data are selected for training (`data_selection`). Furthermore, submitters can initialize the training algorithm’s state (`init_optimizer_state`) and must provide a batch size for each workload (`get_batch_size`). An implementation of the submission functions may make use of a limited API to get some basic information about the workload. A detailed description of the submission functions signatures<sup>9</sup> and the benchmark API can be found in the rules.<sup>9</sup>

In addition to implementing the four submission functions, a training algorithm may have hyperparameters that will be tuned following one of the rule sets described in [Section 4.4](#). Submissions to the external tuning ruleset described in [Section 4.4.1](#) must also specify a hyperparameter search space.

Apart from defining the submission functions and hyperparameter search space, a submission may not modify any other parts of the training pipeline. The rest of the training program (e.g., implementations of data pipelines, model architecture, training loop, calculation of validation and test metrics, and measurement of runtime) are implemented by the benchmark. We intentionally restrict which parts of the training program a submission may modify for two main reasons. First, we prohibit certain types of changes in order to isolate speedups due to training algorithm changes ([Section 4.2.1](#)). Second, we hope to deter submissions that are over-specialized to particular benchmark workloads and are unlikely to be generally useful on new problems ([Section 4.2.2](#)).

The benchmark exposes a limited API of functions that a submission may call, within the four submission functions, to get information about a training workload at execution time. The workload-specific information available to submissions is restricted, since the goal of this benchmark is to identify training algorithms that are generally useful across many workloads. One of the ways in which a submission may be workload-specific is by providing a different batch size to be used for each workload. This is allowed since the benchmark hardware, including accelerator memory capacity, is fixed, and different training algorithms may involve storing and updating different state.

It is impossible to design an API and comprehensive rules that would prohibit all possible cases of submissions circumventing the spirit of the benchmark. Instead, we will prohibit these submissions in the abstract and will defer rulings about specific submissions to a “spirit [of the rules] jury.” Similarly, there may be modifications that we would allow in

9. See the “Valid Submission” Section in <https://github.com/mlcommons/algorithmic-efficiency/blob/main/RULES.md#valid-submissions>.

principle, but which are currently not practically feasible within the provided API. Since this is a practical benchmark, we must accept that we cannot guarantee a perfect overlap between what is allowed and what is possible. However, we may modify the API in future iterations of this benchmark to accommodate a larger set of allowed modifications.

#### 4.2.1 ISOLATE THE TRAINING ALGORITHM

A training algorithms benchmark should prohibit modifications outside of the training algorithm in order to disentangle improvements due to the training algorithm from other beneficial pipeline changes (see [Section 2.2.2](#)). However, exactly what constitutes part of the training algorithm is not always clear. The distinction between the model and the training algorithm can be quite subtle, as well as the distinction between improving the implementation quality of the submission versus improving its software dependencies in a way that could apply to all submissions.

**Model vs. training algorithm.** There is not always an obvious separation between the model and the training algorithm. In our benchmark, we use the basic rule of thumb that anything that can be applied to a generic workload is part of the training algorithm. On the other hand, if something only applies to some workloads or is otherwise inherently workload-specific, it should not be considered part of the training algorithm.

One delicate area is regularization. Because our benchmark focuses on validation and test set performance, we need to include regularization as part of the training algorithm in our benchmark and allow submitters at least some control over it since some training algorithms may implicitly have a regularizing effect. However, submissions can only be allowed complete control over model-agnostic regularization. Therefore, we allow submissions to tune the *strength* of regularization methods predefined by the workloads, and we do not allow submissions to introduce new regularization *strategies* that require modifications to the data preprocessing or model architecture. For example, submissions can set or tune the dropout rates of models that already have dropout layers. However, they cannot introduce additional dropout layers, as this would require knowledge of the model architecture. Similarly, the submitted training algorithms receive raw loss values and can add any workload-agnostic regularization term, e.g.,  $L^2$  regularization of the desired strength.

Data augmentation strategies often improve generalization performance. However, most data augmentations strongly depend on the input data modality and are therefore workload-dependent. Submissions cannot introduce new data augmentation strategies to the benchmark workloads. However, they do have control over batching and could potentially filter, reorder, or otherwise prioritize training data.

Another tricky technique to handle is batch normalization, which can be seen either as a training algorithm component or—as it is commonly implemented—as a model layer and thus part of the model architecture. In our benchmark, submissions cannot introduce additional batch normalization layers to the workloads or change the location of existing normalization layers. However, they do have control over whether the batch normalization statistics should be updated. This is important, for example, for line search approaches that search over multiple candidate updates before applying the most appropriate one. In this case, such a submission might not want to update the batch normalization statistics after each candidate evaluation, but only once an acceptable point has been selected.

**Introducing software dependencies.** Submissions are not allowed to use software engineering approaches to speed up low-level, primitive operations in JAX, PYTORCH, their dependencies, or the operating system. For example, it is prohibited to introduce new compiler functionality, using faster GPU kernels, or make similar modifications that could generally benefit any submission. Submissions must also use the versions of PYTORCH or JAX (and their dependencies) specified by the benchmark.

Submissions are free to add software packages that support novel algorithmic and mathematical ideas, as long as they do not circumvent the intention of the benchmark. For example, submitters are allowed to use packages such as BACKPACK (Dangel et al., 2020), which extracts additional information from the backward pass. We also recognize that the way a method is implemented will impact its performance in the benchmark, and it is generally acceptable to make clever and efficient use of the JAX and PYTORCH APIs from within the submission functions. For example, it would be acceptable to use CUDA streams to schedule the transfer of data from CPU to GPU while performing other computations. However, under the rules there are periodic untimed model evaluations that do not contribute to the submissions score and it would not be acceptable for a submission to schedule asynchronous computations that are being performed during these untimed evaluations.

#### 4.2.2 INCENTIVIZING GENERALLY USEFUL SUBMISSIONS

We want to disallow submissions if they clearly violate the spirit of the benchmark, even if these submissions perform well in our benchmark. Most importantly, this includes overly benchmark-specific methods that cannot be applied to generic deep learning workloads.

It is impossible to define rules that clearly distinguish between allowed and prohibited submissions in all possible scenarios. This section provides some guidelines to clarify whether or not a submission violates the spirit of the rules, and thus should be disqualified by the spirit jury. As a rule of thumb, a submission should be allowed if it will run and do something reasonable on unseen workloads without requiring additional human engineering effort. Two essential questions can guide this distinction:

1. What **information** is used by the submission?
2. What **action** is the submission code taking based on this information?

Generally, both parts are needed to decide if a particular piece of code is within the spirit of the rules. Below are some specific examples intended to illustrate the policy. Additional cases of allowed and disallowed submissions, along with further clarifications, can be found in the complete rules.<sup>10</sup>

**Using shape and layer information of the model.** Submissions may use the provided model parameter shape information if the resulting action can be applied to generic workloads. Some examples of allowed uses include (a) using the shape information of each layer to switch between a high-memory and a low-memory routine, (b) using different update rules (e.g. ADAM and SGD) for different layer types (e.g. convolutional layer, batch normalization layer, fully-connected layer), or (c) leveraging the order of the layers to train layers in an organized fashion.

10. See the “Valid Submission” Section in <https://github.com/mlcommons/algorithmic-efficiency/blob/main/RULES.md#valid-submissions>.



However, submissions may not use this same information to identify the specific workload and use workload-specific settings. A clear case of using this information in a way that violates the spirit of the benchmark would be using the shapes of the model parameters as a workload “fingerprint” and then loading or looking up the predetermined optimal hyperparameters. In general, any hard-coded behavior based on identifying a specific workload is prohibited. At the same time, it is entirely acceptable to take action based on basic modules such as the layer type. In other words, a submission may run different code whenever it encounters a model with convolutional layers, but it shouldn’t need to specifically know that it is training a ResNet-50 on ImageNet.

**Using expensive offline computations.** Submissions may not circumvent the tuning rules by looking up the workload-specific results of offline computations that have been performed ahead of time. This includes looking up optimal hyperparameter settings for each specific workload or even looking up (pre-trained) model parameters.

In contrast, it is perfectly fine to hard-code a single hyperparameter setting, e.g., a default hyperparameter setting, even when found using an expensive offline search because the hyperparameter will need to perform well on all workloads simultaneously and thus could be expected to have some hope of generalizing to new workloads. We also allow submissions based on learned training algorithms, which may include using a learned set of hyperparameters. In this case, we ask submitters to disclose information about the training set used to develop these learned training algorithms.

### 4.3 Workloads

A *workload* consists of a dataset (including any preprocessing), model, and loss function, along with a target that is defined using some evaluation metric. This evaluation metric can be identical to the workload’s loss function, or it could be a workload-specific metric such as the *word error rate* (WER) or the *mean average precision* (mAP). For example, training RESNET-50 on IMAGENET using the *cross-entropy* loss (CE) until a target error of 34.6% on the test set has been reached, would constitute one workload.

A diverse set of workloads that reflect multiple important application areas is necessary to assess the suitability of an algorithm as a general-purpose training algorithm (Section 2.2.2). We selected our list of workloads to cover several different tasks and data modalities, including image classification, machine translation, speech recognition, and other typical machine learning tasks, focusing on today’s practically relevant workloads. The set of workloads will need to be extended in future iterations of the benchmark in order to remain relevant and to match advancements in the field. We intentionally restricted our current set of workloads to supervised learning tasks (although we could easily accommodate self-supervised tasks) and excluded reinforcement learning problems that might require fundamentally different methods and evaluation protocols.

#### 4.3.1 FIXED AND RANDOMIZED WORKLOADS

In service of identifying generally useful training algorithms, our benchmark includes two types of workloads: fixed workloads and randomized workloads. *Fixed workloads* are fully specified in the benchmark and completely known to the submitters. Table 7 provides an overview of the 8 *fixed* workloads used in the first iteration of this benchmark (Appendix D

contains the individual workload details). Additionally, we also provide *randomized workloads* which are only defined as distributions over workloads. Once all the benchmark submissions are frozen, we will sample specific instances from these randomized workloads that we call *held-out workloads*. A submission’s score is a function of its performance on the fixed workloads as well as these held-out workloads.

Task	Dataset	Model	Loss	Metric	Validation Target	Test Target	Maximum Runtime
Clickthrough prediction	rate CRITEO 1TB	DLRMSMALL	CE	CE	0.123649	0.126060	7703
MRI reconstruction	FASTMRI	U-NET	L1	SSIM	0.7344	0.741652	8859
Image classification	IMAGENET	RESNET-50	CE	ER	0.22569	0.3440	63,008
		ViT	CE	ER	0.22691	0.3481	77,520
Speech recognition	LIBRISPEECH	CONFORMER	CTC	WER	0.078477	0.046973	101,780
		DEEPSPEECH	CTC	WER	0.1162	0.068093	92,509
Molecular property prediction	OGBG	GNN	CE	mAP	0.28098	0.268729	18,477
Translation	WMT	TRANSFORMER	CE	BLEU	30.8491	30.7219	48,151

Table 7: **Summary of the *fixed* workloads used in our benchmark.** The possible losses are the cross-entropy loss (CE), the mean absolute error (L1), and the Connectionist Temporal Classification loss (CTC). The evaluation metrics additionally include the structural similarity index measure (SSIM), the error rate (ER), the word error rate (WER), the mean average precision (mAP), and the bilingual evaluation understudy score (BLEU).

Each randomized workload introduces minor modifications to an associated fixed *base* workload. These modifications include, for example, altering the data augmentation strategies or modifying aspects of the model architecture, such as the activation function or the number of layers. Each randomized workload defines a *distribution* over workloads. For the first iteration of the benchmark, for convenience, we used particularly simple discrete distributions that sample one concrete workload variant out of a set of three possible, hand-designed variants of the base workload. Only once all submissions are frozen do we select the specific instance of the held-out workload that will be used during scoring. Thus, although submitters know the three possible variants that might be sampled for each randomized workload, they will not know which of the three will be sampled during scoring. Defining distributions instead of directly specifying instances ensures that while the entire process is public and transparent, neither the submitters nor the benchmark organizers know the specific held-out workloads beforehand.

The random held-out workloads function similarly to a held-out test set, discouraging training algorithms that overfit to the fixed workloads. The randomized workloads are intended to simulate novel-but-related workloads to ensure that the proposed training algorithms generalize to unknown problems. Although submitters can enumerate all possible held-out workloads, by creating a larger set of possibilities than the fixed workloads alone, the randomized workloads should encourage more robust training algorithms. Ideally, the randomized workloads should strike a balance between modifying a base workload enough to generate a different workload, but not so much that they produce something impossible



to train. If the modifications are too conservative, then the held-out workload will simply be a copy of the base workload and not provide a generalization challenge that simulates novel workloads. However, if the modifications are too drastic, then the workload could lose its practical relevance.

Consequently, our randomized workload distribution should only introduce “natural changes” that a practitioner might want to experiment with. These include modifications such as changing the activation function, the type of normalization layer, or modifying the number and width of the layers. Any modification that actually improves results is *prima facie* natural, as are modifications that occur in the literature or are obvious extensions of common practice. We want to avoid extremely contrived changes purely designed to make the training problem harder. That said, we are not above changes that might arise based on an insufficiently careful initial weight distribution, badly scaled parameters, or other changes of that nature. At the end of the day, we want changes that may elucidate robustness properties of training algorithms that actually provide value for practitioners.

For the benchmark’s first iteration, we manually designed three different workload variants for each fixed workload (see [Table 11](#)) from which to draw held-out workloads.<sup>11</sup> For each fixed workload, one of these workload variants will be randomly selected after submissions are frozen. The randomized workloads use the same procedure as the fixed workloads to define validation and test targets ([Section 4.1](#)). However, to save computational resources, we only tuned two training algorithms for the randomized workloads, NADAMW and the other best-performing training algorithm on the corresponding base workload.

In total, a training algorithm submitted to the benchmark is evaluated on 8 fixed and 8 held-out workloads. However, scoring uses the held-out workloads only to penalize submissions that can’t handle them, and reserves the fixed workloads for timing measurements. [Section 4.5.3](#) describes the precise way held-out workloads affect submissions scores, but, at a high level, we wanted to prioritize the fixed workloads for timing measurements since they are the most relevant variant and merely use the held-out workloads as a deterrent for brittle submissions. Finally, [Section 6.2](#) describes the experimental protocol for selecting possible workload variants to serve as components of randomized workloads.

#### 4.4 Tuning

Given our goal of evaluating generally applicable training methods, any workload-specific adaptation (or tuning) should be an automated part of the algorithm. In our benchmark, we provided two tuning rulesets that govern how hyperparameters may be tuned. In the more permissive, external tuning ruleset ([Section 4.4.1](#)), each submission may define a list of hyperparameters along with a search space to tune them over. To evaluate a submission using external tuning, we tune its hyperparameters using quasirandom search (with a modest, fixed budget of tuning trials) over the submission’s search space. The runtime for a submission under this ruleset on a given workload is then the best-performing (fastest to reach the target) hyperparameter setting for that workload across the tuning trials. In contrast, the more restrictive, self-tuning ruleset ([Section 4.4.2](#)) does not allow any tun-

---

11. In preliminary experiments, we briefly attempted to construct randomized workload distributions with support over much larger sets of concrete workloads before deciding to take a simpler approach for the first iteration of the benchmark. See [Appendix E](#) for information on these preliminary experiments.

ing outside of the timed operation of the submission itself on the benchmarking system. Instead, the training algorithm must be hyperparameter-free or, equivalently, it must automatically set any hyperparameters it happens to define. The main difference between these two rulesets is that in the self-tuning ruleset every part of the algorithm—which may include any arbitrary internal tuning procedure—is performed “on the clock.” In contrast, the external tuning ruleset allows for some parallelization of hyperparameter tuning, where only the fastest hyperparameter setting for each workload is used for scoring.

Both rulesets are essentially independent competitions, where submissions only compete with methods adhering to the same ruleset. The two rulesets cover two different practical scenarios: tuning on a single machine sequentially, and tuning in parallel in one shot across a modest number of machines. Any valid submission for the self-tuning ruleset is valid under the external tuning ruleset, but not vice-versa. Both tuning rulesets share all other non-tuning benchmark rules, unless otherwise specified. Below we describe the each version of the tuning rules in more detail.

#### 4.4.1 EXTERNAL TUNING

For each workload and each submission requiring external tuning, the hyperparameters are tuned using 20 tuning *trials* drawn with quasirandom search (Bousquet et al., 2017) from the workload-agnostic search space specified in the submission. In lieu of independent samples from their search space, submissions can instead supply a fixed list of 20 hyperparameter points that will be sampled without replacement.<sup>12</sup> Using more than 20 tuning trials would increase the computational burden of the benchmark. At the same time, we want the external tuning ruleset to be sufficiently different from the self-tuning ruleset, which effectively uses a single tuning trial. To produce lower variance scores, the rules require repeating the tuning for five independent *studies*, resulting in a total of 100 trials. For each of the five studies and for each workload, the hyperparameter setting that reaches the validation target the fastest will be selected among the 20 tuning trials. For each workload, the score for a submission is the median of these five per-study training times and becomes an input into the overall benchmark score (Section 4.5.2). While we use the time to reach the *validation* set target for selecting the hyperparameter point, we use the time to reach the *test* set target for that hyperparameter setting for scoring. The five independent studies effectively simulate five hypothetical independent practitioners training the algorithm on a workload using the same search space. Using the median score allows us to assess what training speed the average practitioner can expect from this method.

#### 4.4.2 SELF-TUNING

Submissions to the self-tuning ruleset are not allowed to have user-defined hyperparameters and therefore receive no extra tuning. Instead, the training algorithms in this ruleset need to perform all necessary adaptations to the workload autonomously. This adaptation could, for example, come in the form of inner-loop tuning, e.g. line search approaches, sequential outer-loop tuning, freeze-thaw methods, or algorithms that use the same hyperparameters

---

12. Following Metz et al. (2020), we also refer to this approach as a “learned optimizer list” or, abbreviated, as an OPTLIST.

for all workloads, e.g. ADAM with its default parameters. Any tuning effort will be part of the per-workload score and thus any tuning should save more time than it costs.

Compared to the external tuning ruleset, there are no tuning trials but only a single run per study. Once again, the median training time of the 5 studies represents the per-workload score and is incorporated into the benchmark score. Since we do not use any (external) tuning in this ruleset, the time to reach the validation target is irrelevant and only the time to the test target is considered. To account for the lack of external tuning, submissions have a longer per-workload time budget. Compared to the external tuning ruleset, all maximum allowed runtimes are tripled, i.e.  $3\times$  the maximum runtime reported in [Table 7](#).

At this time, we do not anticipate self-tuning submissions will be competitive with externally tuned methods in terms of their per-workload score. However, closing the gap between fully automatic methods in the self-tuning ruleset and more traditional externally tuned methods could drastically reduce the compute requirements of deep learning.

## 4.5 Scoring and Reporting Results

So far, we have described how to evaluate a submission on a workload and produce a raw training time score. Suppose that for each submission  $s$  and each workload  $w$ , we determine a training time  $t_{s,w} \in [0, \infty)$  using one of the rulesets described in [Section 4.4](#). This training time is the wall-clock runtime it took the submission to first reach the test target on this particular workload. In the case of the external tuning rule set, we measure the time it takes to reach the test target *only* for the hyperparameter setting that achieved the validation target the fastest. If a submission is unable to reach the target within the given runtime budget, it will receive a score of  $t_{s,w} = \infty$  for this particular workload. These raw training time scores are already useful if we care about a single workload, e.g., by ranking submissions on the benchmark workload that is most similar to the real-world problem we are trying to solve. However, the raw training time scores are too detailed in most situations.

Depending on our goal, we will need different ways of summarizing the raw training time data and converting it into scores, visualizations, and summary statistics. In general, in addition to providing guidance to practitioners that only care about a single workload, we need the results we report to help us:

1. Qualitatively compare submissions across all workloads at once,
2. Construct a leaderboard for submissions that aggregates across workloads, and
3. Summarize year-over-year benchmark progress.

### 4.5.1 AGGREGATION USING PERFORMANCE PROFILES

A natural way to compare a given submission  $s$  within a larger pool of competing submissions is to look at the fraction of workloads where  $s$  trains the fastest (i.e.,  $t_{s,w}$  is the smallest). However, if there are workloads where a few strong submissions get nearly the same raw training time score, it might make sense to look at the fraction of workloads where  $s$  is either the fastest or *nearly* the fastest by being within, say, 1% of the runtime of the best submission on that workload. Since our choice of 1% is arbitrary, we also might ask the same question with different notions of whether a submission is sufficiently close to the fastest. Performance profiles ([Dolan and Moré, 2002](#)) conveniently generalize this

idea. Specifically, a performance profile is a plot of the fraction of workloads where a given method is within some ratio of the best per-workload training time.

Performance profiles are straightforward to compute given the raw training times,  $t_{s,w}$ , for a set of  $k$  submissions  $\mathcal{S} = \{s_1, s_2, \dots, s_k\}$  measured on a set of  $n$  workloads  $\mathcal{W} = \{w_1, w_2, \dots, w_n\}$ . Specifically, for a submission  $\bar{s}$  on a particular workload  $\bar{w}$  we define its performance ratio as:

$$r_{\bar{s},\bar{w}} = \frac{t_{\bar{s},\bar{w}}}{\min_{s \in \mathcal{S}} t_{s,\bar{w}}}. \quad (1)$$

In other words, the performance ratio  $r_{\bar{s},\bar{w}}$  expresses how much time the submission  $\bar{s}$  took on workload  $\bar{w}$  compared to the best submission on  $\bar{w}$ . For example, if a submission takes twice as long on a particular workload compared to the best submission, it will receive a performance ratio of 2. By definition,  $r_{\bar{s},\bar{w}} \geq 1$  for all submissions  $\bar{s}$  and workloads  $\bar{w}$ .

The performance profile  $\rho_{\bar{s}}(\tau)$  for a submission  $\bar{s}$  is the probability that, on a random workload  $w$  drawn uniformly from  $\mathcal{W}$ ,  $\bar{s}$  will have a performance ratio  $r_{\bar{s},w}$  of at most  $\tau$ . Specifically, for  $\tau \in [1, \infty)$  the performance profile is defined as  $\bar{s}$ :

$$\rho_{\bar{s}}(\tau) = \left(\frac{1}{n}\right) \cdot |\{\bar{w} : r_{\bar{s},\bar{w}} \leq \tau\}|. \quad (2)$$

Since  $\rho_{\bar{s}}(\tau)$  expresses the fraction of workloads where a submission is less than  $\tau$  away from the optimal submission, it is piecewise constant, monotonically non-decreasing with  $\tau$ , and bounded between 0 and 1. A perfect submission that always achieves the fastest training time on every workload would have a performance profile that immediately jumps to 1 at  $\tau = 1$ . [Figure 5](#) in [Section 7](#) shows performance profiles for our baseline submissions.

Performance profiles are a convenient way to summarize the overall performance of a submission relative to other submissions on the benchmark workloads. Since performance ratios are relative to the best submission on each workload, performance profiles are also inherently relative and will change as submissions are added or removed from the comparison set. Although performance profiles are great to qualitatively compare submissions, to construct a leaderboard we need a single scalar benchmark score.

#### 4.5.2 INTEGRATING PERFORMANCE PROFILES FOR THE BENCHMARK SCORE

To calculate the scalar benchmark score  $B_{\bar{s}}$  of a submission  $\bar{s}$ , we integrate its performance profile up to a maximum ratio  $r_{\max}$ :

$$B_{\bar{s}} = \frac{1}{r_{\max} - 1} \int_1^{r_{\max}} \rho_{\bar{s}}(\tau) d\tau. \quad (3)$$

Since we normalize by  $r_{\max} - 1$ ,  $B_{\bar{s}} \in [0, 1]$ , with higher benchmark scores being better. A benchmark score of  $B_{\bar{s}} = 1$  would indicate that the submission  $\bar{s}$  was the fastest on every workload. [Table 13](#) presents the benchmark scores of the baselines shown in [Section 7](#).

We set the upper integration limit to  $r_{\max} = 4$ , which also serves as the right-hand limit of any performance profile plot. This choice means that any submission that requires more than four times the runtime of the fastest submission will not get any credit on this workload and will be treated the same as a training algorithm that is unable to successfully reach

the target within the maximum allowed runtime budget. Although the exact integration bound is somewhat arbitrary, we want to encourage algorithms that are robust to different workloads, and in practice we are likely to rank the best-performing submissions similarly for most reasonable values of  $r_{\max}$ . If there exists a specialized algorithm that is four times faster than any generic training algorithm on a particular problem, it seems likely that practitioners will prefer it even if it is only useful for a small number of problems.

