DataPerf: Benchmarking Data for Better ML

DataPerf Working Group

MLCommons

Abstract

We propose DataPerf, a benchmark suite for ML datasets and algorithms for working with datasets. Historically, ML research has focused primarily on models, and simply used the largest existing dataset for common ML tasks without considering the dataset's breadth, difficulty, and fidelity to the underlying problem. This underfocus on data has led to a range of issues, from data cascades in real applications, to saturation of existing dataset-driven benchmarks for model quality impeding research progress. In order to catalyze increased research focus on data quality and foster data excellence, we propose a suite of benchmarks that evaluate the quality of training and test data, and the algorithms for constructing or optimizing such datasets, such as core set selection or labeling error debugging, across a range of common ML tasks. We plan to leverage the DataPerf benchmarks through challenges and leaderboards supported by the MLCommons Association and invite participation in further defining the benchmark suite.

DRAFT

1 Introduction

1.1 Why Benchmark Data?

Machine learning (ML) research has focused more on creating better models than on creating better datasets. We have seen massive progress in ML model technology driven by datasets used as benchmarks to measure model performance. In this way, large public datasets such as ImageNet [Deng et al., 2009], Freebase [Bollacker et al., 2008] and SQuAD [Rajpurkar et al., 2016] have provided compasses for ML research. However, we have often eagerly adopted the largest existing dataset without considering the datasets' breadth, difficulty, and fidelity to the underlying problem because of our overwhelming focus on the model.

As a field, we need to invest more effort in data. As the models improve and transition from the lab to the wild, we are finding discrepancies between performance in the lab and in the wild leading to reduced accuracy, fairness and bias issues [Buolamwini and Gebru, 2018, Denton et al., 2020, Mehrabi et al., 2021], challenges in applications such as health [Wilkinson et al., 2020], data cascades [Sambasivan et al., 2021] and data reuse [Koch et al., 2021] – issues which are often caused not by the model itself but by the data used to train it.

To further illustrate the importance of datasets in enabling progress in ML, consider the case of automated question answering. The SQuAD dataset enabled rapid experimentation, and its leaderboard allowed crisp and rapid assessment of progress in question answering. As researchers made progress on this dataset, they discovered that it was missing some classes of questions, and subsequently, many new datasets were built: SQuAD 2.0 [Rajpurkar et al., 2018], QuAC [Choi et al., 2018], CoQA [Reddy et al., 2019], Natural Questions [Kwiatkowski et al., 2019]. Yatskar [2018] points out that none of these question answering datasets contain questions that would require generating an answer (as opposed to extracting it from a given text). As of this writing, automated

Figure 1: ML benchmark saturation compared to human performance (plot from [Kiela et al., 2021]).

systems are "above human-level performance" on all but the last dataset [Kiela et al., 2021]. Without continually evolving datasets, we cannot continue to evolve the models.

In recognition of these challenges, burgeoning research communities[1][2] have emerged that focus on the science and engineering of data for machine learning [Aroyo et al., 2021, Raji et al., 2021, Paullada et al., 2021, **?**]. We need to grow these communities and provide them with better tools for evolving the datasets on which the whole field depends if we want improvements in machine learning in the lab to reliably translate into machine learning solutions that work in the real world.

1.2 DataPerf: Benchmarks for Data

To understand the scope, quality, and limitations of datasets and accelerate subsequent improvements, we are introducing the DataPerf benchmark suite and a related set of leaderboards and challenges. DataPerf is a suite of benchmarks that measure the quality of training and test datasets for machine learning tasks and the quality of algorithms used to construct such datasets. Benchmarks are instruments for measuring aspects of quality with metrics. For example, in the systems community, there are benchmarks for everything from measuring CPU speed to quantifying a cell phone's battery life. Each benchmark evaluates a specific artifact-under-test, such as a dataset or algorithm, using a particular methodology and metric. For instance, the quality of a training dataset is evaluated by training a set of models on the dataset and then measuring the model's accuracy on gold standards. The benchmarks in DataPerf can be used stand-alone during research and/or development; we also plan to host a range of challenges and leaderboards supported by MLCommons.

1.3 DataPerf drives the ML data cycle

In conventional model-centric ML, the term "benchmark" is often used to mean a standard, fixed dataset used to measure the performance of different models and compare them. For example, ImageNet is a benchmark for models like ResNet. Figure 1 shows that such existing model benchmarks are saturating, i.e., the models are attaining perfect or "human-level" performance when evaluated on these benchmarks [Kiela et al., 2021].

This raises some profound questions for this methodology: 1) To what extent is ML research making real progress on the underlying capabilities or just overfitting to the benchmark datasets or suffering from data artifacts [e.g., Gururangan et al., 2018, Poliak et al., 2018, Tsuchiya, 2018, Ribeiro et al., 2018, Belinkov et al., 2019, Geva et al., 2019, Wallace et al., 2019]? And most importantly, 2) How do we evolve the benchmarks rapidly to push the frontier for ML research forward for making real progress on capabilities?