#### 4.5.3 USING HELD-OUT WORKLOADS IN SCORING

The benchmark score computation is based on a performance profile over only the fixed workloads. However, we penalize submissions that perform poorly on the held-out workloads. If a submission does not perform well enough on a given held-out workload, then we score the submission on the corresponding fixed workload as if that submission did not reach the fixed-workload target. Specifically, for a submission to get credit for a finite training time on a particular fixed workload, it must:

1. Reach the validation and test target on the fixed workload within the runtime budget.
2. Reach the validation and test target on the fixed workload within  $4\times$  of the fastest submission.
3. Reach the validation and test target on the held-out workload corresponding to the fixed workload within the maximum runtime.
4. Reach the validation and test target on the held-out workload corresponding to the fixed workload within  $4\times$  of the fastest submission.<sup>13</sup>

Only if all four requirements are met does the submission get credit for a finite score on that particular workload. Otherwise, a submission will have an infinite training time on the fixed workload. This rule means that being unable to successfully train a held-out workload will disqualify a submission from getting credit for a good training time on the corresponding fixed workload. We thus require submissions to be robust enough to not completely fail when faced with minor workload variations. This ensures that we prioritize the fixed workloads for scoring since they are the most relevant version of that workload in practice. However, we also protect the benchmark from egregious workload-specific tuning and penalize brittle methods that break with slight modifications of the workload.

#### 4.5.4 MEASURING YEAR-OVER-YEAR BENCHMARK PROGRESS

Scores derived from performance profiles make sense for relative comparisons within a set of submissions, but the benchmark scores  $B_s$  of the winning submission between different iterations of the benchmark tell us nothing about how much faster training algorithms have become in an absolute sense. In order to measure year-over-year progress in reducing training time, we can use the geometric mean across workloads of the raw training times for the best submission.

---

13. To determine the fastest submission on a held-out workload, we only consider submissions that reached the target on the corresponding fixed workload. This protects us against extremely fast submissions that only work on a specific held-out workload and are useless as general algorithms.

In general, we recommend that anyone who reports results on the benchmark should report raw training times in addition to any performance profiles and benchmark scores. The raw scores allow other researchers to include the submissions in their own performance profile comparisons or compute geometric means of any speedups if they want to measure progress through time. Ideally, the raw times would be measured on the official competition hardware. However, if practical considerations require using a different system, additional baseline results from previous work should also be reported to facilitate comparisons. In the event that future iterations of the benchmark change the competition hardware, we can run the most crucial previous submissions on the new system.

## 5. Target-Setting Experiments

As a time-to-result benchmark (see [Section 4.1](#)), we need to set validation and test targets for all workloads. To set the targets, we considered 4 training algorithms, ADAMW, NADAMW, NESTEROV, HEAVY BALL. We tuned all four target-setting training algorithms over relatively broad search spaces (see [Table 8](#) for exact search spaces). For each algorithm, we sampled 200 trials quasirandomly from its search space and selected the trial with the best validation metric. [Table 9](#) shows these resulting validation evaluation metric values for each algorithm, with the boldface values denoting the best for each workload.

As mentioned in [Section 4.1](#), each algorithm ran for  $0.75 \times$  the maximum runtime shown in [Table 7](#). More precisely, since these four algorithms happen to have nearly identical step times, in these experiments we used a maximum number of steps as a proxy for runtime.<sup>14</sup> We determined the maximum number of steps that would fit within the given runtime budget, and used this same number of steps for every training algorithm.

Hyperparameter	AdamW	NadamW	Heavy Ball	Nesterov
Base LR	Log [1e-5,1e-1]	Log [1e-5,1e-1]	Log [1e-3,10]	Log [1e-3,10]
Weight decay	Log [1e-5,1]	Log [1e-5,1]	Log [1e-7,1e-2]	Log [1e-7,1e-2]
$1 - \beta_1$	Log [1e-3,1]	Log [1e-3,1]	Log [1e-3,1]	Log [1e-3,1]
$1 - \beta_2$	Log [1e-3,1]	Log [1e-3,1]	NA	NA
Schedule	warmup + cosine decay	warmup + cosine decay	warmup + linear decay + constant	warmup + linear decay + constant
Warmup	{2%, 5%, 10%}	{2%, 5%, 10%}	5%	5%
Decay factor	-	-	{1e-2, 1e-3}	{1e-2, 1e-3}
Decay steps	-	-	Linear [0.8, 1.0]	Linear [0.8, 1.0]
Dropout	{0.0, 0.1}	{0.0, 0.1}	{0.0, 0.1}	{0.0, 0.1}
Aux. dropout	{0.0, 0.1}	{0.0, 0.1}	{0.0, 0.1}	{0.0, 0.1}
Label smoothing	{0.0, 0.1, 0.2}	{0.0, 0.1, 0.2}	{0.0, 0.1, 0.2}	{0.0, 0.1, 0.2}

Table 8: **Hyperparameter search spaces for the target-setting training algorithms.** Descriptions of the learning rate schedules can be found in [Appendix A.1](#). The regularization hyperparameters are tuned only for those workloads where they are applicable.

14. On all workloads, the step time differences between these four optimizers are negligible compared to the time required to complete one gradient calculation.

Workload	Criteo 1TB DLRMsmall	fastMRI U-Net	ImageNet ResNet-50	ImageNet ViT
Metric	CE↓	SSIM↑	Error Rate↓	Error Rate↓
ADAMW	0.123675	0.734330	0.23034	0.22614
NADAMW	<b>0.123609</b>	0.734523	0.22702	<b>0.22534</b>
HEAVY BALL	0.125913	0.733828	<b>0.22534</b>	0.24486
NESTEROV	0.126139	<b>0.734645</b>	0.22660	0.24318

Workload	LibriSpeech Conformer	LibriSpeech DeepSpeech	OGBG GNN	WMT Transformer
Metric	WER↓	WER↓	mAP↑	BLEU↑
ADAMW	0.078327	0.114152	0.277534	30.6939
NADAMW	<b>0.077790</b>	<b>0.113950</b>	0.280012	<b>30.8534</b>
HEAVY BALL	0.132797	0.161977	0.276148	30.6431
NESTEROV	0.130823	0.171137	<b>0.283124</b>	30.1074

Table 9: **Performance of the target-setting training algorithms on each of the workloads.** Each entry contains the validation performance for the best tuning trial for a given algorithm and workload. We often see some “reversion to the mean” when rerunning this trial 20 times (cf. [Table 23](#)). The boldface values indicate the best result in each column.



After running the 200 hyperparameter trials, for every workload we took the best configuration (training algorithm and hyperparameter settings) on the validation set (detailed in Table 10) and retrained it 20 times with different random seeds. The final validation targets were the median values achieved over these 20 repetitions, while the test targets were the worst-case test set performance achieved across those 10 repetitions that hit the validation target. The results of these 20 repetitions, for each workload, along with the associated training algorithm, hyperparameter settings, and additional statistics are included in Tables 23 and 24 in Appendix B. Table 7 shows the final targets for all fixed workloads.

	Criteo 1TB fastMRI		ImageNet		LibriSpeech		OGBG	WMT
	DLRMsmall	U-Net	ResNet-50	ViT	Conformer	DeepSpeech	GNN	Transformer
Algorithm	NADAMW	NESTEROV	HEAVY BALL	NADAMW	NADAMW	NADAMW	NESTEROV	NADAMW
Base LR	0.003331	0.028609	4.131896	0.000844	0.001308	0.004958	2.491773	0.001749
Weight decay	0.003578	0.000577	5.67e-6	0.081354	0.163753	0.114739	1.29e-7	0.081216
$\beta_1$	0.948	0.981543	0.927476	0.889576	0.973133	0.863744	0.944937	0.932661
$\beta_2$	0.998793			0.99785	0.998123	0.629185		0.995516
Warmup	2%	5%	5%	5%	10%	2%	5%	2%
Decay factor	-	0.01	0.001	-	-	-	0.001	-
Decay steps	-	0.984398	0.900777	-	-	-	0.861509	-
Dropout	0.1	0	0	0	0	0	0.1	0.1
Aux. dropout	-	-	-	-	0	0.1	-	0.1
Label smoothing	-	-	0.2	0.2	-	-	0	0

Table 10: **Optimal training algorithm and hyperparameter settings used for target setting on each workload.** The reported configuration (training algorithm and hyperparameter setting) achieved the best performance on the validation set.

For the randomized workloads, we used nearly an identical procedure to set targets for each workload variant. To save computational resources, we only tuned two training algorithms instead of four. For each workload variant, we used NADAMW and the other best-performing training algorithm on the corresponding base workload.

**Comparison to results in the literature** The ideal target-setting procedure would produce targets that compare favorably to results reported in the literature for similar setups, although in many cases it will not be possible to find published results that are an exact match for our workloads (i.e. training budget, evaluation metrics, data preparation, or model architecture might be different). Therefore, this comparison is neither a comprehensive literature review, nor—by itself—a direct assessment of our target-setting procedure. Instead, our goal is to provide context for a holistic evaluation of the relevance and competitiveness of our targets.

- **Criteo 1TB** On this workload, our target-setting procedure resulted in a validation target (binary cross entropy) loss of 0.123649 and a test target loss of 0.126060. In comparison, Sterbenz (2017) reported a combined loss on our validation and test sets of 0.1250 (our combined loss would be 0.1248545), leading to a reported AUC score of 0.8002. It is worth noting that Sterbenz (2017) used a larger model and trained it for more than three times as long as we did. NVIDIA (2023) trained a model with the



same dimensions as the one used in our workload for one epoch and reported AUC scores on our validation set between 0.802509 and 0.802784. However, it is important to note that their test datasets used frequency thresholding, which we did not apply.

- **fastMRI** The target-setting procedure achieved an SSIM of 0.7344 on our validation set and 0.741652 on our test set. The paper introducing the FASTMRI benchmark provided a U-NET baseline with an SSIM score of 0.72 on our combined validation and test set (Zbontar et al., 2018, Table 8, note that model selection was done using NMSE and not SSIM), compared to our slightly better combined score of 0.738. In the 2019 FASTMRI challenge (Knoll et al., 2020), the winning submission for the single-coil knee dataset achieved an SSIM of 0.751 on our combined validation and test set, using a custom i-RIM model (Putzky et al., 2019).
- **ImageNet** For the RESNET-50 workload, our target error rate on the validation set of 22.57% improves over the performance reported in the original RESNET paper by He et al. (2016a) (their most similar setup reaches 24.7%). More recent studies have achieved lower error rates by using variants of the original RESNET architecture or by employing improved training recipes. Bello et al. (2021, RESNET-RS-50 in Table 7) report a validation error of 21.2%, and Wightman et al. (2021, A2 in Table 1) report an error of 20.2% when training for roughly  $2.67\times$  the training budget of our workload. With a slightly shorter training time than our workload, Wightman et al. (2021, A3 in Table 1) achieved an error rate of 21.9% using a lower training resolution, RandAugment, CutMix, and Mixup compared to our workload. For the ViT workload, our target-setting procedure yielded a validation error rate of 22.69%. The paper introducing Vision Transformers reported an error rate of 22.09% for a slightly larger model trained exclusively on IMAGENET (Dosovitskiy et al., 2021, ViT-B/16 in Table 5). We can compare our results with a training budget of roughly 112 epochs, to Beyer et al. (2022), which reported error rates of 23.5% and 21.5% for the same ViT-S/16 model used in our workload for a training budget of 90 and 150 epoch respectively (note that they used our validation set as a test set and trained only on 99% of the training data to use the remaining 1% as their validation set).
- **LibriSpeech** Our target-setting procedure for the LIBRISPEECH dataset resulted in a validation word error rate (WER) of 0.1162 for the DEEPSPEECH workload and 0.078477 for the CONFORMER workload. For easier comparison with the literature, we can consider the WERs on our test set (the `test_clean` split), which are 0.067976 and 0.046696 for DEEPSPEECH and CONFORMER, respectively. The original paper introducing the LIBRISPEECH dataset reported a baseline with a WER of 0.0551 (Panayotov et al., 2015, Table 3), although the model uses a completely different approach than our workloads. Subsequently, Amodei et al. (2016, Table 4) reported an improvement to 0.0515 with the DEEPSPEECH 2 model. The DEEPSPEECH 2 model uses a much larger training set beyond just LIBRISPEECH and uses beam search decoding, unlike our DEEPSPEECH workload. Gulati et al. (2020, Table 2, Conformer(S) without LM) introduced the CONFORMER architecture and reported 0.027 test WER. However, Gulati et al. (2020) used a much larger CONFORMER model, used beam search decoding instead of greedy decoding, trained for more steps, and included a few other more minor model differences.

- **OGBG** The target-setting procedure for the OGBG workload resulted in a mean average precision (mAP) on the validation set of 0.28098. When introducing the OPEN GRAPH BENCHMARK (OGB), [Hu et al. \(2020\)](#) also provided six baselines for the OGBG-MOLPCBA dataset used in our workload. The best baseline reached a validation mAP of 0.2798. Subsequent approaches were able to achieve higher results with different models and training techniques, such as a 0.3012 mAP by [Wang et al. \(2022\)](#), or 0.3252 mAP using additional training data ([Wang et al., 2021](#)).
- **WMT** For our WMT workload, we follow the setup adopted by the FLAX ([Heek et al., 2023](#)) WMT example. WMT datasets for different years aim to establish benchmarks for different problems in the machine translation domain, e.g. low-resource translation, domain & style of translation, or long-sequence translation. Consequently, over the past decade, the neural machine translation literature has used different combinations of WMT datasets depending on the desired language pair, recency of the data, amount of data, and underlying translation problem. Since our goal is to provide a training algorithms benchmark, we decided to stay close to a high quality open source example (in this case the FLAX WMT example). Specifically, the models are trained on “train” split of WMT2017 translation dataset ([Bojar et al., 2017](#)) for German to English (De → En), use the “dev” set from the WMT2014 dataset ([Bojar et al., 2014](#)) and use `newstest2014` as the test set. Our models achieve a BLEU score of 30.72 on `newstest2014`. [Gao et al. \(2022, Table 10\)](#) also evaluate their models on `newstest2014` De → En, and achieve higher BLEU scores (in the range 33.60 – 35.15). However, they pre-train the models in a bidirectional manner (De → En and En → De) before fine-tuning on the De → En translation direction. Additionally, [Ma et al. \(2023, Table 6\)](#) evaluate their models on `newstest2014` De → En direction and report BLEU score in the range 31.33 – 32.35, although they made architectural changes to the attention mechanism.

## 6. Randomized Workloads Experiments

The primary goal of our randomized workloads is to help deter brittle submissions that only work on the original—and relatively standard—fixed workloads, and instead encourage more robust training algorithms that also perform well on novel deep learning workloads. An additional concern is that by using only a small set of 8 (fixed) workloads, submissions in the external tuning ruleset (which are permitted to sample 20 different hyperparameter points per workload) could effectively perform workload-specific tuning. This risk is amplified since we permit submissions with a fixed hyperparameter list (i.e., OPTLIST approaches) instead of search spaces. For instance, a submission could use a hyperparameter list of the best hyperparameter configurations for each fixed workload. Such approaches could potentially lead to generally useful training algorithms and thus should be allowed. However, unless it generalizes to novel workloads the expensive offline computation it would require cannot be justified. Without held-out workloads, the benchmark would not test whether submissions that tune over a list of configurations that perform well on the fixed workloads can generalize to workloads outside of the 8 fixed workloads. To encourage robust submissions and avoid vitiating the limits on workload-specific tuning, we manually created three variations of each fixed, base workload. The three variants together form a randomized workload that

will be used to sample one specific concrete variant to use as a held-out workload during scoring. Since held-out workload sampling occurs after all submission code has been frozen, submissions may need to perform well on any of the possible variants.

### 6.1 Desiderata for Workload Variants

Although any workload change could potentially pose challenges for some hypothetical submissions, an ideal workload variant intended for use as part of a randomized workload would have the following properties.

**Representative of real workloads** The best workload variants would be as representative as possible of changes that occur in practice. Extremely contrived changes to the base workload do not help the community develop robust and general training algorithms. The most natural changes are ones that already occur in the wild (e.g., architectural changes described in the literature). However, when judging how natural a particular modification is for a workload, it is important to distinguish between changes that affect the optimization dynamics and more surface level architectural changes. Even if we would never expect a particular change to be applied in a real application, if it results in optimization dynamics that *do* occur in the wild, it might still be worth considering. The “attention temperature” change described below is arguably a bit contrived in terms of model architecture, but it reproduces a type of training instability that occurs in practice, especially as Transformer models become larger, and therefore training algorithms that handle it well could be useful (Gilmer et al., 2023; Kim et al., 2021; Deghani et al., 2023).

**Trainable in the original runtime budget** We want workload variants that are trainable. Furthermore, they should reach reasonable validation and test evaluation metric values *within the runtime budget of the base workload*. Specifically, there should exist a training algorithm with some hyperparameter setting that achieves good results. It is very easy to make changes that completely ruin a workload and produce a model incapable of performing well on its task no matter how it is trained, but these models are not useful. Exhibiting a configuration of the hyperparameters that achieves a good validation error on the original task proves that it is possible for the workload variant to yield a useful trained model.

**Distinct from the base workload (and other variants)** A workload variant only adds information to the benchmark if it is distinct enough from the original fixed workload it is based on. Ideally, the entire pool of variants of a given workload would be mutually distinct from each other—and the base workload—while each presenting an interesting new challenge for submissions. Although there are many ways to define distinctiveness, we tried to create variants that meaningfully changed the optimal hyperparameters of the training algorithms used during target setting. In this way, we hoped to encourage submissions that are easier to tune than current popular techniques that use straightforward search spaces, as represented by our target-setting algorithms.

Achieving all the desiderata listed above simultaneously, on command, is unfortunately difficult. Nonetheless, workload variants that meet these requirements definitely exist, especially if we make a few practical concessions, so giving up does not seem appropriate either. Originally, we hoped to specify randomized workloads via distributions with support over a very large number of variants by randomizing different pieces of the workload definition.

However, without a better scientific understanding of the effects of various workload modifications, constructing such distributions was simply too onerous (see [Appendix E](#) for details on some of our attempts), which is why we elected to manually design a small number of specific, concrete workload variants instead. Even this goal was more challenging than we expected, but *not* because it is impossible to construct modifications that require re-tuning the hyperparameters. Instead, the challenge in designing interesting workload variants comes from the conjunction of requirements that need to be achieved simultaneously.

## 6.2 Creating and Testing Workload Variants

While creating workload variants, we explored various modifications to the base workloads. The following families of modifications ended up being used in the benchmark variants:

- **Activation function:** Most base workloads employed ReLU as the activation function and we explored alternative activation functions such as GELU, SiLU, or TanH.
- **Pre-LN vs Post-LN:** For Transformer-based models, the base workload was usually the PRE-LAYER NORM (PRE-LN) ([Xiong et al., 2020](#)) version. We changed these to POST-LAYER NORM (POST-LN, see [Figure 3](#)).
- **Attention temperature:** For the WMT TRANSFORMER, we modified the attention layers to compute  $\text{Softmax}\left(\frac{cXW^Q(XW^K)^\top}{\sqrt{D/H}}\right)$  where  $c$  is a constant scalar denoting the attention temperature. The default self-attention implementation sets  $c = 1$ . In order to artificially induce instabilities similar to those faced by larger versions of these models, we set  $c = 1.6$ .
- **Initialization scales:** For the DLRM model, changing the scale of the initial weights of the embedding layer resulted in a variant. For the RESNET model, changing the initial batch normalization layer scale weights resulted in a workload variant.
- **Normalization layer:** We changed the type of normalization layer employed in the model. Common changes included interchanging batch normalization with layer normalization, as well as instance normalization with layer normalization.
- **Width, depth, and channels:** We explored changing model width, depth, and number of channels, as applicable.
- **Input pipeline:** On LIBRISPEECH, we found changing SPECAUGMENT strength to be an effective strategy.
- **Residual connection structure and scaling:** For the DLRM model, we created a variant with additional residual connections. For DEEPSPEECH, we removed residual connections from the model.
- **Pooling layer type:** Changing the pooling layer type from *global average pooling* to *max average pooling* resulted in a variant for the ViT workload.

We used a relatively permissive protocol to decide whether a candidate change to a workload produced an *acceptable* variant, since strictly achieving the strongest versions of all of our desiderata is quite difficult. Specifically, we tested whether the optimal hyperparameter setting for one of the more robust, standard algorithms (NADAMW) were sufficiently

different between the candidate variant and its base workload. We chose NADAMW as the training algorithm to measure variant distinctiveness because, in our experiments with our search spaces, its optimal hyperparameters tended to transfer much better than, for instance, HEAVY BALL or NESTEROV. NADAMW also happened to be the algorithm that set targets on the most workloads (5/8).

In addition to testing distinctiveness, we also rejected variants that degraded validation performance too much. As a rule of thumb, we accepted workload variants for which we could achieve a performance within 10% of the original validation performance (in the original target setting budget). More precisely, given a base workload  $w$  and a candidate workload variant  $w'$ , we re-ran the exact same collection of 200 hyperparameter settings that were used for target-setting (Table 8) for NADAMW on the candidate variant. Let  $H = \{h_1, \dots, h_{200}\}$  denote this set of points in hyperparameter space. Furthermore, let  $H(w) = \{h_i(w)\}_{i=1}^{200}$ , where  $h_i(w)$  denotes the validation evaluation metric of hyperparameter setting  $h_i$  on workload  $w$ . Let  $h^*(w)$  and  $h^*(w')$  denote the best hyperparameter setting for  $w$  and  $w'$ , respectively. Additionally, let  $\text{rank}(h, H(w))$  denote the rank of the hyperparameter setting  $h$  in the set  $H(w)$  when the elements are ordered according to validation performance on workload  $w$  (lower ranks indicate better performance). In general, we sought variants  $w'$  for which the following quantity was as large as possible:

$$\min \{ \text{rank}(h^*(w), H(w')), \text{rank}(h^*(w'), H(w)) \}.$$

In other words, when both ranks are greater than  $c$ , then the optimal hyperparameter setting on  $w$  performs worse on  $w'$  than at least  $c$  other hyperparameter points, and the optimal hyperparameter setting on  $w'$  performs worse on  $w$  than at least  $c$  other hyperparameters as well. Note, if  $w$  and  $w'$  share the same optimal hyperparameter, then these ranks will be 0. In the rest of this section, to simplify notation we write  $\text{rank}(w \rightarrow w')$  and  $\text{rank}(w' \rightarrow w)$  instead of  $\text{rank}(h^*(w), H(w'))$  and  $\text{rank}(h^*(w'), H(w))$ , respectively.

Ultimately, we used human judgement instead of strict thresholds to decide if  $\text{rank}(w \rightarrow w')$  and  $\text{rank}(w' \rightarrow w)$  were large enough for a given candidate variant and if it achieved a tolerable validation performance. In some cases (for the RESNET and CONFORMER models) we did not find variants that substantially changed the rank of the optimal hyperparameter point. In these cases, we fell back to creating variants that had different optimal base learning rates in simple one-dimensional learning rate sweeps. See Appendix D for additional details and results from our variant testing protocol.

Our procedure for testing and rejecting candidate variants is not guaranteed to produce a set of variants that achieve all of our desiderata. Although, by design, all our variants are at least somewhat representative of real workloads and are trainable in the original budget (perhaps with some degradation in performance), we potentially sacrificed distinctiveness. We made no attempt to ensure that different variants of the same base workload were *mutually* distinct from each other in terms of optimal hyperparameters, although we did not repeat the same modifications within a set. Furthermore, our operational definition of distinctiveness depends on the particular set of hyperparameter points we used and our choice of training algorithm. Finally, we did not always produce variants that made very drastic changes to the ranks of the best hyperparameter points.

### 6.3 Workload Variants of the Benchmark

Table 11 contains a brief description of the changes that produced the three variants of every fixed workload. See Appendix D for additional details about the changes and variants, as well as workload-specific results of the variant testing protocol (described above). Although it was hard to predict how candidate variants would perform in our tests, in hindsight some patterns emerged. For example, after switching the activation function from ReLU to GELU in the CONFORMER model, higher learning rates performed better (and in many cases the resulting model performance itself was improved).

We found producing variants which successfully changed the optimal NADAMW hyperparameters to be surprisingly difficult. This speaks to the robustness of NADAMW to workload variations, a desirable property that perhaps helps explain the success of preconditioned training algorithms. In contrast, the optimal learning rate for momentum methods is highly sensitive to seemingly trivial variations in the workload (e.g. the WIDE RESNET stride change experiment in Section 2.2.1). Note, this does not mean that the performance of NADAMW was completely robust to changes in the workloads. In fact, it was quite easy to design a workload variant which was more difficult for NADAMW to optimize. However, for many such variants the optimal NADAMW hyperparameters did not change.

There is at least one known case for which we expect the optimal NADAMW hyperparameters to change: when the batch size changes (Nado et al., 2021). However, this was not a valid option for us because the submission hardware is fixed and the submitter is allowed to choose the batch size. We were initially hopeful that other methods for increasing stochasticity in optimization problem would decrease the optimal NADAMW learning rate. Indeed, although it isn't part of the workload definition, changing the value of dropout can have this effect. However, in many cases when we tried to achieve similar effects by increasing the severity of data augmentation or adding label noise, we did not see a significant change in the optimal learning rate, although we did manage to hurt overall performance.

Another surprising failure was playing with residual scales. For example we tried modifying the traditional residual connection of  $x + F(x)$  to  $\alpha x + (1 - \alpha)F(x)$  for  $\alpha \in [0, 1]$ . By sweeping  $\alpha$  between 0 and 1, we can interpolate between a model with no residual connections ( $\alpha = 0$ ) to the default setting  $\alpha = 0.5$  to a model which ignores all intermediate blocks  $\alpha = 1.0$ —varying  $\alpha$  between these 3 extremes allows us to explore interesting variations of residual connections. Indeed we found  $\alpha$  to be a very important parameter for overall performance (somewhat surprisingly  $\alpha = 0.5$  was not always the optimal value), however the optimal learning rate for any given  $\alpha$  did not seem to change. We also tried  $\alpha x + F(x)$  which did not work either, though did improve performance for  $\alpha = 2$  and  $\alpha = 4$ .

Finally we would like to highlight some of our variants that actually outperform the base workload in validation error. The changes which we saw leading to improvement in models were changing from ReLU to GELU (RESNET), changing from ReLU to SiLU (RESNET and GNN), introducing GLU (ViT) and *removing* residual connections (DEEPSPEECH).

## 7. Baseline Submissions

We constructed baseline submissions using eight different training algorithm families: ADAMW, NADAMW, NESTEROV, HEAVY BALL, LAMB, ADAFACTOR, SAM(w. ADAM) and DISTRIBUTED SHAMPOO. For each training algorithm family, we designed one or more



Base Workload	Variant	Variant Description	Validation Target	Test Target
<b>Criteo 1TB</b>				
DLRMSMALL	EMBED INIT SCALE	Changes initialization scale of the embedding layer from $1/\sqrt{\text{vocab size}}$ to 1	0.124286	0.126725
	LAYERNORM	Adds layer normalization to the network	0.123744	0.126161
	RESIDUAL	Adds residual connections to the network	0.124027	0.126470
<b>fastMRI</b>				
U-NET	CHANNELS & POOLING	Increases number of channels and decreases number of pool layers	0.734376	0.741547
	TANH	Replaces all Leaky ReLU activations with TanH	0.729633	0.736727
	LAYERNORM	Replaces instance normalization with layer normalization with learnable parameters	<b>0.734861</b>	<b>0.741982</b>
<b>ImageNet</b>				
RESNET-50	SiLU	Replaces all ReLU activations with SiLU	<b>0.220090</b>	<b>0.342600</b>
	GELU	Replaces all ReLU activations with GELU	<b>0.220770</b>	<b>0.340200</b>
	BN INIT SCALE	Increases the scale of the initialization of batch normalization scale variables	0.234740	0.357700
ViT	POST-LN	Uses POST-LN instead of PRE-LN	0.246880	0.371400
	MAP	Changes pooling type from global to max average	0.228860	0.347700
	GLU	Include GLU in the MLPBLOCK	<b>0.223300</b>	<b>0.345500</b>
<b>LibriSpeech</b>				
CONFORMER	GELU	Replaces all ReLU activations with GELU	0.077958	0.047643
	LAYERNORM CHANGE	The LayerNorm before the final readout layer was removed	0.085371	0.053096
	ATTENTION TEMP	Increases attention temp from 1 to 1.6	0.082665	0.050168
DEEPSPEECH	TANH	Replaces all ReLU activations with TanH	0.133449	0.079810
	NO RESIDUAL	Removes residual connections.	<b>0.105042</b>	<b>0.060388</b>
	NORM & SPECAUGMENT	Removes decoder layer normalization layer & replaces batch normalization with layer normalization. Changes SPECAUGMENT specifications.	0.131553	0.082442
<b>OGBG</b>				
GNN	GELU	Replaces all ReLU activations with GELU	0.277710	0.262926
	SiLU ALTERED LAYERS	Replaces all ReLU activations with SiLU Adds a hidden layer, decreases latent dimension. Reduces the number of message passing steps and changes layer normalization to batch normalization.	<b>0.282178</b> 0.269446	<b>0.272144</b> 0.253051
<b>WMT</b>				
TRANSFORMER	POST-LN	Uses POST-LN instead of PRE-LN	30.2003	29.8982
	ATTENTION TEMP	Increases attention temperature from 1 to 4.0	30.0756	29.8094
	GLU & TANH	Uses GLUs in the MLP blocks and replaces all ReLU activations to TanH	30.0002	29.8139

Table 11: **Overview of workload variants used for randomized workloads.** See [Appendix D](#) for additional details of each workload variant. Targets that are better than the corresponding base workload target are in bold. Unsurprisingly, the variants where the validation targets improved were the same as the ones where the test targets improved.



search spaces to create valid submissions for the external tuning ruleset. For ADAMW, NADAMW, NESTEROV and HEAVY BALL, we compared submissions derived from search spaces that fixed the relevant first moment parameter ( $\beta_1$ ) to a default value with submissions that tuned it, denoted by names ending with FIXED  $\beta_1$  or TUNED  $\beta_1$ , respectively. Additionally, for these four training algorithm families, we also constructed baselines that use a list of 20 specific hyperparameter configurations to sample from, without replacement. We denote baselines making use of these kinds of search spaces by names ending with OPTLIST (e.g. ADAMW OPTLIST).

## 7.1 Baseline Creation Procedure

We used the results of the target-setting experiments to guide the creation of search spaces for baseline submissions. Since LAMB, ADAFACTOR, SAM(w. ADAM) and DISTRIBUTED SHAMPOO were not used in the target-setting procedure, we collected similar tuning data for them by sampling 200 trials from a broad search space for each algorithm. We used search spaces with the same ranges as in Table 8, except defined over the analogous hyperparameters in LAMB, ADAFACTOR, SAM(w. ADAM) and DISTRIBUTED SHAMPOO (i.e. the ADAMW  $1 - \beta_1$  range was used as the  $1 - \beta_1$  range for the analogous first moment parameters in the other algorithms). SAM(w. ADAM) has an additional  $\rho$  hyperparameter that we search over using the recommended discrete search space of  $\{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$ . Since the target setting procedure violates the submission rules by using too many tuning trials, we needed to tighten the search spaces used in the 200-trial searches to produce strong baselines for the 20 trial budget allowed in the external tuning ruleset. We used the following recipe to produce narrower search spaces for all baseline submissions (see Table 12 for the resulting exact search spaces).

1. **Learning rate and weight decay:** For a given algorithm, for every fixed workload, we first found the hyperparameter setting with the best validation error among the 200 trials, yielding a set of at most 8 hyperparameter points (up to one per fixed workload). We set the boundaries for the search spaces for the (base) learning rate and weight decay to include the maximum and minimum values observed in this set, plus a little extra on each side. While this procedure led to a small search space for preconditioned algorithms such as ADAMW and NADAMW, the search spaces for HEAVY BALL and NESTEROV were prima facie too wide to allow for efficient sampling, so we narrowed them by excluding outliers when building the set of per-workload hyperparameter points. In particular, we removed hyperparameter points from consideration when we could find an alternative hyperparameter point in the target-setting experiments with very similar performance, but within our narrowed-down search space.
2.  **$\beta_1$  and  $\beta_2$ :** As mentioned above, some baseline submissions searched over  $\beta_1$  (e.g. ADAMW TUNED  $\beta_1$ , HEAVY BALL TUNED  $\beta_1$ , etc.) and others fixed it to a default value (e.g. ADAMW FIXED  $\beta_1$ ).<sup>15</sup> When tuning  $\beta_1$ , we followed the same procedure as described for the learning rate to find the extreme values among the best per-workload hyperparameter settings, and set the search space to be within these

---

15. For the purposes of this paper, we call the momentum parameter  $\beta_1$  in HEAVY BALL and NESTEROV although it obviously is a different hyperparameter in each of those algorithms (as well as in ADAMW).

(rounded) bounds. As is common, we reparameterized the search space to explore  $1 - \beta_1$  in logspace. For  $\beta_2$  (where applicable) we used its default value of 0.999.

3. **Learning rate schedule parameters:** We used a fixed 5% linear warmup from 0.0 for all runs.<sup>16</sup> Baselines using the Linear Decay + Constant schedule family always searched the decay factor in the set  $\{0.01, 0.001\}$  (as in target-setting). In target-setting, both of these values appeared in top-performing hyperparameter settings in roughly equal proportion. For the decay steps factor parameter, we used a fixed value of 0.9; we saw the best trials in the 200-trial searches concentrated near that value and preliminary experiments seemed to suggest tuning it did not matter very much.
4. **Regularization Parameters:** For dropout, we noticed that, in the best trials, dropout and aux. dropout parameters tended to be the same (either both 0.0 or both 0.1). Therefore, we tied these parameters together in our baselines and searched them in the discrete set  $\{0.0, 0.1\}$ . For label smoothing, we noticed that whenever the workload allows it, the best-performing trials used either 0.1 or 0.2, therefore we searched within that set whenever the workload supported label smoothing.
5. **OptList baselines:** To build OPTLIST baselines, we ranked the 200 hyperparameter configurations from the broad searches independently on each of the 8 fixed workloads. Then we greedily grew a set of 20 hyperparameter configurations by cycling through workloads in an arbitrary, round-robin order and adding the top configurations on each workload that wasn't already in the set.

## 7.2 Baseline Timing

In order to run more extensive experiments, we used Google TPUs (Jouppi et al., 2020) whenever we could, thus deviating from the official,  $8 \times$  NVIDIA V100 GPUs benchmark system. To estimate training time for our baselines on the official benchmark system, we timed each training algorithm on the true competition system, on each workload, for a reduced number of training steps. Our timing measurements have some noise, possibly exacerbated by non-local disks. However, variations smaller than the time between off-the-clock evaluations (typically about 1% of total allowed runtime) are unlikely to have a large effect on the benchmark results. To reduce the time needed on the official benchmark system, we ran each algorithm (except DISTRIBUTED SHAMPOO) for 20% of the allowed number of steps, extrapolated to the full number of steps, and averaged across two runs on two different machines (see Table 28 in the Appendix for the timing results). As of the time of writing, we were not able to run DISTRIBUTED SHAMPOO on all workloads on the official benchmark system due to various memory and configuration issues that we hope to correct in future versions of this work. Therefore, we omit DISTRIBUTED SHAMPOO when presenting results with respect to runtime. Since we *were* able to run DISTRIBUTED SHAMPOO on TPUs, we have included DISTRIBUTED SHAMPOO in any results that don't require runtime measurements on the official benchmark system.

All the baseline algorithms should take roughly the same time per step, except for SAM(w. ADAM) and DISTRIBUTED SHAMPOO. In practice, SAM(w. ADAM) and DIS-

---

16. Although using a learning rate of zero on the first step is somewhat perverse, it simplifies the logic and is common in many codebases.

Hyperparameter	AdamW	NadamW	Heavy Ball	Nesterov
Base LR	Log [1e-4,1e-2]	Log [1e-4,1e-2]	Log [1e-1,10]	Log [1e-1,10]
Weight decay	Log [5e-3,1]	Log [5e-3,1]	Log [1e-7,1e-5]	Log [1e-7,1e-5]
1 - $\beta_1$ [Default]	0.1	0.1	0.1	0.1
1 - $\beta_1$ [Tuned]	Log [2e-2,0.5]	Log [4e-3,0.1]	Log [5e-3,0.3]	Log [5e-3,0.3]
1 - $\beta_2$	0.999	0.999	NA	NA
Schedule	warmup + cosine decay	warmup + cosine decay	warmup + linear decay	warmup + linear decay
Warmup	5%	5%	5%	5%
Decay factor	NA	NA	{1e-2,1e-3}	{1e-2,1e-3}
Decay steps	NA	NA	0.9	0.9
Label smoothing	{0.1, 0.2}	{0.1, 0.2}	{0.1, 0.2}	{0.1, 0.2}
Dropout (Tied)	{0.0, 0.1}	{0.0, 0.1}	{0.0, 0.1}	{0.0, 0.1}

Hyperparameter	LAMB	Adafactor	SAM(w. Adam)	Distributed Shampoo
Base LR	Log [1e-4,1e-2]	Log [1e-4,1e-2]	Log [1e-4,1e-2]	Log [1e-4,1e-2]
Weight decay	Log [1e-3,1]	Log [1e-3,1]	Log [1e-2,0.2]	Log [5e-3,1]
1 - $\beta_1$ [Default]	0.1	0.1	0.1	0.1
1 - $\beta_1$ [Tuned]	Log [2e-2,0.5]	Log [1e-2,0.45]	Log [5e-2,0.43]	Log [1e-2,0.15]
1 - $\beta_2$	0.999	0.999	0.999	0.999
$\rho$	NA	NA	{0.01, 0.02, 0.05}	NA
Schedule	warmup + cosine decay	warmup + cosine decay	warmup + cosine decay	warmup + cosine decay
Warmup	5%	5%	5%	5%
Decay factor	NA	NA	NA	NA
Decay steps	NA	NA	NA	NA
Label smoothing	{0.1, 0.2}	{0.1, 0.2}	{0.1, 0.2}	{0.1, 0.2}
Dropout (Tied)	{0.0, 0.1}	{0.0, 0.1}	{0.0, 0.1}	{0.0, 0.1}

Table 12: **Hyperparameter search space for the baseline submissions.** Descriptions of the learning rate schedules can be found in [Appendix A.1](#). The regularization hyperparameters are tuned only for those workloads where they are applicable. Dropout and aux. dropout are set to the same value.

TRIBUTED SHAMPOO typically require  $1.5\times-2\times$  the time per step (for DISTRIBUTED SHAMPOO this measurement is on TPU), depending on the workload. Note however that DISTRIBUTED SHAMPOO can amortize these costs by increasing the batch size which we leave for future work as it requires careful work for fair comparisons. Although all of our main, target-setting training algorithms had timing results in line with our expectations, specifically on the competition hardware using GPUs, our implementation of ADAFACTOR took around 10-20% more time per step than ADAMW on multiple workloads. In contrast, on TPU, ADAFACTOR step times were much closer to ADAMW, although they still seemed slightly slower. Due to these unexpectedly slower step times and the max runtime limit, our ADAFACTOR baseline misses a couple of targets that it would be able to hit if it ran for the same number of steps as ADAMW. Although ADAFACTOR was not one of the target-setting algorithms and we did not find any obvious issues in the implementation, we would like to investigate this slowdown more in the future.

On the competition hardware, using our current implementations, three out of the eight workloads, namely CRITEO 1TB DLRMSMALL, FASTMRI U-NET and OGBG GNN, are data-pipeline-bound when training with any of the target-setting algorithms. In other words, reading, preprocessing, and preparing batches of training data takes much longer than computing gradients and weight updates. The exact bottleneck varies across workloads, and isn't necessarily transferring data from disk. For example, on OGBG GNN creating and padding batches of graphs is a bottleneck. On data-pipeline-bound workloads, techniques such as data echoing (Choi et al., 2019a) can accelerate training. Additionally, on such workloads, optimizers that perform more work per batch and might otherwise increase the time-per-step (e.g. SAM(w. ADAM)) can reclaim idle accelerator time. Although data-pipeline-bound workloads exist in the wild and are a legitimate part of a representative benchmark suite,<sup>17</sup> as researchers iterate on a specific workload, they tend to make changes that reduce, or eliminate, idle accelerator time. For example, they might optimize the input pipeline code, run multiple independent copies of the data pipeline in parallel, or switch to a larger, more computationally intensive model. For these reasons, input pipeline bottlenecks are best viewed not as an immutable property of the problem, but instead as a consequence of a particular model size, amount of engineering resources, and amount of computational resources. Although a small number are fine, we generally view data-pipeline-bound workloads as undesirable for a training algorithms benchmark because such workloads tend to favor expensive training algorithms that wouldn't be as useful when we have the resources to remove the bottleneck. In the future, we would like to optimize the data pipelines for CRITEO 1TB DLRMSMALL, FASTMRI U-NET and OGBG GNN in hopes of removing bottlenecks and increasing GPU utilization.

### 7.3 Baseline Results

Figure 5 shows performance profiles for ADAMW, NADAMW, NESTEROV, HEAVY BALL, LAMB, ADAFACTOR, SAM(w. ADAM) and DISTRIBUTED SHAMPOO using the search spaces described in Table 12 with the  $\beta_1$  parameter tuned. Figure 5a shows the perfor-

---

17. Given the end of Moore's law, we should expect accelerator improvements to continue to outpace improvements in general purpose processors. For that reason, the problem of bottlenecks upstream of the part of the training pipeline that runs on the accelerator(s) is not likely to go away.

mance profile with respect to runtime, [Figure 5b](#) shows the same performance profile with respect to steps, and [Table 13](#) shows the benchmark scores measuring the normalized area under the performance profile curves. Unlike the official scoring procedure we will use for real submissions, baselines in [Figure 5](#) ignore the held-out workload criterion described in [Section 4.5.3](#), which will depend on the specific held-out workloads sampled during official scoring, after submission code has been frozen. Since the performance profile (and thus the benchmark score) for any individual baseline depends upon the entire set of baselines included in the performance profile, we computed performance profiles and benchmark scores for the complete set of baselines presented across all sections, together (e.g. including those presented in [Section 7.3.1](#)) to ensure consistency. Throughout the paper, when comparing a subset of baselines, we might elide some baselines from a figure to improve readability, but the curves are always computed based on the full set. Please see [Table 27](#) in the appendix for the complete set of benchmark scores for all baselines considered in the paper.

Although our benchmark is designed around measuring time-to-result, looking at performance profiles based on steps as well can sometimes be illuminating. Given that all target-setting algorithms take roughly the same time per step, the relative ranking of the baselines derived from those algorithms does not change when measuring steps instead of runtime. Since SAM(w. ADAM) and DISTRIBUTED SHAMPOO typically require  $1.5 \times - 2 \times$  the time per step as the target-setting algorithms, in order to create performance profiles based on steps as well as runtime, we ran SAM(w. ADAM) and DISTRIBUTED SHAMPOO for the same number of steps as the other baselines and used learning rate schedules based on this step budget. These step-budget-optimized schedules likely decay the learning rate too slowly to be ideal when measured with respect to runtime. With faster implementations and runtime-tuned learning rate schedules, it might be possible to make these training algorithms competitive in runtime as well as in number of steps. That said, SAM(w. ADAM) has a very large deficit to make up, and might still struggle to be competitive. Regardless, we hope proponents of these algorithms create submissions that achieve the best possible runtime results on our benchmark.

We can observe several interesting results from [Figure 5](#), [Table 13](#) and the workload-specific breakdown in [Table 25](#) and [Table 26](#) presented in the appendix.

- There is no single baseline training algorithm that hits the targets on every workload, but on every workload there exists at least one (and usually more than one) baseline that hits the target. DISTRIBUTED SHAMPOO and NADAMW both reach the target on 7 out of 8 workloads. However, as described above, DISTRIBUTED SHAMPOO has an unfair budget advantage, so NADAMW hits the targets on the largest number of workloads among baselines adhering to the strict runtime budget. DISTRIBUTED SHAMPOO misses the target on CRITEO 1TB DLRMSMALL, whereas NADAM misses the target on IMAGENET RESNET-50.
- With our search spaces, NADAMW performs significantly better than ADAMW both in terms of runtime and steps. Despite this impressive result, it seems to be much less popular than ADAMW.
- We found that ADAFACTOR performs somewhat worse than ADAMW in terms of steps to target. ADAFACTOR is able to hit targets on  $\frac{4}{8}$  workload. ADAMW hits the same targets and additionally the target on CRITEO 1TB DLRMSMALL.

- LAMB hits the target on only 2 of our workloads, LIBRISPEECH DEEPSPEECH and WMT TRANSFORMER and thus does not do well in terms of the benchmark score. However, for WMT TRANSFORMER in particular, LAMB is the fastest algorithm to the target (by a 7% margin).
- The non-preconditioned baselines, HEAVY BALL and NESTEROV, are not able to hit the target on any workload. Two factors explain this surprisingly poor result: (1) how we constructed search spaces for these training algorithms, and (2) the way our benchmark rules require consistent performance and force training algorithms to account for the workload-specific tuning they require to achieve good results. Both HEAVY BALL and NESTEROV have some trials which hit the target on IMAGENET RESNET-50 and FASTMRI, but they are not frequent enough to show up in the median over the five studies. Indeed, search spaces for these algorithms that are better tailored to the tuning budget can hit more targets, which we show later in [Section 7.3.1](#). However, as the target-setting experiments suggest, outperforming the pre-conditioned baselines would be a tall order.

Submission	Benchmark Score	
	Runtime	Steps
ADAMW	0.600141	0.596116
NADAMW	<b>0.849960</b>	0.830414
NESTEROV	0.0	0.0
HEAVY BALL	0.0	0.0
LAMB	0.248619	0.248494
ADAFCTOR	0.236111	0.475760
SAM(w. ADAM)	0.120368	0.731717
DISTRIBUTED SHAMPOO	-	<b>0.854210</b>

Table 13: **The benchmark scores for our baseline submissions.** These are the integrated performance profiles shown in [Figure 5a](#) (runtime) and [Figure 5b](#) (steps).

### 7.3.1 BASELINE RESULTS COMPARING SEARCH SPACES

In order to reveal more about the role of the hyperparameter search space, we compared baselines using the target-setting algorithms (ADAMW, NADAMW, NESTEROV, HEAVY BALL) with different search spaces. For each algorithm, we compared two search spaces (specified in [Table 12](#)): one tuning the  $\beta_1$  parameter, and the other keeping the  $\beta_1$  parameter fixed to 0.9. Additionally, for each algorithm, we also prepared an OPTLIST baseline, as described above, that samples without replacement from a list of hyperparameter configurations that performed well on at least one workload during target-setting. Tables [29](#), [30](#), [31](#) and [32](#) in the Appendix contain the complete list of configurations used in the OPTLIST baselines. [Figure 6](#) shows the resulting performance profiles for this comparison and [Table 14](#) shows the corresponding benchmark scores given by the (normalized) area

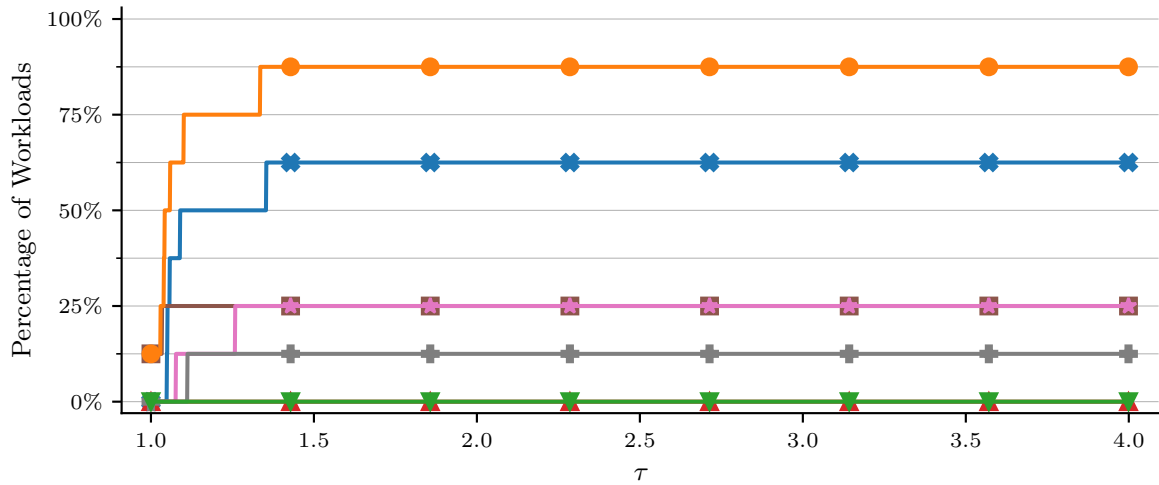
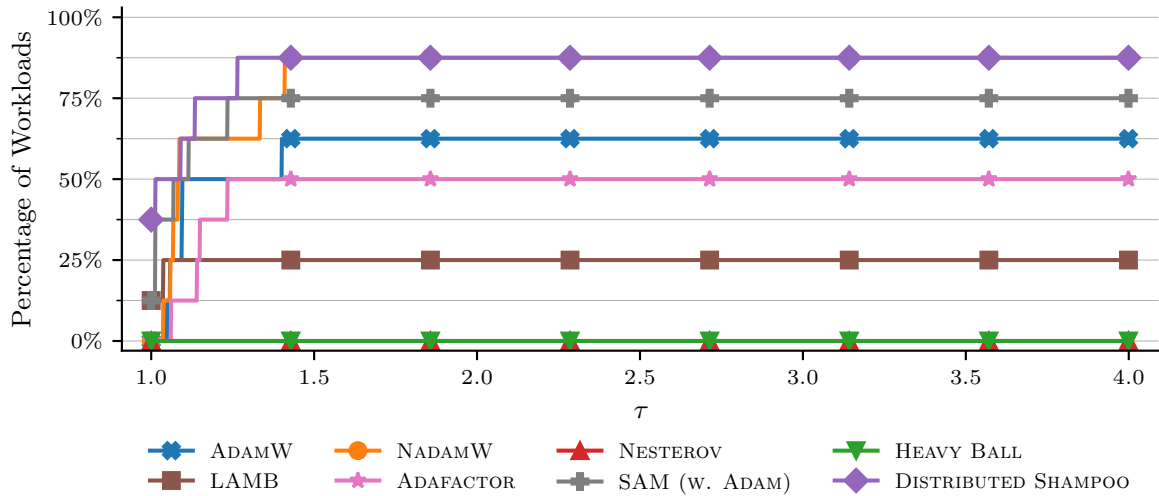
(a) Performance profiles when measuring **runtime** to target(b) Performance profiles when measuring **steps** to target

Figure 5: **Performance profiles of our baseline submissions.** Each line in these plots is the performance profile of a single baseline submission. A step in any line occurring at a value  $\tau$  indicates that for one additional workload the corresponding submission achieves the target within a  $\tau$  factor of the runtime of the best submission. For example in (b), NADAMW (—) has a bump just before  $\tau = 1.5$ . This indicates that on one additional workload, NADAMW requires a bit less than  $1.5\times$  the number of steps as the fastest submission to reach the target on this workload. If a submission does not reach the target on one or more workloads, its performance profile will not reach 100% at the very right of the plot. A flat line at 0% indicates that for all workloads the submission either did not hit the target at all or did so in time/steps that is at least 4 times worse than the time/steps to target of the best submission. As mentioned earlier, we computed the performance profile for the entire set of baselines presented in the paper together and have removed the profiles for other baselines from this figure for readability.



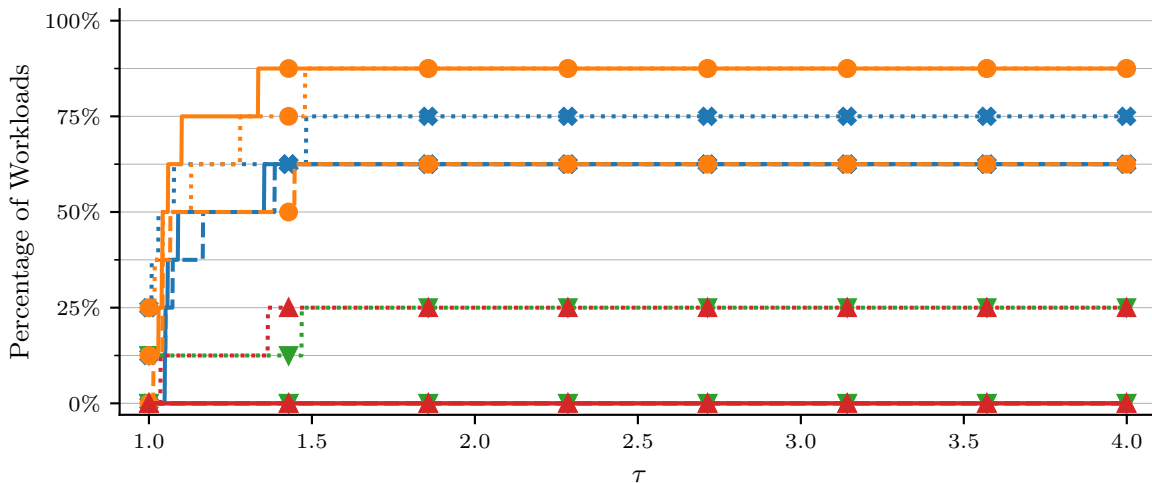
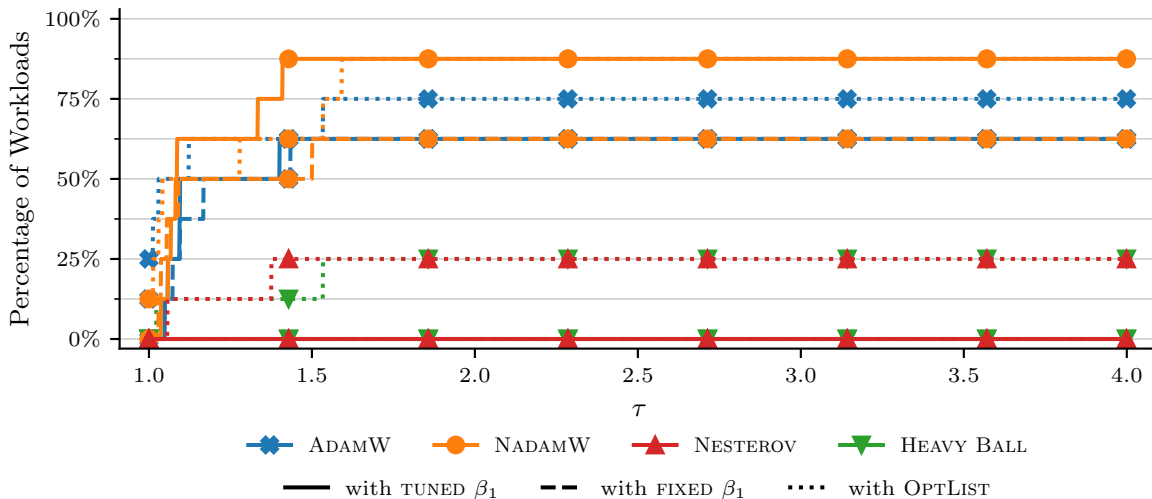
(a) Performance profiles when measuring **runtime** to target(b) Performance profiles when measuring **steps** to target

Figure 6: **Performance profiles of our baseline submissions with different search spaces.** Each line in these plots is the performance profile of a single baseline submission (analogously to Figure 5). As mentioned earlier we computed the performance profile for the entire set of baselines presented in the paper together and have suppressed the profiles for other baselines from this figure for readability.

under the curve. The raw data used to prepare these plots can be found in Tables 25 and 26 in the Appendix.

As expected, the search space plays a large role in the results.

- Although overall NADAMW with a search space that tunes  $\beta_1$  performed the best among these training algorithms in terms of runtime (and steps), NADAMW performed roughly the same as ADAMW when the  $\beta_1$  parameter isn't tuned (fixed to the default

value of 0.9). On the other hand, for ADAMW, tuning  $\beta_1$  shows little advantage (at least at the limited number of 20 tuning trials).

- The OPTLIST baselines for NADAMW and ADAMW seem to be roughly the same in terms of performance as compared to the best box polytope search spaces for those training algorithms. On the other hand, for NESTEROV and HEAVY BALL, the OPTLIST baselines perform substantially better (and are the only ones to hit any target). This result is likely a consequence of our search space construction procedure and the diversity of hyperparameter values that worked well for NESTEROV and HEAVY BALL. In other words, our procedure for constructing box polytope search spaces needed to cover much wider hyperparameter ranges for NESTEROV and HEAVY BALL that also contain a large volume of bad points, making it harder to get good results when sampling from these types of search spaces, for these algorithms.

Submission	Version	Benchmark Score	
		Runtime	Steps
ADAMW	TUNED $\beta_1$	0.600141	0.596116
	FIXED $\beta_1$	0.596985	0.593047
	OPT-LIST	0.725260	0.721035
NADAMW	TUNED $\beta_1$	<b>0.849960</b>	<b>0.830414</b>
	FIXED $\beta_1$	0.599691	0.595478
	OPT-LIST	0.835602	0.813194
NESTEROV	TUNED $\beta_1$	0.0	0.0
	FIXED $\beta_1$	0.0	0.0
	OPT-LIST	0.233373	0.232048
HEAVY BALL	TUNED $\beta_1$	0.0	0.0
	FIXED $\beta_1$	0.0	0.0
	OPT-LIST	0.230504	0.226860

Table 14: **The benchmark scores for our baseline submissions.** These are the integrated performance profiles shown in [Figure 6a](#) (runtime) and [Figure 6b](#) (steps).

## 8. Discussion

### 8.1 Target Setting

Fundamentally, our target setting procedure is a *generic* recipe for training neural networks that achieves (arguably) a near state-of-the-art result on all of our fixed workloads and could, in principle, be applied to any workload. The procedure assumes the model architecture, dataset, loss function, and other workload details are given and also assumes we have a known limit for the maximum runtime allowed for the training program. Furthermore, it assumes we can afford to sample hundreds of trials from the search spaces we designed for the various training algorithms. Although it depends on a lot of assumptions, it *does* give us an automatic process for getting good results on our fixed workloads.

Naturally, we might ask ourselves how good this procedure is compared to other alternatives. Unfortunately, our procedure is the only one of its kind we are aware of in the deep learning literature. Although [Godbole et al. \(2023\)](#) make an attempt to systematize the applied deep learning workflow, the process they describe is still far from a repeatable, mechanical procedure we can use for setting targets. Instead of generic procedures for training and tuning neural networks, the literature is filled with particular configurations that worked well on specific problems, only occasionally accompanied by the tuning search spaces that helped discover them. These workload-specific training recipes typically do not generalize to new workloads. This gap in the literature makes it difficult to know how competitive the validation and test targets we set with our procedure actually are, or how realistic they are to reach for completely general, workload-agnostic training algorithms.

Two steps in the deep learning workflow seem to be especially neglected by the literature. First, we are not aware of algorithms for determining how long to train, or papers that tackle the concomitant philosophical issues surrounding how best to frame this choice. Second, existing tools for automatic hyperparameter tuning still require search spaces as input, i.e., we need to somehow determine which hyperparameters to tune and what set of values they should be allowed to take on. Although the blackbox optimization literature has explored methods that automatically grow or shrink search spaces, the deep learning literature contains shockingly little research on how to design the best search spaces for use with existing tools and arbitrary budgets. Indeed, this profound gap in the literature was part of our motivation for forcing submissions in our own benchmark to grapple with tuning challenges directly.

Our own target-setting procedure is far from the final word on these issues. In an effort to keep the protocol workload-agnostic and generic, we ended up with broad search spaces that required a large number of tuning trials, even though preliminary experiments showed that some hyperparameters only needed to be tuned on a subset of the workloads. For example, supporting and tuning dropout was not always necessary or helpful. Furthermore, our search spaces ([Table 8](#)), despite being relatively broad, had to be constructed through manual trial and error. In several cases, we had to refer to previous published results on similar workloads to determine that our initial search spaces needed to be adjusted. We hope that the community scrutinizes our target setting procedure and comes up with better alternatives. Although we tried our best to create an objective protocol that minimized human judgement, we were unable to remove it entirely. Even if full automation is out of reach, we hope the community takes the problem of hyperparameter search space construction, specifically in deep learning, much more seriously.

Stepping back, our target setting procedure essentially simulates a competition between training algorithms to get the best validation performance within a fixed runtime budget. This protocol is a mirror image of our time-to-result benchmark. One possibility might be to formalize this relationship and alternate competitions to improve targets given fixed runtime budgets, and runtimes given fixed targets. Even if we don't take things quite so far, we could revise targets in future benchmark iterations by incorporating winning submissions into the target-setting protocol. The targets we selected using our procedure represent what is possible with currently popular training algorithms applied in a workload-agnostic manner. These targets are generally *not* going to be the state-of-the-art error rates on the tasks our workloads cover.

## 8.2 Randomized Workloads

As described in [Section 6](#), although we found realistic workload variants that simultaneously changed the best-performing NADAMW hyperparameters in our search spaces and reached reasonable error rates in the original runtime budget, discovering such variants was surprisingly difficult and far too labor intensive. Our difficulties highlight the desperate need for new research to predict when hyperparameter settings that perform well on one workload will transfer to a related workload and, more generally, predict the effect of various workload modifications on the optimal hyperparameters. While there has been some recent work in this vein, such as [Yang et al. \(2021\)](#) which focuses on model size, we are still far from being able to reliably predict when good hyperparameters will transfer. A theory of hyperparameter transfer for all of the most common hyperparameters (and most popular training algorithms) would not just make it easier to produce interesting workload variants, but would likely go a long way towards achieving the underlying purpose of our randomized workloads: encouraging training algorithms that are robust to the exact details of the workload and are easy to tune. Even if such a theory did not immediately lead to more robust and convenient algorithms, it would let us save tuning effort by extrapolating to new experiments from related tuning results.

Our randomized workloads built from workload variants highlight an important set of trade-offs in the design of our benchmark. The more workloads in the benchmark, the more expensive running a submission through the benchmark becomes, especially in the more expensive external tuning ruleset. Both fixed workloads and randomized workloads contribute to this cost, but because we sample a single variant from each randomized workload to construct the held-out workloads, the fixed and randomized workloads currently contribute roughly equally to the total cost. Similarly, the more tuning trials we allow in the external tuning ruleset, the more expensive the benchmark becomes. Since larger tuning budgets let submissions adapt more to specific workloads, they also increase the importance of using a large enough and diverse enough set of workloads, potentially through randomized workloads. However, without a better understanding of when existing training algorithms need to be re-tuned, it is very hard to know when a new workload or workload variant is worth the increase in cost. Even worse, a variant that provides a useful challenge for one training algorithm might be completely redundant for another.

## 8.3 Baselines

Our baseline results show that our per-workload targets are achievable, reveal clear gaps between different training algorithms, and suggest performance on the benchmark is far from saturated. On each workload, at least one baseline was able to reach the target, but no baseline reached the targets on all workloads simultaneously. NADAMW with our tuned  $\beta_1$  search space reached the target on seven out of the eight fixed workloads, and constitutes a provisional state of the art on our benchmark. The performance profiles that form the basis of the benchmark scores showed stark differences between the tested methods.

Our baseline results also showed that the current practice of trying to compare abstract update rules with free parameters, divorced from the tuning protocol that would instantiate them, is doomed to produce perpetually conflicting results. Hyperparameter tuning search spaces (and tuning protocols more generally) play a crucial role in the effectiveness of

a training algorithm. By picking specific search spaces, even when they are reasonable *a priori*, we could claim our results showed that ADAMW is better than NADAMW, or vice versa, when in reality we are only showing that NADAMW *with a particular tuning protocol* is better than ADAMW with some other specific tuning protocol. We need to view the tuning protocol as an inseparable part of the training algorithm if we are ever going to determine a meaningful notion of the state-of-the-art training algorithm. Similarly, although HEAVY BALL or NESTEROV was the target-setting algorithm on a combined three out of the eight workloads, our HEAVY BALL or NESTEROV with box polytope search spaces failed to reach the target reliably on any workloads. This is not a general judgment about HEAVY BALL or NESTEROV as update rules. Instead, this result once again highlights the importance of the tuning protocol and, if anything, might indicate that these algorithms require more workload-specific tuning effort. New training algorithms with hyperparameters that must be tuned should ideally provide (budget-dependent) tuning procedures, or at a minimum, guidance on how to tune at a variety of budgets.

#### 8.4 Benchmark Limitations

The benchmark we presented in this work, like any benchmark, has a variety of limitations. These limitations fall into several, broad categories. First, the benchmark has limited coverage of possible submissions. In other words, the benchmark rules and software end up prohibiting, or effectively prohibiting, some potentially interesting submissions that we would have preferred to allow, in principle. Second, there are limitations that could affect whether benchmark scores truly measure what we intend them to measure (i.e. benchmark “validity” in the parlance of psychometrics). Third, there are limitations of the scope of the benchmark that could affect its relevance to the actual practice of deep learning. Finally, there are limitations that affect the accessibility of the benchmark, primarily in terms of how easy and affordable it is for researchers to score new submissions.

**Coverage of the space of potentially interesting submissions** Benchmark submissions must adhere to a specific training algorithm API and interoperate with workload implementations in either JAX or PYTORCH. Although this restriction is an essential design choice intended to, among other things, help isolate the effects of the training algorithm, it does make it so some potentially interesting submissions cannot be supported under the current rules. For example, although submissions can employ arbitrary hyperparameter tuning procedures while being timed, submissions adhering to the external tuning ruleset and making use of parallel tuning resources must use random search. Instead of being allowed to employ more sophisticated black-box optimization algorithms, they must only rely on their control over the tuning search space. Although we could potentially relax this limitation in future versions, allowing submissions complete control over how they utilize the tuning resources would require much more complicated orchestration code and APIs.

Similarly, although not against the rules, the API does not provide a way to implement model parallelism since it isn’t situated in the submission code, as we define it, even if we could imagine a hypothetical exotic training algorithm that depended on it somehow. As another example, submissions are not allowed to access arbitrary information about the current workload. Instead, they can only obtain basic layer metadata and dataset information. This restriction precludes submissions based on optimizers, such as K-FAC (Martens

and Grosse, 2015), that require detailed architectural information. Although K-FAC itself is non-trivial to apply to new model architectures even when detailed architectural information is available, hypothetical generic training algorithms that required such information, but worked for any arbitrary neural network, would likely also face difficulties.

Submissions are also constrained from a software implementation standpoint. They must interoperate with either JAX or PYTORCH workload implementations, prohibiting other frameworks and software stacks. Although in theory there is nothing stopping us from porting our workloads to additional frameworks, in practice it is far too costly in terms of engineering resources.

Ultimately, our initial rules and API err on the side of caution to make sure we isolate the effects of the training algorithm. Moving forward, we intend to keep a close eye on the types of algorithmic modifications that are of interest to the community, but are not possible within our rules or API, and solicit suggestions on ways they could be accommodated.

**Experimental protocol** Although we designed our benchmark to prioritize producing convincing measurements, there are still ways the experimental protocol could be strengthened. Specifically in the external tuning ruleset, the ratio between the number of workloads (eight fixed along with eight more held-out workloads) and the number of tuning trials allowed per study (twenty) could allow for too much workload-specific tuning and thus risk rewarding submissions that overfit to the particular suite of workloads. Although the randomized workloads we draw held-out workloads from are designed to mitigate this issue, the workload variants did not fully achieve all of our desiderata. Although all three variants of a given base workload will require different hyperparameter settings than the base workload, they might not require mutually distinct hyperparameter settings from each other, risking a set of variants that doesn’t challenge submissions as much as initial appearances might suggest. Additionally, the variants of different *base* workloads often repeat similar changes. For example, we generated variants of multiple base workloads by making activation function changes. Moving beyond workload overfitting concerns, neural network training is a noisy process, especially when hyperparameter tuning using random search is involved. Our strategy of repeating measurements with different random seeds during scoring can help, but the best-performing optimization hyperparameters for the most popular optimizers are often near the “edge of stability” (Cohen et al., 2022), so some submissions can get unlucky and have training diverge more than is typical when we only have a small number of repetitions.

**Scope and relationship to current practice** Our choices of what workloads and conditions to study in our benchmark necessarily constrain the set of situations the results will be most relevant for. Even restricting our attention to supervised and self-supervised learning, we cannot hope to cover every practically relevant data modality, let alone every practically relevant dataset and model. For example, our benchmark currently does not contain any workloads for object detection from point cloud data, weather prediction, language modeling, video understanding, or image generation. Our choice to prioritize currently popular, easily-accessible, and well-studied datasets and models could end up reinforcing existing selection effects. Perhaps because currently popular training algorithm co-evolved with the most popular application domains, they work unusually well together. In this case, a benchmark emphasizing existing popular workloads will make it hard to break out of what could



be a methodological local optimum. The advent of new training algorithms could potentially unlock efficient training for entirely different models, ones that are not covered by our benchmark. In the near-term, it is impossible to resolve these types of counterfactuals, but we could potentially create a more diverse—along every dimension we can measure—set of benchmark workloads, if we saved resources in other ways. Additionally, as discussed in [Section 7.2](#), we would like to have slightly fewer data-pipeline-bound workloads. Moving away from the limitations of our particular set of current workloads, the benchmark today includes only a single hardware weight class. Although we believe results at this scale are relevant at a variety of interesting scales, the gold standard would be to directly measure much larger and smaller scales. That said, even if the benchmark results are informative for many different scales, the relative performance of different training algorithms might systematically vary with batch size. By only using a single specific system with a particular amount of accelerator memory, the benchmark effectively covers only the narrow range of batch sizes for each workload that are close to the largest batch size that can fit in memory for the most competitive submissions.

**Accessibility** In order for our benchmark to be useful, scoring the most intriguing new training algorithms developed by the research community needs to be feasible. A large part of whether it is feasible for a particular group to evaluate a submission on the benchmark is the compute costs of the scoring protocol, running on the official benchmark system. [Table 15](#) lists the provisional<sup>18</sup> number of machine-hours required, on the official benchmark system using 8×NVIDIA V100 GPUs, to evaluate submissions under both tuning rulesets. Although that tuning can run on a different machine and need not use the official benchmark system, it is still worth computing the number of machine-hours on the benchmark system as a reference point. There are a variety of ways these costs could come down. For example, we could switch to a system with newer GPUs that speeds up training enough to be worth any concomitant increase in the price per machine-hour. Or, as submissions become more competitive, we could shrink the maximum runtime allowed.

Nevertheless, as it stands, for groups that don’t have several 8×NVIDIA V100 GPUs machines on premises, cloud costs to score submissions could be a significant hurdle. On the other hand, groups that train very large language models routinely run experiments eclipsing these scoring costs by orders of magnitude. Ultimately, whether or not evaluating a submission is affordable is a relative question that will depend on the specific research group. For now, our solution is to obtain compute sponsorship to help groups with more limited resources evaluate and score promising submissions.

In order to allocate limited sponsorship funding among submissions from groups unable to self-fund scoring costs, we plan to use performance on a qualification set of workloads that excludes some of the most expensive workloads. Specifically, the qualification set consists of the CRITEO 1TB DLRM SMALL, OGBG GNN, and WMT TRANSFORMER workloads, without any held-out workloads. Submitters that do not have the resources to self-report results on the full set of workloads may instead report results on this smaller qualification

---

18. We plan to reduce the running time of some of the more expensive workloads before issuing a call for submissions, so these costs are likely a bit of an overestimate, although they should be the correct order of magnitude.



set. As shown in Table 15, evaluating submissions on the qualification set is about an order of magnitude cheaper than the full set of workloads.

Stepping back, part of the issue of affordability is that our benchmark is all-or-nothing. To compute a benchmark score, we need to run submissions on all workloads. Thus, for any specific level of compute resources necessary for scoring, some groups will struggle to participate without compute support. It is hard to imagine this approach scaling to thousands of realistic workloads, especially if we consider tuning costs, without excluding far too many groups. Ideally, we would have a benchmark with a flexible mechanism for incrementally investing computational resources to gather more and more information about the performance of a submission, eventually culminating in a complete set of experiments on a large, highly-diverse set of fixed and held-out workloads. Even with our current design, we can compute an upper bound on a submission’s benchmark score based on training on a subset of workloads and/or with a reduced runtime limit, but our scoring procedure would probably need to be revised if we wanted it to scale to a much larger number of still-relatively-costly workloads.

Setting	Time (h)
<b>External Tuning Ruleset</b>	
One hyperparameter	232.23
Scoring a submission	1161.13
Tuning a submission	23,222.61
<i>Qualification Set</i>	
One hyperparameter	20.65
Scoring a submission	103.24
Tuning a submission	2064.75
<b>Self-tuning Ruleset</b>	
One hyperparameter	696.68
Scoring a submission	3483.39
<i>Qualification Set</i>	
One hyperparameter	61.94
Scoring a submission	309.71

Table 15: **Estimated required runtime of the benchmark.** *One hyperparameter* refers to running all eight fixed and eight held-out workloads once, i.e. with a single hyperparameter. *Scoring a submission* involves repeating this process for each study, i.e. five times. To fully *tune a submission*, each study uses twenty tuning trial to identify the best hyperparameter setting (note that this need not use the benchmark system). Running a single hyperparameter is more expensive in the self-tuning ruleset since it has a three times larger runtime budget for every workload to compensate for the lack of external tuning. The *qualification set* only consists of three (out of the eight) fixed workloads, without held-out workloads, and thus offers a reduce cost.

## 8.5 Future Work

In addition to work on improving the benchmark itself and building new, and even stronger, baselines (especially for the self-tuning ruleset), there are several related areas that would benefit from more research. One appealing area of future work motivated by our challenges

with building randomized workloads, as discussed previously in [Section 8.2](#), would be a theory to predict the effect of various workload modifications on the optimal hyperparameters, paving the way for a better understanding of how to perform workload-specific tuning, how to build randomized workloads with support over a combinatorial space of variants, and when workload-specific tuning is even necessary. Another direction with immense practical potential is developing completely workload-agnostic training recipes that, when given an arbitrary neural network training workload (model, dataset, and loss function) along with a budget, produce the best possible result. Even though we could use such recipes for target-setting, to be maximally interesting, they would need to go far beyond our target-setting procedure and be useful enough to be adopted by practitioners.

Although we restricted our attention to training algorithms in this work, there are other parts of the deep learning training pipeline that affect training speed. For the purposes of our training algorithms benchmark, the preprocessing, model architecture, and loss function are fixed components of the workload, but they could all benefit from algorithmic improvements. The initial proposal for the benchmark rules also included a separate, time-to-result model benchmark that measured training speedups due to model changes. In order to isolate the effects of model changes, this model benchmark required submissions to train using a small set of standard training algorithms. Unlike in the training algorithm benchmark, models would only be required to perform well on a single task, albeit across different datasets (e.g. different language pairs in machine translation). Consequently, task-specific components of the machine learning pipeline, such as data augmentation, could be incorporated as part of the submission in a hypothetical future `ALGOPERF MODELS` benchmark. With results from both training algorithms and separate model benchmarks, we could determine the relative responsibility for speedups on different tasks of the training algorithms vs the model architecture, and develop a complete picture of the most promising directions for accelerating neural network training.

## 9. Conclusion

Neural networks must be trained to be useful, making training algorithms essential for creating useful deep learning models. Unfortunately, due to the lack of a standard, convincing protocol for empirically comparing training algorithms, progress on better training algorithms has stalled. To address this pressing issue, in this work, we introduced the `ALGOPERF: TRAINING ALGORITHMS` benchmark, a competitive, time-to-result benchmark covering multiple realistic workloads, running on fixed hardware. This benchmark represents the collective efforts, over multiple years, of the members of the `MLCommons Algorithms` working group to remove the measurement barriers frustrating progress on neural network training algorithms.

Although we believe the benchmark we have created represents an important advance, it is far from perfect and has a variety of limitations. We extend an open invitation to the entire community—and in particular, those who disagree with our design choices—to join the working group and collaborate on improving the benchmark further, either before we issue the initial call for submissions, or after. We would also be delighted by any novel solutions for the challenges with training algorithm comparisons we described in [Section 2](#), whether

they are possible to incorporate into ALGOPERF: TRAINING ALGORITHMS or necessitate an additional benchmark with a fundamentally different approach.

We urge researchers developing new training algorithms to submit them to the ALGOPERF: TRAINING ALGORITHMS benchmark competition once the call for submissions has been issued. Crafting a valid submission might require thinking more about hyperparameter tuning than researchers inventing new optimizers are used to, but this work will help us break free of the cycle of hype and abandonment currently facing new algorithms. Outside the competition schedule, researchers can still use the benchmark to measure the performance of new algorithms as long as they adhere to the rules and report raw workload times, in addition to recomputing unofficial benchmark scores. Researchers should feel free to reach out to the working group for guidance, as needed, or to find potential collaborators with the resources to run larger, more comprehensive experiments. Finally, we believe that the presented benchmark constitutes a new state of the art for empirical comparisons of training algorithms, and should be viewed as a first step towards a more reproducible and empirically rigorous scientific literature on neural network training algorithms.

## Author Contributions and Acknowledgments

- **George E. Dahl:** Founded and chaired the working group. Co-authored the initial rules proposal and shaped the rules of the benchmark. Recruited contributors to complete the necessary engineering work. Co-led paper experiments and co-authored the codebase used for paper experiments. Co-led the paper’s writing process and supervised the writing contributions from other authors. Directly contributed to every aspect of the writing process including outlining, drafting sections, creating figures, and editing. Served as overall project coordinator.
- **Frank Schneider:** Chaired the working group. Significantly influenced the rules of the benchmark. Co-led the paper’s writing process and supervised the writing contributions from other authors. Directly contributed to every aspect of the writing process including outlining, drafting sections, creating figures, and editing.
- **Zachary Nado:** Lead engineer, tech lead, and supervisor for building the benchmark codebase and cloud infrastructure, including the implementation of the workloads in both JAX and PYTORCH. Coordinated verifying the correctness across infrastructure implementations. Co-authored the codebase for paper experiments. Co-authored the initial rules proposal. Made major writing contributions to the paper.
- **Naman Agarwal:** Co-led paper experiments. Designed and ran a large number of critical experiments including target-setting experiments, baselines, workload variants, and many others. Worked on initial implementations of the FASTMRI and VIT workloads. Made major writing contributions to the paper. Significantly influenced the rules of the benchmark.
- **Chandramouli Shama Sastry:** Implemented both LIBRISPEECH workloads in PYTORCH. Developed test suites to compare the implementations of workloads across frameworks that caught several implementation deficiencies. Contributed to the benchmark infrastructure.
- **Philipp Hennig:** Significantly influenced the rules of the benchmark and made major writing contributions to the paper.
- **Sourabh Medapati:** Assisted with the writing of the paper. Implemented both LIBRISPEECH workloads in JAX and helped debug them in PYTORCH. Conducted several experiments. Helped maintain the [paper experiment codebase](#). Contributed to the benchmark infrastructure.
- **Runa Eschenhagen:** Significantly contributed to the benchmark infrastructure, leading the PYTORCH development for a large portion of the development cycle. Implemented the WMT workload in PYTORCH and the OGBG workload in both PYTORCH and JAX. Debugged numerous critical issues in the [benchmark codebase](#) and helped refine APIs.
- **Priya Kasimbeg:** Led the timing experiments between our PYTORCH and JAX implementations and debugged several critical issues in the [benchmark codebase](#). Contributed to the benchmark infrastructure. Helped maintain the [paper experiment codebase](#).

- **Daniel Suo:** Implemented the initial version of the FASTMRI and VIT workloads in JAX. Created the performance profile and scoring infrastructure. Made a large number of essential contributions to the [main codebase for paper experiments](#). Contributed to the benchmark infrastructure. Assisted with the writing of the paper.
- **Juhan Bae:** Implemented the VIT, FASTMRI, and CRITEO 1TB workloads in PYTORCH. Debugged several critical issues in the [benchmark codebase](#). Contributed to the benchmark infrastructure and documentation.
- **Justin Gilmer:** Made major writing contributions to the paper. Conducted experiments highlighting the need for workload standardization and designed several held-out workload variants. Co-authored the initial rules proposal. Co-authored the [codebase for paper experiments](#).
- **Abel L. Peirson:** Explored ways to define randomized workload distributions for the OGBG base workload and implemented the CIFAR-10 development workload for JAX. Assisted with the writing.
- **Bilal Khan:** Conducted the baseline experiments for DISTRIBUTED SHAMPOO, SAM, ADAFACTOR, and LAMB. Assisted with the writing of the paper.
- **Rohan Anil:** Co-authored the initial rules proposal. Advised on the setup of the JAX baselines, and optimizer configuration details, generally helped with performance tuning and debugging JAX/PYTORCH differences, and assisted with writing the paper.
- **Mike Rabbat:** Significantly influenced the rules of the benchmark, assisted with the PYTORCH implementation, provided support for the FASTMRI workload, and made major writing contributions to the paper.
- **Shankar Krishnan:** Assisted with the writing of the paper. Conducted several experiments.
- **Daniel Snider:** Implemented the RESNET-50 workload in JAX. Conducted preliminary experiments for randomized workloads on OGBG, wrote code for logging, and explored options for serving the benchmark. Drafted rules around software dependencies.
- **Ehsan Amid:** Assisted with the writing of the paper. Supported the implementation of CRITEO 1TB for JAX.
- **Kongtao Chen:** Implemented the initial drafts of the DEEPSPEECH workload in PYTORCH and the WMT workload in JAX.
- **Chris J. Maddison:** Co-authored the initial rules proposal.
- **Rakshith Vasudev:** Supported the implementation of the CRITEO 1TB workload for PYTORCH.
- **Michal Badura:** Assisted with the writing of the paper. Wrote the initial JAX implementation of the OGBG workload in the [main paper experiment codebase](#).
- **Ankush Garg:** Supported the implementation of the WMT workload. Assisted with the writing of the paper.

- **Peter Mattson:** Co-authored the initial rules proposal.

The Brain team at Google Research supported the initial work on the benchmark and the JAX baselines.

The authors would like to express their gratitude to David Kanter and the entire MLCommons organization for their support throughout the project. Many thanks to Toby Boyd for his help in preparing a request for compute sponsorship and additional logistical support. We are thankful to Hanlin Tang for his help in designing the initial PYTORCH API requirements and implementing the RESNET-50 workload in PYTORCH. We thank Leda Sari for providing a reference implementation and her expertise for the LIBRISPEECH workloads. Thanks to Guodong Zhang for helpful suggestions regarding the rules and the submission API. We thank Dami Choi and Roger Grosse for helpful discussions. Furthermore, we would like to thank Varun Godbole for helpful feedback on this manuscript, Lucas Nestler for his help in implementing LIBRISPEECH workloads in JAX, and Kamal Raj for helping formulate Docker instructions. Finally, we'd like to especially thank all the members of the MLCommons Algorithms working group.

Frank Schneider is supported by funds from the Cyber Valley Research Fund. Philipp Hennig and Frank Schneider gratefully acknowledge financial support by the European Research Council through ERC StG Action 757275/PANAMA; the DFG Cluster of Excellence "Machine Learning - New Perspectives for Science", EXC 2064/1, project number 390727645; the German Federal Ministry of Education and Research (BMBF) through the Tübingen AI Center (FKZ:01IS18039A); and funds from the Ministry of Science, Research and Arts of the State of Baden-Württemberg. Daniel Snider is supported by funds from the Canada Foundation for Innovation JELF grant, NSERC Discovery grant, AWS Machine Learning Research Award (MLRA), Facebook Faculty Research Award, Google Scholar Research Award, and VMware Early Career Faculty Grant.

# Appendices

## A. Experimental Details for [Section 2](#)

### A.1 Learning Rate Schedules

Throughout the experiments in this paper (not only those reported in [Section 2](#)), we use two types of learning rate schedules. Both schedules are illustrated in [Figure 7](#) and their details are provided in the following.

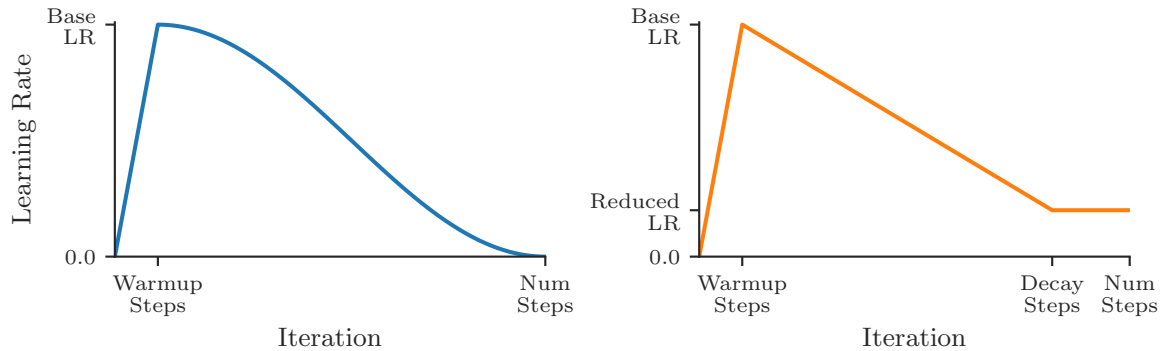


Figure 7: **The two learning rate schedules used in our experiments.** We use a cosine decay schedule with a learning rate warmup phase (**warmup + cosine decay**, *left*, —) and a linear decay schedule with a learning rate warmup phase and a constant phase at the end (**warmup + linear decay + constant**, *right*, —). The schedules are scaled to fill the entire NumSteps and are further parameterized by the parameters shown in the respective figures.

#### A.1.1 WARMUP COSINE DECAY

The cosine decay with warmup (denoted *warmup + cosine decay*) is characterized by a linear warmup phase followed by a cosine decay of the learning rate. It is parameterized by the following parameters:

- **Base LR:** The base learning rate. It is used as the peak learning rate of the warmup part and the cosine decay.
- **Num Steps:** The total number of steps of the schedule.
- **Warmup Steps:** The number of warmup steps over which the learning rate is linearly increased from zero to the base learning rate. Here, the warmup steps, characterize the absolute number of steps over which the warmup is performed. For simplicity, in our code, we instead provide the warmup phase in terms of the percentage of the total number of steps in the schedule (see for example [Table 10](#)).



Mathematically, the learning rate LR at any step  $t \geq 0$  of the *warmup + cosine decay* schedule is defined by

$$\text{LR}(t) = \begin{cases} \text{BaseLR} \frac{t}{\text{WarmupSteps}} & t \leq \text{WarmupSteps} \\ \frac{\text{BaseLR}}{2} \left( 1 + \cos \left( \pi \left( \frac{t - \text{WarmupSteps}}{\text{NumSteps} - \text{WarmupSteps}} \right) \right) \right) & \text{otherwise} . \end{cases}$$

### A.1.2 WARMUP LINEAR DECAY CONSTANT

The other learning rate schedule (denoted *warmup + linear decay + constant*) also employs a linear warmup phase, this time followed by a *linear* decay and a final phase of constant learning rate. It is parameterized by the following parameters:

- **Base LR:** The base learning rate. It is used as the upper learning rate of the warmup part and the linear decay.
- **Num Steps:** The total number of steps of the schedule.
- **Warmup Steps:** The number of warmup steps over which the learning rate is linearly increased from zero to the base learning rate. Here, the warmup steps, characterize the absolute number of steps over which the warmup is performed. For simplicity, in our code and the rest of the paper, we instead provide the warmup phase in terms of the percentage of the total number of steps in the schedule (see the entry *warmup* for example in [Table 10](#)).
- **Reduced LR:** The reduced learning rate is the lower bound of the linear decay and is used for the final constant phase of the schedule. Here, the reduced learning rate denotes the absolute learning rate. For simplicity, in our code and the remaining paper, we characterize the reduced learning rate by a *decay factor* (see the entry *decay factor* for example in [Table 10](#)) relative to the base learning rate.
- **Decay Steps:** The absolute number of steps of the linear decay (including the warmup steps). Once again, for simplicity in our code and throughout the paper, we use a slightly different version. There, we define the *decay steps* relative to the total number of steps excluding the warmup phase (see the entry *decay steps* for example in [Table 10](#)).

Mathematically, the learning rate LR at any step  $t \geq 0$  of the *warmup + linear decay + constant* schedule is defined by

$$\text{LR}(t) = \begin{cases} \text{BaseLR} \frac{t}{\text{WarmupSteps}} & t \leq \text{WarmupSteps} \\ \text{BaseLR} \frac{\text{DecaySteps} - t}{\text{DecaySteps} - \text{WarmupSteps}} + \text{ReducedLR} \frac{t - \text{WarmupSteps}}{\text{DecaySteps} - \text{WarmupSteps}} & \text{WarmupSteps} < t \leq \text{DecaySteps} \\ \text{BaseLR} \text{DecayFactor} & \text{otherwise} . \end{cases}$$

## A.2 Details for Training Curves that Cross

The training curves presented in [Figure 1](#) in [Section 2.1](#) were drawn from preliminary target-setting experiments for RESNET-50 trained on IMAGENET using ADAMW. They use the

workload configuration described in [Appendix D.3.1](#). The trials were selected from two arbitrary preliminary tuning studies that each had a budget of 100 trials. To select the two trials, we plotted all the training curves on a single plot, selecting two arbitrary trials that crossed and also achieved relatively good final validation error rates (in the top 20 trials for both studies). It was not necessary to find trials from different studies given how plentiful training curves that cross are in typical tuning studies. The search spaces were the same as the ADAMW search space listed in [Table 8](#), except they were from experiments that predate the final choice of allowed options for learning rate warmup lengths, disabled dropout, and tuned label smoothing on a continuous range of  $[0.0, 0.2]$ . These preliminary experiments also happened to be from studies using a batch size of 8192 and 32768, respectively. The particular trials selected used a warmup length of 15% and 10%; the trial with the better final validation error used a batch size of 8192 and a warmup length of 15%. Many pairs of trials in these studies cross, and the specific trials we selected were typical well-performing ones. It is easy to find training curves that cross multiple times among the top dozen trials in almost any of our larger tuning studies.

### A.3 Details for Sensitivity of Optimizer Ranking to the Model Architecture

#### A.3.1 WIDE RESNET WITH STRIDE CHANGES

The standard WIDE RESNET 28-10 architecture consists of 3 groups of 4 residual blocks, each block containing 2 convolutional layers. The default strides used in each of the three groups are  $1 \times 1$ ,  $2 \times 2$ , and  $2 \times 2$  in groups 1, 2, and 3 respectively. Our modified architectures changes the strides in group 3 from  $2 \times 2$  to  $1 \times 1$ . The strides are what reduce the height and width dimensions in the embedded image tensor, so changing the strides from  $2 \times 2$  to  $1 \times 1$  means that no reduction occurs in the final group. The final operation before the dense layer is an  $8 \times 8$  average pooling operation followed by a flatten operation. So all remaining height ( $H$ ), width ( $W$ ) and channel ( $C$ ) dimensions are flattened into a single dim of size  $H \times W \times C$ . In the stride 1x1 variant both  $H$  and  $W$  increase by a factor of 9, resulting in a factor of 81 increase in the fan-in of the final dense layer (in particular the dense layer shape changes from  $640 \times 10$  to  $51840 \times 10$ ). This dense layer is initialized with the LeCun normal initialization, which makes the variance of the forward pass invariant to the fan-in. However, the statistics of the backward pass are not invariant to the fan-in (and more importantly the loss curvature itself is not invariant) and so the resulting model has an instability particularly with respect to the final output layer (see [Figure 8](#)).

#### A.3.2 ARCHITECTURAL MODIFICATIONS OF TRANSFORMER MODELS

This section provides additional data for the experiment presented in [Section 2.2.1](#). Specifically, [Table 16](#) is a variant of [Table 3](#) but showing the cross-entropy loss instead of the BLEU score.

### A.4 Details for Comparing Instances of Training Algorithms

#### A.4.1 TRAINING ALGORITHMS WITH DIFFERENT HYPERPARAMETERS

This section provides additional details for the experiments described in [Section 2.3.1](#) and summarized in [Table 4](#). We report the performance of the *per-workload optimal* hyper-

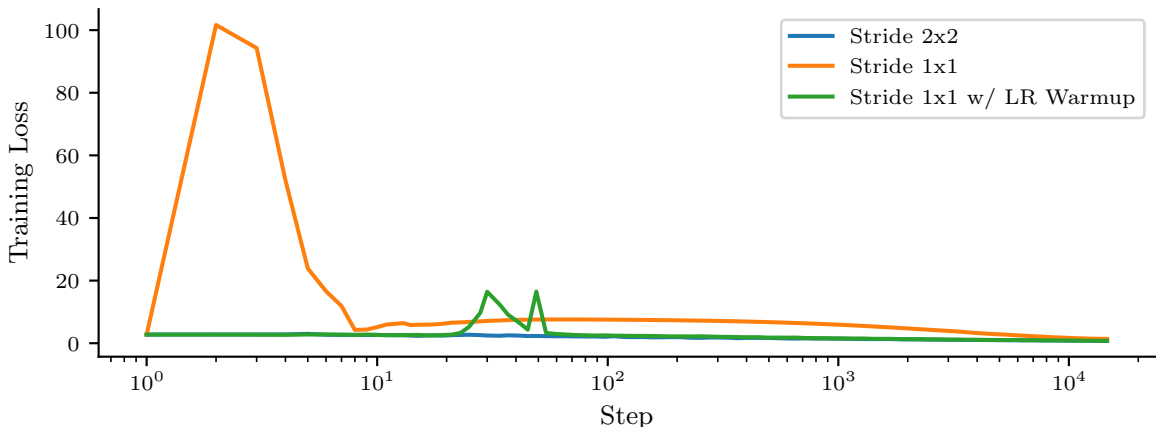


Figure 8: **Visualizing the training instability caused by changes to the Wide ResNet convolutional stride.** For the standard Stride 2x2 architecture (—), the training loss decreases almost monotonically throughout training. However, the Stride 1x1 architecture (—) shows a brief period of training instability, followed by a recover all within the first 10 steps of training. Despite the recovery, the long term optimization trajectory is affected and the stride change under-performs the standard architecture. Learning rate warmup helps mitigate this instability and allows the Stride 1x1 architecture to match the performance of the original model (—). All runs are shown for a learning rate of 0.215.

Training Algorithm	Pre-LN		Post-LN	
	Best	Confidence interval	Best	Confidence interval
ADAMW	1.3093	1.3449 ± 0.0337	1.3632	1.4934 ± 0.1327
NADAMW	1.3070	1.3308 ± 0.0240	1.3539	1.5072 ± 0.1822
NESTEROV	1.5668	1.6362 ± 0.1047	1.9577	2.5451 ± 0.6894
SHAMPOO	1.4096	1.4400 ± 0.0318	1.4516	1.5330 ± 0.1452

Table 16: **Test set cross-entropy score for Pre-Layer Norm (Pre-LN) and Post-Layer Norm (Post-LN) architectures for different training algorithms.** The architectural modification of PRE-LN vs. POST-LN affects NESTEROV momentum, ADAMW, NADAMW, and SHAMPOO differently. This is the same as Table 3 but showing the cross-entropy loss instead of the BLEU score.

parameters, of the *overall optimal* hyperparameters, as well as the resulting relative performance degradation  $\phi(H)$  from using a single shared hyperparameter setting, which is defined as

$$\phi_w(H) = \left| \frac{\text{val}(w, h^*) - \text{val}_H(w)}{\text{val}_H(w)} \right|$$

where  $h^*$  is the overall optimal hyperparameter setting. It holds that  $\max_w \phi_w(H) = \Phi(H)$  as reported in Section 2.3.1. Tables 17 to 20 show the quantities mentioned above for the four target-setting training algorithms.

Workload		Performance of the		$\phi_w(H)$
		<i>Per-Workload Optimal Hyperparameters</i>	<i>Overall Optimal Hyperparameters</i>	
CRITEO 1TB	DLRMSMALL	0.123675	0.124022	0.002806
FASTMRI	U-NET	0.734330	0.731403	0.003986
IMAGENET	RESNET 50	0.230340	0.267460	0.161153
	ViT	0.226140	0.263300	0.164323
LIBRISPEECH	CONFORMER	0.078327	0.093633	<b>0.195425</b>
	DEEPSPEECH	0.114152	0.133201	0.166873
OGBG	GNN	0.277534	0.252594	0.089862
WMT	TRANSFORMER	30.693876	29.067333	0.052992

Table 17: **Performance of the *per-workload optimal* and *overall optimal* hyperparameters for AdamW.** A column with  $\phi_w(H) = 0$  indicates that the overall optimal hyperparameters achieved the same performance as the per-workload optimal hyperparameters on this workload. The  $\phi_w$ -value in bold indicates the largest value observed across all workloads, which is reported as  $\Phi(H)$  in Table 4.

Workload		Performance of the		$\phi_w(H)$
		<i>Per-Workload Optimal Hyperparameters</i>	<i>Overall Optimal Hyperparameters</i>	
CRITEO 1TB	DLRMSMALL	0.123609	0.135988	0.100147
FASTMRI	U-NET	0.734523	0.710679	0.032462
IMAGENET	RESNET 50	0.227020	0.261320	0.151088
	ViT	0.225300	0.263420	<b>0.169197</b>
LIBRISPEECH	CONFORMER	0.077790	0.089341	0.148486
	DEEPSPEECH	0.113950	0.121887	0.069654
OGBG	GNN	0.280012	0.274704	0.018954
WMT	TRANSFORMER	30.853422	29.747388	0.035848

Table 18: **Performance of the *per-workload optimal* and *overall optimal* hyperparameters for NadamW.** A column with  $\phi_w(H) = 0$  indicates that the overall optimal hyperparameters achieved the same performance as the per-workload optimal hyperparameters on this workload. The  $\phi_w$ -value in bold indicates the largest value observed across all workloads, which is reported as  $\Phi(H)$  in Table 4.

Workload		Performance of the		$\phi_w(H)$
		<i>Per-Workload Optimal Hyperparameters</i>	<i>Overall Optimal Hyperparameters</i>	
CRITEO 1TB	DLRMSMALL	0.126139	0.144934	0.149004
FASTMRI	U-NET	0.734645	0.733440	0.001640
IMAGENET	RESNET 50	0.226600	0.278720	<b>0.230009</b>
	ViT	0.243180	0.278700	0.146065
LIBRISPEECH	CONFORMER	0.130823	0.130823	0.0
	DEEPSPEECH	0.171137	0.192623	0.125546
OGBG	GNN	0.283124	0.226850	0.198761
WMT	TRANSFORMER	30.107387	26.977169	0.103968

Table 19: **Performance of the *per-workload optimal* and *overall optimal* hyperparameters for Nesterov.** A column with  $\phi_w(H) = 0$  indicates that the overall optimal hyperparameters achieved the same performance as the per-workload optimal hyperparameters on this workload. The  $\phi_w$ -value in bold indicates the largest value observed across all workloads, which is reported as  $\Phi(H)$  in Table 4.

Workload		Performance of the		$\phi_w(H)$
		<i>Per-Workload Optimal Hyperparameters</i>	<i>Overall Optimal Hyperparameters</i>	
CRITEO 1TB	DLRMSMALL	0.125913	0.145933	0.158998
FASTMRI	U-NET	0.733828	0.731964	0.002539
IMAGENET	RESNET 50	0.225340	0.279280	<b>0.239372</b>
	ViT	0.244860	0.286660	0.170710
LIBRISPEECH	CONFORMER	0.132797	0.134879	0.015684
	DEEPSPEECH	0.161977	0.186385	0.150687
OGBG	GNN	0.276148	0.223901	0.189199
WMT	TRANSFORMER	30.643066	26.705241	0.128506

Table 20: **Performance of the *per-workload optimal* and *overall optimal* hyperparameters for Heavy Ball.** A column with  $\phi_w(H) = 0$  indicates that the overall optimal hyperparameters achieved the same performance as the per-workload optimal hyperparameters on this workload. The  $\phi_w$ -value in bold indicates the largest value observed across all workloads, which is reported as  $\Phi(H)$  in Table 4.

## A.4.2 TRAINING ALGORITHMS WITH DIFFERENT HYPERPARAMETER SEARCH SPACES

In [Table 21](#), we report additional data for the ADAMW search space comparison presented in [Section 2.3.2](#). Specifically, we show the results presented in [Table 6](#) but showing a different budget of tuning trials, i.e. 5 trials instead of the 20 trials reported in the main text.

Workload		AdamW Narrow			AdamW Broad		
		Median	$Q_1$	$Q_3$	Median	$Q_1$	$Q_3$
CRITEO 1TB	DLRMSMALL	<b>0.124039</b>	0.124016	0.124074	0.124319	0.124118	0.124634
FASTMRI	U-NET	<b>0.734512</b>	0.734335	0.734682	0.733791	0.733094	0.734065
IMAGENET	RESNET-50	<b>0.23382</b>	0.23298	0.2357	0.26842	0.24708	0.291495
	ViT	<b>0.22324</b>	0.22108	0.22526	0.24690	0.23924	0.27028
LIBRISPEECH	CONFORMER	<b>0.077553</b>	0.076817	0.079047	0.090477	0.085484	0.106039
	DEEPSPEECH	<b>0.115082</b>	0.112909	0.117964	0.131462	0.124718	0.155395
OGBG	GNN	<b>0.279426</b>	0.277748	0.281602	0.269102	0.265382	0.275955
WMT	TRANSFORMER	<b>31.1849</b>	30.9471	31.2992	30.4967	29.2948	30.9648

Table 21: **Performance across multiple workloads for AdamW with two different hyperparameter search spaces.** Shown are the median, as well as the lower and upper quartiles ( $Q_1$  and  $Q_3$ ) of the best observed validation metric. The results are for a budget of  $T=5$  trials across 1000 simulations. This table is similar to [Table 6](#) but showing a different budget of tuning trials (5 instead of 20).

## A.4.3 TRAINING ALGORITHMS WITH DIFFERENT TUNING GOALS

In this section, we provide more details from the experiment presented in [Section 2.3.2](#) and illustrated in [Figure 4](#). For this, we ran hyperparameter tuning studies for RESNET-50 on IMAGENET trained with ADAMW with two different training step budgets, 186, 666 and 233, 333. Both studies used the same search space and used 100 tuning trials of quasirandom search. They both used a cosine decay schedule and a linear learning rate warmup. Although the cosine decay schedule is a function of the maximum number of training steps, the only learning rate schedule parameter that was tuned beyond the peak learning rate was the length of the warmup, which was tuned over three discrete options: 2%, 5%, or 10% of the step budget (for the complete search space see [Table 8](#), ADAMW). We used the same seed for quasirandom search in both studies to generate the exact same set of 100 hyperparameter points. With the search space parameterized in this way to be relative to the training step budget and with the exact same set of 100 hyperparameter points, the hyperparameter setting achieving the best validation error happened to be the same across the two studies ([Table 22](#)). We selected the best trial based on minimum validation error achieved at *any* point during training, not just at the end of training, but with a cosine learning rate decay schedule we should expect the best result to be at—or near—the end of training.

B. Details for Target-Setting Experiments ([Section 5](#))

[Tables 23](#) and [24](#) provide the results of all 20 reruns of the target-setting training algorithm for each workload.

Hyperparameter	Value
Base LR	0.00040908031497988146
Weight decay	0.5107969085979827
$\beta_1$	0.9978657786056152
$\beta_2$	0.9961526971493336
Warmup	5%
Label smoothing	0.1986402587653765

Table 22: **Hyperparameter values for both runs shown in Figure 4.** Although both trials use a different step budget (186,666 vs. 233,333) the hyperparameter values found after tuning are the same.

Workload	Criteo 1TB fastMRI		ImageNet		LibriSpeech		OGBG	WMT
	DLRMsmall	U-Net	ResNet-50	ViT	Conformer	DeepSpeech	GNN	Transformer
Metric	CE↓	SSIM↑	Error Rate ↓		WER ↓		mAP↑	BLEU↑
0	0.123585	0.733429	0.22484	0.22436	0.076226	0.113374	0.275715	30.7062
1	0.123589	0.733868	0.22488	0.2255	0.07658	0.115416	0.276308	30.7142
2	0.123609	0.733994	0.2249	0.22606	0.076726	0.115487	0.277504	30.726
3	0.123619	0.734157	0.2251	0.22618	0.076862	0.115659	0.277586	30.7814
4	0.123624	0.734161	0.22512	0.22626	0.077372	0.11573	0.277711	30.7992
5	0.123631	0.734162	0.22512	0.22628	0.077453	0.11574	0.278202	30.7993
6	0.123634	0.734205	0.22518	0.22632	0.077517	0.115922	0.278523	30.8069
7	0.123635	0.734223	0.22534	0.22674	0.077599	0.115982	0.278594	30.8229
8	0.123637	0.734276	0.22546	0.2268	0.077653	0.116134	0.279844	30.829
9	0.123649	0.73432	0.22562	0.22688	0.078145	0.116144	0.280716	30.8458
10	0.123649	0.73448	0.22576	0.22694	0.078809	0.116255	0.281243	30.8524
11	0.123662	0.734483	0.2258	0.22726	0.078881	0.116478	0.281428	30.8571
12	0.123664	0.734532	0.22586	0.22766	0.079063	0.116791	0.281518	30.8631
13	0.123668	0.734587	0.22598	0.22776	0.079718	0.117135	0.281907	30.8787
14	0.123673	0.734605	0.22598	0.22796	0.079718	0.117378	0.282107	30.8835
15	0.123687	0.734658	0.22608	0.2284	0.084566	0.117398	0.282539	30.9064
16	0.1237	0.734674	0.22628	0.2284	0.091796	0.117438	0.282713	30.9196
17	0.123715	0.734712	0.22664	0.22854	0.093352	0.117812	0.282918	30.9267
18	0.123728	0.734804	0.22706	0.22886	0.098336	0.131573	0.283148	30.9446
19	0.12374	0.735031	0.22746	0.2293	0.104256	0.171481	0.288737	31.0684
min	0.123585	0.733429	0.22484	0.22436	0.076226	0.113374	0.275715	30.7062
25th	0.12363	0.734162	0.22512	0.226275	0.077433	0.115737	0.278079	30.7993
median	0.123649	0.7344	0.22569	0.22691	0.078477	0.1162	0.28098	30.8491
75th	0.123676	0.734619	0.226005	0.22807	0.08093	0.117383	0.282215	30.8892
max	0.12374	0.735031	0.22746	0.2293	0.104256	0.171481	0.288737	31.0684
max – min	0.000155	0.001602	0.00262	0.00494	0.028031	0.058107	0.013022	0.3621
$\frac{\max - \min}{\min}$ (%)	0.13	0.22	1.17	2.20	36.77	51.25	4.72	1.18

Table 23: **Validation evaluation metric results of the 20 reruns of the top training algorithm and hyperparameter combination for each workload.**



Workload	Criteo 1TB	fastMRI	ImageNet		LibriSpeech		OGBG	WMT
	DLRMsmall	U-Net	ResNet-50	ViT	Conformer	DeepSpeech	GNN	Transformer
Metric	CE↓	SSIM↑	Error Rate ↓		WER ↓		mAP↑	BLEU↑
0	0.125989	0.7404	0.3371	0.3396	0.044773	0.065546	0.263254	30.6455
1	0.126002	0.74106	0.3405	0.3403	0.044849	0.066671	0.265461	30.7219
2	0.126029	0.741231	0.3412	0.3408	0.045175	0.066862	0.266552	30.7898
3	0.126035	0.741236	0.3415	0.3414	0.04571	0.066883	0.267611	30.8858
4	0.126039	0.741361	0.3417	0.3416	0.046055	0.066926	0.268155	30.8996
5	0.126042	0.741451	0.3425	0.3422	0.046189	0.067499	0.268729	30.9216
6	0.126043	0.74157	0.343	0.3422	0.046265	0.067584	0.268964	30.9228
7	0.126044	0.741643	0.3432	0.3426	0.046476	0.067732	0.26905	30.947
8	0.126051	0.741652	0.3432	0.3428	0.046552	0.067944	0.269792	30.9692
9	0.126052	0.741671	0.3434	0.3432	0.046629	0.067944	0.269988	30.9959
10	0.126054	0.741699	0.3436	0.3432	0.046763	0.068008	0.270238	30.9987
11	0.12606	0.741721	0.3437	0.3437	0.046954	0.068029	0.270672	31.005
12	0.126071	0.741738	0.3438	0.3438	0.046973	0.068093	0.270756	31.0129
13	0.126076	0.741778	0.3438	0.3439	0.047184	0.068369	0.271233	31.0221
14	0.126076	0.741827	0.3439	0.3444	0.048638	0.068454	0.271476	31.0408
15	0.126087	0.7419	0.344	0.3451	0.050551	0.069112	0.272067	31.0412
16	0.126087	0.741909	0.3449	0.3451	0.055086	0.069282	0.272105	31.1161
17	0.126126	0.741915	0.3453	0.3457	0.057286	0.069494	0.27285	31.1407
18	0.126128	0.742025	0.3462	0.3466	0.060749	0.081402	0.276531	31.1446
19	0.126136	0.742195	0.3481	0.3481	0.066317	0.110948	0.276976	31.1679
min	0.125989	0.7404	0.3371	0.3396	0.044773	0.065546	0.263254	30.6455
25th	0.126041	0.741428	0.3423	0.34205	0.046155	0.067355	0.268586	30.9161
median	0.126053	0.741685	0.3435	0.3432	0.046696	0.067976	0.270113	30.9973
75th	0.126079	0.741845	0.343925	0.344575	0.049116	0.068618	0.271624	31.0409
max	0.126136	0.742195	0.3481	0.3481	0.066317	0.110948	0.276976	31.1679
max - min	0.000148	0.001796	0.011	0.0085	0.021544	0.045402	0.013722	0.5225
$\frac{\text{max} - \text{min}}{\text{min}}$ (%)	0.12	0.24	3.26	2.50	48.12	69.27	5.21	1.70

Table 24: Test evaluation metric results of the 20 reruns of the top training algorithm and hyperparameter combination for each workload.

## C. Details for Baseline Experiments (Section 7)

In this section, we provide additional results from the baseline experiments presented in Section 7. Tables 25 and 26 report the total runtime and number of steps needed to achieve the target performance for all baselines. The benchmark scores for all baselines presented in the paper are listed in Table 27. Table 28 presents the estimated runtimes for a given number of steps of the different training algorithms used in the baselines. We report the 20 hyperparameter points used in our OPTLIST baselines in Tables 29 to 32.

Algorithm	Criteo 1TB	fastMRI	ImageNet		LibriSpeech		OGBG	WMT
	DLRMsmall	U-Net	ResNet-50	ViT	Conformer	DeepSpeech	GNN	Transformer
<b>AdamW</b>								
TUNED $\beta_1$	5622	inf	inf	62,667	95,222	80,106	inf	40,534
FIXED $\beta_1$	<b>5320</b>	7473	inf	62,667	inf	81,946	inf	41,499
OPT-LIST	5471	<b>6415</b>	inf	64,213	88,135	<b>76,427</b>	inf	44,391
<b>Heavy Ball</b>								
TUNED $\beta_1$	inf	inf	inf	inf	inf	inf	inf	inf
FIXED $\beta_1$	inf	inf	inf	inf	inf	inf	inf	inf
OPT-LIST	inf	inf	<b>57,321</b>	inf	inf	inf	inf	43,984
<b>LAMB</b>								
TUNED $\beta_1$	inf	inf	inf	inf	inf	78,966	inf	<b>29,962</b>
<b>NadamW</b>								
TUNED $\beta_1$	5850	8559	inf	62,005	92,558	79,569	<b>11,441</b>	30,822
FIXED $\beta_1$	5544	inf	61,049	60,457	inf	79,569	inf	43,329
OPT-LIST	5544	8205	inf	<b>59,682</b>	<b>87,475</b>	77,721	12,914	44,291
<b>Nesterov</b>								
TUNED $\beta_1$	inf	inf	inf	inf	inf	inf	inf	inf
FIXED $\beta_1$	inf	inf	inf	inf	inf	inf	inf	inf
OPT-LIST	inf	8750	59,330	inf	inf	inf	inf	inf
<b>Adafactor</b>								
TUNED $\beta_1$	inf	inf	inf	inf	inf	82,214	inf	37,679
<b>SAM(w. Adam)</b>								
TUNED $\beta_1$	5910	inf	inf	inf	inf	inf	inf	inf

Table 25: **Total runtime to achieve the target performance for different baselines.** These numbers are used to plot the performance profiles in Figure 5a and Figure 6a. A value of inf indicates that that baseline was unable to achieve the target within the maximum allowed runtime.

## D. Workload Details

### D.1 Criteo 1TB

We train on the CRITEO 1TB Click Logs dataset (Lab, 2014) to train a standard ads recommender model, DLRM (Naumov et al., 2019) to predict the CTR. The dataset contains examples with 13 numerical features and 26 categorical. The numerical features are log-transformed and the 26 categorical features are hashed into a single embedding table. The data is from 24 days of click data, split into one file per day. We use the first 23 days as the training split, resulting in 4,195,197,692 training examples. We then use the first half of

Algorithm	Criteo 1TB	fastMRI	ImageNet		LibriSpeech		OGBG	WMT
			DLRMsmall	U-Net	ResNet-50	ViT		
<b>AdamW</b>								
TUNED $\beta_1$	7881	inf	inf	151,146	250,604	69,600	inf	111,972
FIXED $\beta_1$	<b>7455</b>	30,324	inf	151,146	inf	71,200	inf	114,638
OPT-LIST	7668	<b>25,992</b>	inf	154,878	231,942	<b>66,400</b>	inf	122,636
<b>Distributed Shampoo</b>								
TUNED $\beta_1$	inf	32,851	181,002	<b>138,084</b>	<b>229,276</b>	67,200	<b>35,200</b>	90,644
<b>Heavy Ball</b>								
TUNED $\beta_1$	inf	inf	inf	inf	inf	inf	inf	inf
FIXED $\beta_1$	inf	inf	inf	inf	inf	inf	inf	inf
OPT-LIST	inf	inf	169,806	inf	inf	inf	inf	122,636
<b>LAMB</b>								
TUNED $\beta_1$	inf	inf	inf	inf	inf	68,800	inf	<b>79,980</b>
<b>NadamW</b>								
TUNED $\beta_1$	8094	34,656	inf	149,280	242,606	68,800	49,600	85,312
FIXED $\beta_1$	7668	inf	181,002	145,548	inf	68,800	inf	119,970
OPT-LIST	7668	33,212	inf	143,682	<b>229,276</b>	67,200	56,000	122,636
<b>Nesterov</b>								
TUNED $\beta_1$	inf	inf	inf	inf	inf	inf	inf	inf
FIXED $\beta_1$	inf	inf	inf	inf	inf	inf	inf	inf
OPT-LIST	inf	35,739	175,404	inf	inf	inf	inf	inf
<b>Adafactor</b>								
TUNED $\beta_1$	inf	inf	inf	158,610	261,268	70,400	inf	98,642
<b>SAM(w. Adam)</b>								
TUNED $\beta_1$	8307	inf	<b>166,118</b>	147,414	231,942	67,200	inf	98,642

Table 26: **Number of steps to achieve the target performance for different baselines.** These numbers are used to plot the performance profiles in [Figure 5b](#) and [Figure 6b](#). A value of inf indicates that that baseline was unable to achieve the target in the maximum allowed step budget.

Submission	Version	Benchmark Score	
		Runtime	Steps
ADAMW	TUNED $\beta_1$	0.600141	0.596116
	FIXED $\beta_1$	0.596985	0.593047
	OPT-LIST	0.725260	0.721035
DISTRIBUTED SHAMPOO	TUNED $\beta_1$	-	<b>0.85421</b>
HEAVY BALL	TUNED $\beta_1$	0	0
	FIXED $\beta_1$	0	0
	OPT-LIST	0.230504	0.22686
LAMB	TUNED $\beta_1$	0.248618	0.248494
NADAMW	TUNED $\beta_1$	<b>0.849960</b>	0.595478
	FIXED $\beta_1$	0.599691	0.830414
	OPT-LIST	0.835602	0.813194
NESTEROV	TUNED $\beta_1$	0	0
	FIXED $\beta_1$	0	0
	OPT-LIST	0.233373	0.232048
ADAFCTOR	TUNED $\beta_1$	0.236111	0.47576
SAM(w. ADAM)	TUNED $\beta_1$	0.120368	0.731717

Table 27: Benchmark scores for all baselines presented in the paper.

Algorithm	Criteo 1TB fastMRI		ImageNet		LibriSpeech		OGBG	WMT
	DLRMsmall	U-Net	ResNet-50	ViT	Conformer	DeepSpeech	GNN	Transformer
Steps	10,667	36,189	186,667	186,667	133,333	80,000	80,000	133,333
ADAMW	7602	8906	62,827	77,380	101,368	92,064	18,361	48,261
HEAVY BALL	7042	8962	<b>63,008</b>	76,535	105,506	91,460	18,303	47,818
LAMB	7166	8972	67,797	85,254	101,982	91,802	19,330	49,911
NADAMW	<b>7703</b>	8935	62,958	<b>77,520</b>	<b>101,780</b>	<b>92,509</b>	18,438	<b>48,151</b>
NESTEROV	7097	<b>8859</b>	63,137	78,791	105,108	90,874	<b>18,477</b>	47,855
ADAFCTOR	7419	9058	67,381	93,895	122,920	93,404	20,379	50,896
SAM(w. ADAM)	7583	9048	117,997	146,507	196,144	180,832	20,852	95,139

Table 28: **Runtime measurements (rounded to the nearest second) for different training algorithms used in baselines.** For each workload and training algorithm we measure the runtime for the mentioned number of steps. The highlighted entry for each workload corresponds to the runtime for the target-setting algorithm for that workload and thus defines our maximum allowed runtime for each workload.

Learning Rate	$\beta_1$	$\beta_2$	Weight Decay	Warmup	Dropout	Aux. Dropout	Label Smoothing
0.002995	0.960644	0.998968	0.006842	5	0	0	0
0.003331	0.948	0.998793	0.003578	2	0.1	0.1	0
0.007502	0.86932	0.989658	0.000071	10	0	0	0
0.000644	0.931783	0.969068	0.677756	5	0	0	0
0.002827	0.885906	0.820717	0.185407	2	0	0	0
0.001308	0.973133	0.998123	0.163753	10	0	0	0.1
0.018978	0.966607	0.996816	0.015654	5	0.1	0.1	0.2
0.001949	0.532796	0.981278	0.037642	5	0	0	0.1
0.000845	0.889576	0.99785	0.081354	5	0	0	0.2
0.000987	0.99139	0.993211	0.00835	10	0	0	0.2
0.002107	0.823119	0.877457	0.275905	2	0.1	0.1	0
0.002234	0.751323	0.612946	0.215092	5	0	0	0
0.003926	0.813913	0.987628	0.028657	2	0	0	0
0.001949	0.532796	0.981278	0.037642	5	0	0	0
0.004958	0.863744	0.629185	0.114739	2	0	0	0
0.000584	0.962501	0.998687	0.000148	5	0.1	0.1	0.1
0.000504	0.923701	0.99475	0.000012	5	0.1	0.1	0.2
0.000388	0.674831	0.946874	0.124668	10	0.1	0.1	0
0.001749	0.932661	0.995516	0.081216	2	0.1	0.1	0
0.000275	0.788345	0.076547	0.023545	2	0.1	0.1	0.2

Table 29: **OptList for AdamW**. The learning rate schedule used is warmup + cosine decay.

Learning Rate	$\beta_1$	$\beta_2$	Weight Decay	Warmup	Dropout	Aux. Dropout	Label Smoothing
0.003331	0.948	0.998793	0.003578	2	0.1	0.1	0
0.001614	0.959792	0.998463	0.000033	5	0	0	0
0.010937	0.974179	0.998111	0.007607	5	0.1	0.1	0
0.005146	0.994362	0.994663	0.246009	10	0	0	0
0.002827	0.885906	0.820717	0.185407	2	0	0	0
0.001308	0.973133	0.998123	0.163753	10	0	0	0.1
0.018978	0.966607	0.996816	0.015654	5	0.1	0.1	0.2
0.000845	0.889576	0.99785	0.081354	5	0	0	0.2
0.001949	0.532796	0.981278	0.037642	5	0	0	0.1
0.000987	0.99139	0.993211	0.00835	10	0	0	0.2
0.001308	0.973133	0.998123	0.163753	10	0	0	0
0.000845	0.889576	0.99785	0.081354	5	0.1	0.1	0
0.004958	0.863744	0.629185	0.114739	2	0	0	0
0.003528	0.819231	0.495851	0.043397	10	0	0	0
0.001308	0.973133	0.998123	0.163753	10	0	0	0
0.003296	0.996693	0.998649	0.003729	10	0.1	0.1	0.1
0.000584	0.962501	0.998687	0.000148	5	0.1	0.1	0.1
0.000279	0.991934	0.997984	0.000324	10	0.1	0.1	0.1
0.001749	0.932661	0.995516	0.081216	2	0.1	0.1	0
0.001011	0.712472	0.966607	0.000069	5	0.1	0.1	0.1

Table 30: **OptList for NadamW**. The learning rate schedule used is warmup + cosine decay.

Learning Rate	$\beta_1$	Weight Decay	Warmup	Decay Steps	End Factor	Dropout	Aux. Dropout	Label Smoothing
0.333132	0.948	1.40e-7	5	0.942079	0.01	0.1	0.1	0
0.082037	0.980735	1.01e-6	5	0.891621	0.01	0.1	0.1	0
0.810523	0.898228	1.00e-7	5	0.842587	0.01	0.1	0.1	0
0.028609	0.981543	5.77e-4	5	0.984398	0.01	0	0	0
0.416058	0.970426	1.99e-5	5	0.936585	0.01	0	0	0
4.131896	0.927476	5.67e-6	5	0.900777	0.001	0	0	0.2
0.191165	0.995978	3.83e-6	5	0.871275	0.01	0.1	0.1	0.2
1.376742	0.736477	5.09e-6	5	0.977277	0.01	0	0	0.2
0.032559	0.988578	3.32e-6	5	0.876362	0.001	0	0	0.1
0.130821	0.973133	2.90e-7	5	0.816545	0.001	0	0	0.2
0.022941	0.984057	2.40e-7	5	0.924988	0.01	0.1	0.1	0
0.010036	0.986308	3.22e-5	5	0.994571	0.01	0	0	0
0.026287	0.992389	3.88e-4	5	0.945944	0.01	0.1	0.1	0
0.014244	0.970264	4.22e-4	5	0.940451	0.01	0	0	0
0.019827	0.95789	2.41e-4	5	0.80861	0.001	0.1	0.1	0
2.491773	0.944937	1.30e-7	5	0.861509	0.001	0.1	0.1	0
2.051309	0.917965	4.58e-6	5	0.82041	0.001	0.1	0.1	0.1
1.897755	0.966607	6.90e-7	5	0.987857	0.01	0.1	0.1	0.1
0.169804	0.99636	1.03e-6	5	0.998233	0.001	0.1	0.1	0.1
0.253647	0.989819	1.15e-6	5	0.932109	0.01	0	0	0.1

Table 31: **OptList for Nesterov.** The learning rate schedule used is warmup + linear decay + constant.

Learning Rate	$\beta_1$	Weight Decay	Warmup	Decay Steps	Decay Factor	Dropout	Aux. Dropout	Label Smoothing
0.299534	0.960644	1.10e-7	5	0.839739	0.01	0.1	0.1	0
5.133865	0.673928	2.50e-7	5	0.98745	0.01	0	0	0
1.042065	0.862619	3.90e-7	5	0.920716	0.01	0.1	0.1	0
0.028609	0.981543	5.77e-4	5	0.984398	0.01	0	0	0
0.416058	0.970426	1.99e-5	5	0.936585	0.01	0	0	0
4.131896	0.927476	5.67e-6	5	0.900777	0.001	0	0	0.2
7.263293	0.860749	2.81e-6	5	0.889586	0.01	0.1	0.1	0.2
1.376742	0.736477	5.09e-6	5	0.977277	0.01	0	0	0.2
1.111547	0.589943	6.31e-6	5	0.911763	0.01	0	0	0.1
0.392622	0.813913	6.62e-6	5	0.854795	0.01	0	0	0.2
0.130821	0.973133	2.90e-7	5	0.816545	0.001	0.1	0.1	0
0.022941	0.984057	2.40e-7	5	0.924988	0.01	0.1	0.1	0
0.142788	0.961398	1.83e-6	5	0.84625	0.01	0	0	0
0.027867	0.991934	3.20e-7	5	0.923564	0.001	0	0	0
0.026287	0.992389	3.88e-4	5	0.945944	0.01	0.1	0.1	0
2.051309	0.917965	4.58e-6	5	0.82041	0.001	0.1	0.1	0.1
1.897755	0.966607	6.90e-7	5	0.987857	0.01	0.1	0.1	0.1
2.491773	0.944937	1.30e-7	5	0.861509	0.001	0.1	0.1	0
0.426111	0.995127	9.80e-7	5	0.934347	0.001	0	0	0.2
0.169804	0.99636	1.03e-6	5	0.998233	0.001	0.1	0.1	0.1

Table 32: **OptList for Heavy Ball.** The learning rate schedule used is warmup + linear decay + constant.

the 24th day as the test set, and the second half of the 24th day as a validation set, resulting in 89,137,319 validation and 89,137,318 test examples. See [here](#) for the implementation of our CRITEO 1TB input pipeline. Unlike many other Criteo 1TB pipelines, instead of the AUC we use the sigmoid binary cross entropy loss as an evaluation metric. We do this to have a metric that decomposes elementwise, which avoids requiring submitters to run expensive AUC evaluations that would have required maintaining arrays the size of the roughly 89 million evaluation examples.

We split the training split into files with 5,000,000 lines each and randomly shuffle them. However, the training time budget we set typically only allows for around 60% of an epoch to be consumed at the batch sizes we can fit into memory. In our [paper experiment codebase](#), we reshuffle the dataset each time we have a preemption, which could result in repeated examples, but in the [benchmark codebase](#) we do not assume that the code runs on preemptable instances so this is not a concern.

### D.1.1 DLRMSMALL MODEL

The concatenated features form the input layer for the DLRM model ([Naumov et al., 2019](#)). The single embedding table is of size 4M entries with an embedding dimension of 128. The dense features are fed into a three-layer fully-connected network with 512, 256, 128 units per layer. The outputs of this layer are then concatenated to the embedding lookups of the categorical features, and fed into the cross-interaction layer. Finally, the cross-interaction output is passed into a five-layer fully-connected network with 1024, 1024, 512, 256, 1 units per layer. All layers use a ReLU activation, except for the final 1d output. A dropout layer follows the 512 dimensional layer in the second network, after the ReLU activation. See [here](#) for our DLRM model in Jax and [here](#) for our DLRM model in PyTorch.

### D.1.2 CRITEO 1TB DLRMSMALL WORKLOAD VARIANTS

The three workload variants for the DLRMSMALL model are:

- **Embed Init Scale:** We changed the initialization of the embedding layer from  $\mathcal{N}\left(0, \frac{1}{\sqrt{4194304}}\right) \approx \mathcal{N}(0, 0.00049)$  (where 4194304 is the vocabulary size), to  $\mathcal{N}(0, 1)$ .
- **LayerNorm:** Layer normalization was added after the activations of each layer, except for the final 1d output.
- **Residual:** For every layer after the first layer of each subnetwork, instead of being a simple fully-connected layer, it is the residual branch in a residual subnetwork.

## D.2 fastMRI

The FASTMRI ([Zbontar et al., 2018](#)) dataset was released by NYU Langone Health as part of a challenge organized in collaboration with Facebook AI Research (FAIR, now Meta AI). The challenge aims to reduce the time to acquire an MRI scan by up to a factor of ten without any loss in diagnostic quality. In MRI acquisition, a “pulse sequence” of spatially- and temporally-varying magnetic fields induces the subject’s body to emit electromagnetic response fields that are measured by one or more receiver coils placed near the area to be imaged. Using these fine-grained Fourier-space measurements (*k-space*



Variant	Validation Performance	Hyperparameter Performance Transfer			
		BASE $\rightarrow$ VARIANT		VARIANT $\rightarrow$ BASE	
		Performance	Rank	Performance	Rank
BASE WORKLOAD	<b>0.123609</b>				
EMBED INIT SCALE	0.123800	0.123880	3	0.123920	16
LAYERNORM	0.123653	0.123797	14	0.127778	136
RESIDUAL	0.123920	0.124010	5	0.123860	13

Table 33: **Hyperparameter transfer between the base workload and the variants of DLRMsmall on Criteo 1TB.** We show the validation performance of the base workload compared to the performance achieved by the variants. Further, we show the performance (and hyperparameter ranking) when using the optimal hyperparameter point from the base workload on the variants (BASE  $\rightarrow$  VARIANT) and vice-versa (VARIANT  $\rightarrow$  BASE). All runs are from the same search space for NADAMW. See [Section 6.2](#) for a detailed description of our protocol for accepting workload variants.

in the medical literature), clinicians reconstruct volumetric images with high-quality soft tissue contrast. However, while powerful, these images can take 30 minutes or more to produce, causing logistical issues ranging from patient discomfort to low patient throughput. Recent strides in machine learning seek to reduce this time by reconstructing the volumetric images from sub-sampled  $k$ -space measurements. The FASTMRI task provides raw  $k$ -space data, randomly masks this data to simulate sub-sampling, and asks a supplied algorithm to faithfully reconstruct the image as compared to the inverse Fourier transform of the complete (non-sub-sampled)  $k$ -space data.

In this benchmark, we use FASTMRI’s single-coil knee data which is organized into volumes, with each volume coming from one patient and being composed of a collection of slices (i.e., images). In this task, we treat each slice as an independent example. The official data set contains 34,742 training slices (from 973 volumes), and 7,135 validation slices (from 199 volumes). In the data pipeline, we use a fixed random sequence to mask certain columns (down-sampling by a factor of four) from each raw  $k$ -space slice before applying an inverse Fourier transform and normalizing the resulting image. The target ground-truth image is normalized using the same mean and standard deviation. Since the official test set targets are not publicly available, we split the validation set roughly in half to obtain disjoint validation and test sets—the first 100 validation HDF5 files are used for validation and contain 3,554 slices, while the final 99 validation HDF5 are used for test and contain 3,581 slices. See [here](#) for the implementation of our FASTMRI input pipeline.

### D.2.1 U-NET MODEL

We train a U-NET model similar to the one described in [Ronneberger et al. \(2015\)](#). Our U-NET implementation has 32 channels, four down-sampling convolutional blocks, and four up-sampling transpose convolutional/convolutional block pairs. Each layer uses dropout with a rate that may be tuned by the submission; the default dropout rate is 0.0. See [here](#)

for the implementation of our U-NET model in Jax and [here](#) for the implementation of our U-NET model in PyTorch.

### D.2.2 FASTMRI U-NET WORKLOAD VARIANTS

The three workload variants for the U-NET model are:

- **Channels & Pooling:** The base number of channels in each convolution block was increased to 64 (which is multiplied by another factor of 2 with each down or up sample level), and the number of down and up sample layers was decreased to 3.
- **TanH:** Activation functions were swapped to TanH.
- **LayerNorm:** Instance normalization was swapped to layer normalization (which has learnable parameters).

Variant	Validation Performance	Hyperparameter Performance Transfer			
		BASE $\rightarrow$ VARIANT		VARIANT $\rightarrow$ BASE	
		Performance	Rank	Performance	Rank
BASE WORKLOAD	0.734523				
CHANNELS & POOLING	0.734603	0.734172	12	0.734277	8
TANH	0.729743	0.728203	6	0.734438	3
LAYERNORM	<b>0.734968</b>	0.733304	26	0.734364	6

Table 34: **Hyperparameter transfer between the base workload and the variants of U-Net on fastMRI.** We show the validation performance of the base workload compared to the performance achieved by the variants. Further, we show the performance (and hyperparameter ranking) when using the optimal hyperparameter point from the base workload on the variants (BASE  $\rightarrow$  VARIANT) and vice-versa (VARIANT  $\rightarrow$  BASE). All runs are from the same search space for NADAMW. See [Section 6.2](#) for a detailed description of our protocol for accepting workload variants.

## D.3 ImageNet

For our IMAGENET workload we used the ILSVRC 2012 training and validation sets as the training and validation splits ([Deng et al., 2009](#)), and IMAGENET-v2 as the test split ([Recht et al., 2019](#)). For training preprocessing we take a random crop and randomly flip the image, whereas for the validation and test sets we just take a center crop (colloquially referred to as "ResNet preprocessing"). Images are fed into the model normalized to  $[0, 1]$ . We used RandAugment ([Cubuk et al., 2020](#)) and mixup ([Zhang et al., 2018](#)) for our ViT workload but not for RESNET. Code for our IMAGENET input pipelines can be found [here](#) in tf.data, [here](#) in PyTorch, and [here](#) for IMAGENET-v2.

### D.3.1 RESNET-50 MODEL

With the exception of [Section 2.2.1](#), all experiments use the RESNET-50 defined in [He et al. \(2016a, Section 4.1\)](#). We use ghost batch normalization with a virtual batch size of 64

(Hoffer et al., 2017). To further improve optimization stability, the scales in the final batch normalization layer in each residual block are initialized to all zeros. This has the effect of initializing each residual block to be the identity function. See here for our implementation of RESNET-50 in JAX and here for our implementation of RESNET-50 in PYTORCH.

In Section 2.2.1 we ran experiments on an unstable 200-layer RESNETV2 architecture (He et al., 2016b). These experiments also used a virtual batch size of 64. In contrast to the above RESNET-50, we initialize all batch normalization layers in the default way (with scalar factor all initialized to 1). This was necessary to produce the training instability at initialization. The extra batch normalization layer was added right after every residual connection, so each residual block returns  $BN(x + F(x))$  instead of the standard  $x + F(x)$ .

### D.3.2 IMAGENET RESNET-50 WORKLOAD VARIANTS

The three workload variants for the U-NET model are:

- **SiLU**: Changed activation functions to SiLU (Elfwing et al., 2018).
- **GELU**: Changed activation functions to GELU (Hendrycks and Gimpel, 2016).
- **BN Init Scale**: Changed the initialization of the scale variable in the final batch normalization layer in the residual branches from 0.0 to 8.0.

In Table 35 we present the results of our protocol to test our variants (described in Section 6.2). We observe that the ranks of hyperparameter transfers were not as high as variants in other models. In such cases we tested these models along an alternative protocol to ensure that the optimum learning rate is indeed different for these models.

Variant	Validation Performance	Hyperparameter Performance Transfer			
		BASE $\rightarrow$ VARIANT		VARIANT $\rightarrow$ BASE	
		Performance	Rank	Performance	Rank
BASE WORKLOAD	0.22702				
SiLU	<b>0.21996</b>	0.24234	9	0.22980	2
GELU	0.22148	0.22858	3	0.22980	2
BN INIT SCALE	0.23502	0.23502	1	0.22702	1

Table 35: **Hyperparameter transfer between the base workload and the variants of ResNet-50 on ImageNet.** We show the validation performance of the base workload compared to the performance achieved by the variants. Further, we show the performance (and hyperparameter ranking) when using the optimal hyperparameter point from the base workload on the variants (BASE  $\rightarrow$  VARIANT) and vice-versa (VARIANT  $\rightarrow$  BASE). All runs are from the same search space for NADAMW. See Section 6.2 for a detailed description of our protocol for accepting workload variants.

### D.3.3 VISION TRANSFORMER MODEL

For all experiments, we use the S/16 variant of the VISION TRANSFORMER (ViT) (Dosovitskiy et al., 2021) as enumerated in Steiner et al. (2022) with no regularization and `light2`

data augmentation. We chose the S/16 variant given its relatively small size (width of 384, depth of 12, MLP dimension of 1536, with 6 heads and a  $16 \times 16$  image patch size). Steiner et al. (2022) define light2 data augmentation as using mixup (Zhang et al., 2018) per batch, with  $\alpha = 0.2$  and RandAugment (Cubuk et al., 2020) per image with two layers and magnitude 15. See here for our implementation of VISION TRANSFORMER in Jax and here for our implementation of VISION TRANSFORMER in PyTorch.

#### D.3.4 IMAGENET VISION TRANSFORMER WORKLOAD VARIANTS

The three workload variants for the U-NET model are:

- **Post-LN**: Layer normalization was applied after the residual branch was added back into the trunk.
- **MAP**: The pooling layer type was changed from global average pooling to multihead attention pooling.
- **GLU**: We included gated linear units in the MLPBLOCK.

In Table 36 we present the results of our protocol to test our variants (described in Section 6.2). We observe that the ranks of hyperparameter transfers were 0 for two variants. For these variants we tested these models along an alternative protocol to ensure that the optimum learning rate is indeed different for these models.

Variant	Validation Performance	Hyperparameter Performance Transfer			
		BASE $\rightarrow$ VARIANT		VARIANT $\rightarrow$ BASE	
		Performance	Rank	Performance	Rank
BASE WORKLOAD	0.22530				
POST-LN	0.24794	0.24794	14	0.28478	5
MAP	0.23004	0.23004	0	0.22530	0
GLU	<b>0.22258</b>	0.22258	0	0.22530	0

Table 36: **Hyperparameter transfer between the base workload and the variants of ViT on ImageNet.** We show the validation performance of the base workload compared to the performance achieved by the variants. Further, we show the performance (and hyperparameter ranking) when using the optimal hyperparameter point from the base workload on the variants (BASE  $\rightarrow$  VARIANT) and vice-versa (VARIANT  $\rightarrow$  BASE). All runs are from the same search space for NADAMW. See Section 6.2 for a detailed description of our protocol for accepting workload variants.

#### D.4 LibriSpeech

The LIBRISPEECH dataset (Panayotov et al., 2015) is a corpus of read English speech with sampling rate of 16 kHz. We train our speech recognition models on the combination of train-clean-100, train-clean-360, and train-other-500 splits from the LIBRISPEECH dataset giving us 960 hours of raw audio data. For validation we use a combination of

`dev-clean` and `dev-other` splits resulting in 5567 examples in the validation set. We report the word error rates on the `test-clean` split as our test set performance.

We preprocess the data by eliminating any examples with audio length greater than 320k and a target sentence length greater than 256. We then compute logmel spectrogram features for the raw audio input and use the `SentencePiece` (Kudo and Richardson, 2018) tokenizer with a vocab size of 1024 to tokenize the target sentences. We pad the sequences to bring all examples to same length and handle paddings inside the model and while computing metrics like loss and word error rate. See [here](#) for the implementation of our LIBRISPEECH input pipeline.

#### D.4.1 CONFORMER MODEL

CONFORMER (Gulati et al., 2020) is an architecture combining attention and convolution layers to capture both global and local relationships in input audio. We use a 4-layer deep CONFORMER model with model encoder dimension of 512. Model weights are initialized using Xavier uniform initialization. CTC loss (Graves et al., 2006) is used to train the model. See [here](#) for our implementation of CONFORMER in Jax and [here](#) for our implementation of CONFORMER in PyTorch.

**Inference** We use a greedy decoding procedure to generate decoded logits which are then transformed back into sentences using `SentencePiece` (Kudo and Richardson, 2018) tokenizer to compute word error rates.

#### D.4.2 LIBRISPEECH CONFORMER WORKLOAD VARIANTS

The three workload variants for the CONFORMER model are:

- **GELU**: Activations functions were changed to GELU.
- **LayerNorm change** The layer normalization before the final readout layer was removed.
- **Attention Temp** The *output* of the attention softmax was multiplied by a temperature constant of 1.6. Note that this is different than other attention temperature setups where the temperature is multiplied in before the softmax.

In [Table 37](#) we present the results of our protocol to test our variants (described in [Section 6.2](#)). We observe that the ranks of hyperparameter transfers were not as high as variants in other models. In such cases we tested these models along an alternative protocol to ensure that the optimum learning rate is indeed different for these models.

#### D.4.3 DEEPSPEECH MODEL

We use a variant of the DEEPSPEECH (Amodei et al., 2016) model with residual connections, dropout (Srivastava et al., 2014), layer normalization, and SPECAUGMENT (Park et al., 2019) to improve performance. We use a convolution subsampling layer to reduce input dimensions by a factor of 4, which is further passed through 6 bi-directional LSTM layers and 3 feed-forward layers with a model internal dimension of 512 across layers. We use batch normalization inside the LSTM and feed-forward layers as post normalization layers. Model weights are initialized using Xavier uniform initialization. The CTC loss (Graves

Variant	Validation Performance	Hyperparameter Performance Transfer			
		BASE $\rightarrow$ VARIANT		VARIANT $\rightarrow$ BASE	
		Performance	Rank	Performance	Rank
BASE WORKLOAD	<b>0.077790</b>				
GELU	0.079354	0.079391	2	0.085503	3
LAYERNORM CHANGE	0.084175	0.084175	1	0.229800	1
ATTENTION TEMP	0.083092	0.083092	1	0.077790	1

Table 37: **Hyperparameter transfer between the base workload and the variants of Conformer on LibriSpeech.** We show the validation performance of the base workload compared to the performance achieved by the variants. Further, we show the performance (and hyperparameter ranking) when using the optimal hyperparameter point from the base workload on the variants (BASE  $\rightarrow$  VARIANT) and vice-versa (VARIANT  $\rightarrow$  BASE). All runs are from the same search space for NADAMW. See [Section 6.2](#) for a detailed description of our protocol for accepting workload variants.

[et al., 2006](#)) is used to train the model. See [here](#) for our implementation of DEEPSPEECH in Jax and [here](#) for our implementation of DEEPSPEECH in PyTorch. For experiments in this paper, we used a slightly different word piece tokenizer than CONFORMER, but we plan to update the tokenizer before the call for submissions to also use the same `SentencePiece` ([Kudo and Richardson, 2018](#)) tokenizer. We expect the DEEPSPEECH validation WER target to improve from 0.1162 to 0.111825.

#### D.4.4 LIBRISPEECH DEEPSPEECH WORKLOAD VARIANTS

The three workload variants for the DEEPSPEECH model are:

- **TanH:** Activation functions were changed to TanH.
- **No residual:** We removed the residual connections in the model. Interestingly this improved the overall performance of the model.
- **Norm & SpecAugment:** We removed the decoder layer normalization layer. We replaced all other batch normalization layers with layer normalization layers. We changed SPECAUGMENT specifications, specifically the FREQUENCY MASK from 2 to 4 and the TIME MASK from 10 to 15.

## D.5 OGBG

We use the OGBG-MOLPCBA dataset ([Hu et al., 2020](#)) containing molecular graphs and 128 molecular properties. The goal is to predict the properties given the graphs. The graphs contain on average 26 nodes and 28 edges. The train split contains 350,343 examples, and the train and validation splits both contain 43,793 examples.

Because the graphs are of varying sizes, we construct the batches using dynamic batching. We specify a maximum number of nodes and edges per batch and take graphs from the dataset into the batch until one of these thresholds is exceeded. We set these thresholds to

Variant	Validation Performance	Hyperparameter Performance Transfer			
		BASE $\rightarrow$ VARIANT		VARIANT $\rightarrow$ BASE	
		Performance	Rank	Performance	Rank
BASE WORKLOAD	0.113950				
TANH	0.131300	0.177558	32	0.122959	9
NO RESIDUAL	<b>0.105063</b>	0.135881	23	0.116508	2
NORM & SPECAUGMENT	0.130967	0.141442	6	0.117388	3

Table 38: **Hyperparameter transfer between the base workload and the variants of DeepSpeech on LibriSpeech.** We show the validation performance of the base workload compared to the performance achieved by the variants. Further, we show the performance (and hyperparameter ranking) when using the optimal hyperparameter point from the base workload on the variants (BASE  $\rightarrow$  VARIANT) and vice-versa (VARIANT  $\rightarrow$  BASE). All runs are from the same search space for NADAMW. See [Section 6.2](#) for a detailed description of our protocol for accepting workload variants.

BatchSize  $\cdot$  AvgNumNodes and  $2 \cdot$  BatchSize  $\cdot$  AvgNumEdges (the factor of 2 accounts for the bidirectional nature of the edges), respectively. See [here](#) for the implementation of our OGBG input pipeline.

#### D.5.1 GNN MODEL

The model is defined as a graph neural network (GNN) ([Battaglia et al., 2018](#)), which is a generalization of graph architectures such as GIN ([Xu et al., 2019](#)). We use the implementation provided by the JRAPH library ([Godwin et al., 2020](#)). The model first performs an embedding step which transforms the node and edge features with a linear embedding of size 256, and creates global features of size 128 initialized to zeros. Then the model performs 5 message passing steps following [Battaglia et al. \(2018, Algorithm 1\)](#) which update the node, edge, and global features. Each of the update functions is a 1-layer fully-connected network with a dense layer of size 256, layer normalization, ReLU, and dropout. The dropout rate can be tuned by the submissions but it defaults to 0.1. The weights for the model are initialized using LeCun normal initialization ([Klambauer et al., 2017](#)). The final output is read from the global features after the last step. See [here](#) for our implementation of GNN in Jax and [here](#) for our implementation of GNN in PyTorch.

#### D.5.2 OGBG GNN WORKLOAD VARIANTS

The three workload variants for the GNN model are:

- **GELU:** The activation was changed to GELU.
- **SiLU:** The activation was changed to SiLU.
- **Altered Layers:** An additional hidden layer of width 256 was added to each fully connected network, the latent dim was reduced to 128, and the number of message



passing steps was reduced to 3. Also, the layer normalization layers were swapped for batch normalization.

Variant	Validation Performance	Hyperparameter Performance Transfer			
		BASE $\rightarrow$ VARIANT		VARIANT $\rightarrow$ BASE	
		Performance	Rank	Performance	Rank
BASE WORKLOAD	0.280012				
GELU	0.284251	0.261344	25	0.267038	17
SiLU	<b>0.287569</b>	0.258375	28	0.275818	3
ALTERED LAYERS	0.269345	0.252153	15	0.244109	44

Table 39: **Hyperparameter transfer between the base workload and the variants of GNN on OGBG.** We show the validation performance of the base workload compared to the performance achieved by the variants. Further, we show the performance (and hyperparameter ranking) when using the optimal hyperparameter point from the base workload on the variants (BASE  $\rightarrow$  VARIANT) and vice-versa (VARIANT  $\rightarrow$  BASE). All runs are from the same search space for NADAMW. See [Section 6.2](#) for a detailed description of our protocol for accepting workload variants.

## D.6 WMT

The models are trained on the WMT 2017 German $\rightarrow$ English (De $\rightarrow$ En) training dataset ([Bojar et al., 2017](#)) which consists of about 5.9 million parallel sentence pairs. We don’t use any monolingual data or data augmentation methods. The models are evaluated on the ”validation” and ”test” splits of the WMT 2014 De $\rightarrow$ En dataset ([Bojar et al., 2014](#)) which consists of 3000 and 3003 sentence pairs respectively. We use `SentencePiece` ([Kudo and Richardson, 2018](#)) with 32k vocabulary size which is shared on source and target side. Sentences longer than 256 tokens on either the source or the target side are removed from the training data. See [here](#) for the implementation of our WMT input pipeline.

### D.6.1 TRANSFORMER MODEL

We use the TRANSFORMER-big architecture from [Vaswani et al. \(2017\)](#) with some modifications. More specifically, the model has 6 encoder and decoder layers each, model dimension 1024, hidden dimension 4096 and 16 attention heads. The embedding parameters are shared on the encoder and the decoder side. The same embedding matrix is also used for linear readout on the decoder (softmax). For regularization, the model uses label smoothing and dropout where the rates may be tuned by the submissions; the default label smoothing is 0.1 and the default dropout rate is 0.1. A batch size of 128 corresponds to about 280K tokens each on source and target side in one training batch. See [here](#) for our implementation of TRANSFORMER in Jax and [here](#) for our implementation of TRANSFORMER in PyTorch.

**Decoding** To generate samples from the model, we use beam search decoding with beam size 4, and a length penalty of 0.6. For evaluation, we use standard de-tokenized BLEU scores (Papineni et al., 2002) using the sacreBLEU library (Post, 2018).<sup>19</sup>

#### D.6.2 WMT TRANSFORMER WORKLOAD VARIANTS

The three workload variants for the WMT model are:

- **Post-LN:** Layer normalization was applied after the residual branch was added back into the trunk.
- **Attention Temp:** The attention logits were multiplied by a temperature of 4.0, before applying the softmax.
- **GLU & TanH:** Gated Linear Units (GLU) (Dauphin et al., 2017) were used in the MLP blocks, and activation functions were changed to TanH.

Variant	Validation Performance	Hyperparameter Performance Transfer			
		BASE $\rightarrow$ VARIANT		VARIANT $\rightarrow$ BASE	
		Performance	Rank	Performance	Rank
BASE WORKLOAD	<b>30.8534</b>				
POST-LN	30.2011	11.1946	76	29.7474	23
ATTENTION TEMP	30.2643	0	74	29.4278	32
GLU & TANH	30.1595	0	82	29.7135	25

Table 40: **Hyperparameter transfer between the base workload and the variants of Transformer on WMT.** We show the validation performance of the base workload compared to the performance achieved by the variants. Further, we show the performance (and hyperparameter ranking) when using the optimal hyperparameter point from the base workload on the variants (BASE  $\rightarrow$  VARIANT) and vice-versa (VARIANT  $\rightarrow$  BASE). All runs are from the same search space for NADAMW. See Section 6.2 for a detailed description of our protocol for accepting workload variants.

## E. Preliminary Experiments for Randomized Workloads

In order to create randomized workloads based on one of the fixed workloads, we needed to write down a distribution over workload modifications that has the properties we want. Our initial plan was to design a relatively broad distribution over natural changes to the base workload, then draw held-out workload samples, rejecting ones that failed a “trainability” test, and then rejecting ones that were too close to the base workload. Then we would revise the distribution to hopefully reduce the rejection rates and produce a final, official randomized workload. We started this process for the GNN workload on OGBG before eventually abandoning it.

<sup>19</sup>. case.mixed + numrefs.1 + smooth.exp + tok.13a + sacreBLEU 1.3.1

The distribution we started with is defined over the GNN’s hidden dimensions, latent dimensions (node and edge embedding dimension), normalization layer type, activation function, number of message passing steps, and dropout probability<sup>20</sup>, see Table 41. All workload samples from this distribution fit into competition hardware GPU memory with a batch size of 512. The batch size is not part of the workload definition, so in principle we might generate held-out workloads that require adjusting the batch size to remain within memory constraints. In general, we also need to find the best-performing batch size for target setting in order to reach competitive targets.

GNN Parameters	Distribution
Hidden dimension 1	{128, 256, 512, 1024, 2048}
Hidden dimension 2	{None, 128, 256, 512, 1024, 2048}
Latent dimension	{128, 256, 512, 1024, 2048}
Normalization layer	{batch normalization, layer normalization}
Activation function	{ReLU, SiLU, Leaky ReLU}
$N$ message passing steps	{2, 3, 4, 5, 6, 7, 8}
Dropout probability	[0.0, 0.5]

Table 41: **Distribution used in a preliminary experiment for a potential randomized GNN OGBG workload.**

After sampling 60 held-out workload candidates from the distribution described by Table 41, we ran a target-setting procedure on each sample. In other words, for each sample we tuned NESTEROV using 100 hyperparameter points to find the hyperparameters that achieve the best mAP. We considered a sample able to achieve an acceptable result in the runtime budget if, after tuning, it reached an mAP no more than 10% worse than the original base workload mAP target. We found that roughly 30% of the samples were trainable by this definition.

To quickly check whether the surviving 16 trainable workload candidates were sufficiently different from the base workload, we ran the best-performing hyperparameter settings from the base workload on each candidate. Unfortunately, in pretty much every case except for one activation function change, the best-performing hyperparameter settings from the base workload achieved about the same (within re-training noise) mAP as the best we could find after tuning just on the workload sample.

We decided to abandon this process for constructing randomized workload distributions (and not extend it to the other base workloads) for a variety of reasons. First, it was far too costly. Every time we revised the distribution we were considering, we had to run our entire target-setting procedure on every single sample from the distribution (in the experiment described above, this amounted to running 6000 trials for just one version of the distribution). Our procedure above also ignored some of the technicalities around selecting batch sizes for target setting, and we had to be very careful to filter out samples (or write down distribu-

20. Our initial experiments on randomized workloads for OGBG included the dropout probability as part of the workload definition. However, in the competition, submitters are free to determine the dropout probability. Therefore, a final, official randomized workload distribution for our benchmark could not include the dropout probability.

tions) that didn't violate the memory constraints of the competition hardware. We also did not have the understanding of what kinds of changes would cause the training problem to actually change that would be necessary to quickly come up with interesting distributions over workloads. Instead, we decided to adopt a procedure of manually constructing variants one at a time (see [Section 6.2](#)).

## References

- Naman Agarwal, Rohan Anil, Elad Hazan, Tomer Koren, and Cyril Zhang. Disentangling Adaptive Gradient Methods from Learning Rates, 2020.
- Ehsan Amid, Rohan Anil, and Manfred Warmuth. LocoProp: Enhancing BackProp via Local loss optimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.
- Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse H. Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Y. Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Y. Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, and Zhenyao Zhu. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. In *International Conference on Machine Learning (ICML)*, 2016.
- Rohan Anil, Vineet Gupta, Tomer Koren, Kevin Regan, and Yoram Singer. Scalable second order optimization for deep learning, 2020.
- Setareh Ariafar, Justin Gilmer, Zachary Nado, Jasper Snoek, Rodolphe Jenatton, and George E. Dahl. Predicting the utility of search spaces for black-box optimization: a simple, budget-aware approach. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.
- Juhan Bae, Paul Vicol, Jeff Z. HaoChen, and Roger B. Grosse. Amortized proximal optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- Thomas Bartz-Beielerstein, Carola Doerr, Daan van den Berg, Jakob Bossek, Sowmya Chandrasekaran, Tome Eftimov, Andreas Fischbach, Pascal Kerschke, William La Cava, Manuel Lopez-Ibanez, Katherine M. Malan, Jason H. Moore, Boris Naujoks, Patryk Orzechowski, Vanessa Volz, Markus Wagner, and Thomas Weise. Benchmarking in Optimization: Best Practice and Open Issues, 2020.
- Peter Battaglia, Jessica Blake Chandler Hamrick, Victor Bapst, Alvaro Sanchez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andy Ballard, Justin Gilmer, George E. Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Jayne Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks, 2018.
- Irwan Bello, William Fedus, Xianzhi Du, Ekin Dogus Cubuk, Aravind Srinivas, Tsung-Yi Lin, Jonathon Shlens, and Barret Zoph. Revisiting ResNets: Improved Training and Scaling Strategies. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- Lucas Beyer, Xiaohua Zhai, and Alexander Kolesnikov. Better plain ViT baselines for ImageNet-1k, 2022.

- Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 2014.
- Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 2015.
- Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, 2017.
- Olivier Bousquet, Sylvain Gelly, Karol Kurach, Olivier Teytaud, and Damien Vincent. Critical Hyper-Parameters: No Random, No Cry, 2017.
- Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, et al. Symbolic Discovery of Optimization Algorithms, 2023.
- Dami Choi, Alexandre Passos, Christopher J. Shallue, and George E. Dahl. Faster Neural Network Training with Data Echoing, 2019a.
- Dami Choi, Christopher J. Shallue, Zachary Nado, Jaehoon Lee, Chris J. Maddison, and George E. Dahl. On Empirical Comparisons of Optimizers for Deep Learning, 2019b.
- Jeremy M. Cohen, Behrooz Ghorbani, Shankar Krishnan, Naman Agarwal, Sourabh Medapati, Michal Badura, Daniel Suo, David Cardoze, Zachary Nado, George E. Dahl, and Justin Gilmer. Adaptive Gradient Methods at the Edge of Stability, 2022.
- Cody Coleman, Deepak Narayanan, Daniel Kang, Tian Zhao, Jian Zhang, Luigi Nardi, Peter Bailis, Kunle Olukotun, Chris Ré, and Matei Zaharia. DAWNbench: An end-to-end deep learning benchmark and competition. In *ML System Workshop at Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. RandAugment: Practical data augmentation with no separate search. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- Felix Dangel, Frederik Kunstner, and Philipp Hennig. BackPACK: Packing more into Backprop. In *International Conference on Learning Representations (ICLR)*, 2020.
- Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language Modeling with Gated Convolutional Networks. In *International Conference on Machine Learning (ICML)*, 2017.

- Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin F. Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Patrick Collier, Alexey Gritsenko, Vighnesh Birodkar, Cristina Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetić, Dustin Tran, Thomas Kipf, Mario Lučić, Xiaohua Zhai, Daniel Keysers, Jeremiah Harmsen, and Neil Houlsby. Scaling Vision Transformers to 22 Billion Parameters, 2023.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL-HLT)*, 2019.
- Elizabeth D. Dolan and Jorge J. Moré. Benchmarking optimization software with performance profiles. *Mathematical Programming*, 2002.
- Xuanyi Dong and Yi Yang. NAS-Bench-201: Extending the Scope of Reproducible Neural Architecture Search. In *International Conference on Learning Representations (ICLR)*, 2020.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 2021.
- Vijay Prakash Dwivedi, Chaitanya K. Joshi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking Graph Neural Networks. In *Journal of Machine Learning Research*, 2023.
- Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, 1993.
- Stefan Elfving, Eiji Uchibe, and Kenji Doya. Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning. *Neural Networks*, 2018.
- Pengzhi Gao, Zhongjun He, Hua Wu, and Haifeng Wang. Bi-SimCut: A Simple Strategy for Boosting Neural Machine Translation. In *Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL-HLT)*, 2022.
- John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue. TIMIT Acoustic-Phonetic Continuous Speech Corpus. Web download, 1993.



- Justin Gilmer, Behrooz Ghorbani, Ankush Garg, Sneha Kudugunta, Behnam Neyshabur, David Cardoze, George E. Dahl, Zachary Nado, and Orhan Firat. A Loss Curvature Perspective on Training Instabilities of Deep Learning Models. In *International Conference on Learning Representations (ICLR)*, 2021.
- Justin Gilmer, Andrea Schioppa, and Jeremy Cohen. Intriguing Properties of Transformer Training Instabilities. *To Appear*, 2023.
- Varun Godbole, George E. Dahl, Justin Gilmer, Christopher J. Shallue, and Zachary Nado. Deep Learning Tuning Playbook, 2023. URL [http://github.com/google/tuning\\_playbook](http://github.com/google/tuning_playbook). Version 1.0.
- Jonathan Godwin, Thomas Keck, Peter Battaglia, Victor Bapst, Thomas Kipf, Yujia Li, Kimberly Stachenfeld, Petar Veličković, and Alvaro Sanchez-Gonzalez. Jraph: A library for graph neural networks in jax., 2020. URL <http://github.com/deepmind/jraph>.
- Donald Goldfarb, Yi Ren, and Achraf Bahamou. Practical Quasi-Newton Methods for Training Deep Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- Alex Graves, Santiago Fernandez, Faustino Gomez, and Jürgen Schmidhuber. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *International Conference on Machine Learning (ICML)*, 2006.
- Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented Transformer for Speech Recognition, 2020.
- Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned Stochastic Tensor Optimization . In *International Conference on Machine Learning (ICML)*, 2018.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016a.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity Mappings in Deep Residual Networks. In *Computer Vision – ECCV*, 2016b.
- Jonathan Heek, Anselm Levskaya, Avital Oliver, Marvin Ritter, Bertrand Rondepierre, Andreas Steiner, and Marc van Zee. Flax: A neural network library and ecosystem for JAX, 2023. URL <http://github.com/google/flax>.
- Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs), 2016.
- Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 2006.
- Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

- Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open Graph Benchmark: Datasets for Machine Learning on Graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning (ICML)*, 2015.
- Norman P. Jouppi, Doe Hyun Yoon, George Kurian, Sheng Li, Nishant Patil, James Laudon, Cliff Young, and David Patterson. A Domain-Specific Supercomputer for Training Deep Neural Networks. *Communications of the ACM*, 2020.
- Hyunjik Kim, George Papamakarios, and Andriy Mnih. The Lipschitz Constant of Self-Attention. In *International Conference on Machine Learning (ICML)*, 2021.
- Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2015.
- Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- Florian Knoll, Tullie Murrell, Anuroop Sriram, Nafissa Yakubova, Jure Zbontar, Michael Rabbat, Aaron Defazio, Matthew J. Muckley, Daniel K. Sodickson, C. Lawrence Zitnick, and Michael P. Recht. Advancing machine learning for MR image reconstruction with an open competition: Overview of the 2019 fastMRI challenge. *Magnetic Resonance in Medicine*, 2020.
- Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.
- Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.
- Criteo A. I. Lab. Criteo 1TB Click Logs dataset. Web download, 2014.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV*, 2014.
- Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the Variance of the Adaptive Learning Rate and Beyond. In *International Conference on Learning Representations (ICLR)*, 2020.
- James Lucas, Shengyang Sun, Richard S. Zemel, and Roger B. Grosse. Aggregated Momentum: Stability Through Passive Damping. In *International Conference on Learning Representations (ICLR)*, 2019.

- Jerry Ma and Dennis Yarats. On the Adequacy of Untuned Warmup for Adaptive Optimization. In *AAAI Conference on Artificial Intelligence*, 2021.
- Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Luke Zettlemoyer. Mega: Moving Average Equipped Gated Attention. In *International Conference on Learning Representations (ICLR)*, 2023.
- James Martens. Deep learning via Hessian-free optimization. In *International Conference on Machine Learning (ICML)*, 2010.
- James Martens and Roger B. Grosse. Optimizing Neural Networks with Kronecker-factored Approximate Curvature. In *International Conference on Machine Learning (ICML)*, 2015.
- Peter Mattson, Christine Cheng, Gregory Diamos, Cody Coleman, Paulius Micikevicius, David Patterson, Hanlin Tang, Gu-Yeon Wei, Peter Bailis, Victor Bittorf, David Brooks, Dehao Chen, Debo Dutta, Udit Gupta, Kim Hazelwood, Andy Hock, Xinyuan Huang, Daniel Kang, David Kanter, Naveen Kumar, Jeffery Liao, Deepak Narayanan, Tayo Oguntebi, Gennady Pekhimenko, Lillian Pentecost, Vijay Janapa Reddi, Taylor Robie, Tom St John, Carole-Jean Wu, Lingjie Xu, Cliff Young, and Matei Zaharia. MLPerf Training Benchmark. In *Proceedings of Machine Learning and Systems*, 2020.
- Luke Metz, Niru Maheswaranathan, Ruoxi Sun, C. Daniel Freeman, Ben Poole, and Jascha Sohl-Dickstein. Using a thousand optimization tasks to learn hyperparameter search strategies, 2020.
- Thomas Moreau, Mathurin Massias, Alexandre Gramfort, Pierre Ablin, Pierre-Antoine Bannier, Benjamin Charlier, Mathieu Dagr eou, Tom Dupr e la Tour, Ghislain Durif, Cassio F. Dantas, Quentin Klopfenstein, Johan Larsson, En Lai, Tanguy Lefort, Benoit Malr ezieux, Badr Moufad, Binh T. Nguyen, Alain Rakotomamonjy, Zaccharie Ramzi, Joseph Salmon, and Samuel Vaiter. Benchopt: Reproducible, efficient and collaborative optimization benchmarks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- Zachary Nado, Justin M. Gilmer, Christopher J. Shallue, Rohan Anil, and George E. Dahl. A Large Batch Optimizer Reality Check: Traditional, Generic Optimizers Suffice Across Batch Sizes, 2021.
- Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G. Azzolini, Dmytro Dzhulgakov, Andrey Malleovich, Ilia Cherniavskii, Yinghai Lu, Raghuraman Krishnamoorthi, Ansha Yu, Volodymyr Kondratenko, Stephanie Pereira, Xianjie Chen, Wenlin Chen, Vijay Rao, Bill Jia, Liang Xiong, and Misha Smelyanskiy. Deep Learning Recommendation Model for Personalization and Recommendation Systems, 2019.
- Jorge Nocedal and Stephen Wright. *Numerical Optimization*. Springer Science, 1999.

- NVIDIA. DLRM for PyTorch, 2023. URL [https://catalog.ngc.nvidia.com/orgs/nvidia/resources/dlrm\\_for\\_pytorch](https://catalog.ngc.nvidia.com/orgs/nvidia/resources/dlrm_for_pytorch).
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002.
- Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech*, 2019.
- Matt Post. A Call for Clarity in Reporting BLEU Scores. In *Conference on Machine Translation: Research Papers*, 2018.
- Patrick Putzky, Dimitrios Karkaloulos, Jonas Teuwen, Nikita Miriakov, Bart Bakker, Matthan Caan, and Max Welling. i-RIM applied to the fastMRI challenge, 2019.
- Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet Classifiers Generalize to ImageNet? In *International Conference on Machine Learning (ICML)*, 2019.
- Yi Ren and Donald Goldfarb. Tensor normal training for deep learning models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- Yi Ren, Achraf Bahamou, and Donald Goldfarb. Practical Quasi-Newton Methods for Training Deep Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 2015.
- Robin M. Schmidt, Frank Schneider, and Philipp Hennig. Descending through a Crowded Valley – Benchmarking Deep Learning Optimizers. In *International Conference on Machine Learning (ICML)*, 2021.
- Frank Schneider, Lukas Balles, and Philipp Hennig. DeepOBS: A Deep Learning Optimizer Benchmark Suite. In *International Conference on Learning Representations (ICLR)*, 2019.

- Christopher J. Shallue, Jaehoon Lee, Joseph Antognini, Jascha Sohl-Dickstein, Roy Frostig, and George E. Dahl. Measuring the Effects of Data Parallelism on Neural Network Training. *Journal of Machine Learning Research*, 2019.
- Prabhu Teja Sivaprasad, Florian Mai, Thijs Vogels, Martin Jaggi, and Francois Fleuret. Optimizer Benchmarking Needs to Account for Hyperparameter Tuning. In *International Conference on Machine Learning (ICML)*, 2020.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 2014.
- Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. *Transactions on Machine Learning Research*, 2022.
- Andreas Sterbenz. Using Google Cloud Machine Learning to predict clicks at scale, 2017. URL <https://cloud.google.com/blog/products/gcp/using-google-cloud-machine-learning-to-predict-clicks-at-scale>.
- Mingxing Tan and Quoc Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *International Conference on Machine Learning (ICML)*, 2019.
- Ran Tian and Ankur P. Parikh. Amos: An Adam-style Optimizer with Adaptive Weight Decay towards Model-Oriented Scale, 2022.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- Xu Wang, Huan Zhao, Lanning Wei, and Quanming Yao. Pooling Architecture Search for Graph Property Prediction in Open Graph Benchmark. Technical report, AutoGraph team, 2022.
- Yan Wang, Hao Zhang, Jing Yang, Ruixin Zhang, and Shouhong Ding. Technical Report for OGB Graph Property Prediction. Technical report, Tencent Youtu Lab, 2021.
- Ross Wightman, Hugo Touvron, and Hervé Jégou. ResNet strikes back: An improved training procedure in timm, 2021.
- Ashia C. Wilson, Rebecca Roelofs, Mitchell Stern, Nathan Srebro, and Benjamin Recht. The Marginal Value of Adaptive Gradient Methods in Machine Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: a benchmark for molecular machine learning. *Chemical science*, 2018.
- Xingyu Xie, Pan Zhou, Huan Li, Zhouchen Lin, and Shuicheng Yan. Adan: Adaptive Nesterov Momentum Algorithm for Faster Optimizing Deep Models, 2022.

- Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On Layer Normalization in the Transformer Architecture. In *International Conference on Machine Learning (ICML)*, 2020.
- Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How Powerful are Graph Neural Networks? In *International Conference on Learning Representations (ICLR)*, 2019.
- Greg Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- Yang You, Igor Gitman, and Boris Ginsburg. Large Batch Training of Convolutional Networks, 2017.
- Yang You, Jing Li, Shashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. In *International Conference on Learning Representations (ICLR)*, 2020.
- Sergey Zagoruyko and Nikos Komodakis. Wide Residual Networks, 2016.
- Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J. Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdal, Adriana Romero, Michael Rabbat, Pascal Vincent, Nafissa Yakubova, James Pinkerton, Duo Wang, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastMRI: An Open Dataset and Benchmarks for Accelerated MRI, 2018.
- Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations (ICLR)*, 2018.
- Yilong Zhao, Li Jiang, Mingyu Gao, Naifeng Jing, Chengyang Gu, Qidong Tang, Fangxin Liu, Tao Yang, and Xiaoyao Liang. RePAST: A ReRAM-based PIM Accelerator for Second-order Training of DNN, 2022.
- Juntang Zhuang, Tommy Tang, Yifan Ding, Sekhar C. Tatikonda, Nicha Dvornek, Xenophon Papademetris, and James Duncan. AdaBelief Optimizer: Adapting Stepsizes by the Belief in Observed Gradients. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- Hongyu Zhy, Mohamed Akrouf, Bojian Zheng, Andrew Pelegris, Amar Phanishayee, Bianca Schroeder, and Gennady Pekhimenko. TBD: Benchmarking and Analyzing Deep Neural Network Training. In *IEEE International Symposium on Workload Characterization (IISWC)*, 2018.