[1] https://www.eval.how/, https://sites.google.com/view/sedl-workshop, i.a.
[2] https://datacentricai.org/

DRAFT

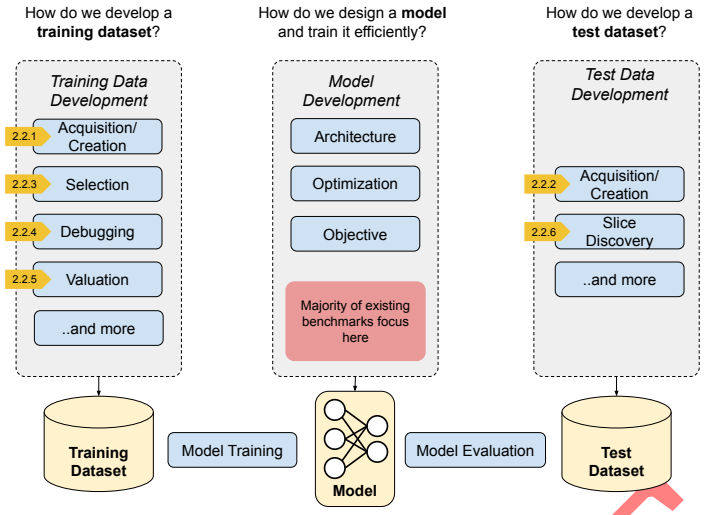

Figure 2: Design overview for DataPerf to illustrate the data-centric ML engineering process and the section references for the corresponding DataPerf benchmarks.

DRAFT

The DataPerf approach can be used to rapidly evolve the training and test data used to benchmark models, and consequently and concurrently, the ML models themselves. We believe future progress in ML will be driven by competition between rapidly evolving ML solutions that combine proprietary models and datasets. For a specific ML problem, multiple vendors will develop competing solutions. Cross-organizational groups will define test sets that serve as challenges for those solutions. As the ML solutions improve and the problem appears solved, the test sets will be made more complete and harder, thereby driving the development of even better solutions. We term this constructive competition the "data ratchet." DataPerf is designed to provide fast-moving data ratchets for the most critical ML problems, such as vision and speech-to-text. Toward this end, DataPerf will be used in challenges to develop training sets and test sets for the same problems so that the training sets can be used to evaluate the test sets and vice versa in an ongoing cycle.

2 DataPerf

We intend DataPerf to help address the lack of fundamental understanding of how to best engineer ML datasets and the absence of high productivity and efficient open data engineering tools in the machine learning community. DataPerf will make building, maintaining and evaluating datasets easier, cheaper, and more repeatable, regardless of which institution(s) you are affiliated with. Figure 2 shows how datasets and models come together during the development of an ML solution using a data-centric ML engineering process and provide section references for the detailed description of DataPerf benchmarks that address specific parts of the process.

DataPerf contains several types of benchmarks, as summarized in Table 1. Each of the benchmark types is described in more detail in the remainder of this section.

Table 1: Benchmark types in DataPerf.

Benchmark Type	Benchmark Methodology	Benchmark Metric
Training Dataset	Replace training set with novel training dataset	Accuracy of models trained on novel training dataset
Test Dataset	Select a fixed number of additional test data from the supplemental data set	Number of test datums submitted which are incorrectly labeled by models and correctly labeled by humans, where credit for each item is divided by number of submissions containing that item
Selection Algorithm	Replace training dataset with subset	Accuracy of models trained on subset
Debugging Algorithm	Identify labeling errors in version of training dataset with some corrupted labels	Accuracy of trained models after identified labels are corrected
Slicing Algorithm	Divide training dataset into semantically coherent slices	Fraction of data assigned to the correct slice
Valuation Algorithm	Estimate accuracy improvement from training on dataset A to training on dataset A + dataset B (where B is without labels at time of estimate)	Absolute difference between predicted accuracy and actual accuracy

DRAFT

2.1 Overview

We intend DataPerf to help address the lack of fundamental understanding of how to engineer ML datasets and the absence of high productivity and efficient open data engineering tools in the machine learning community to make building, maintaining and evaluating datasets easier, cheaper and more repeatable. Figure 1 shows how datasets and models come together during the development of an ML solution using a data-centric ML engineering process and provide section references for the detailed description of DataPerf benchmarks that address specific parts of the process.

2.2 Benchmark Types

2.2.1 Training Dataset

Purpose: Generating data, augmenting data, and other data-centric techniques can transform limited datasets into valuable training sets, but finding the right combination of methods can be a painstaking and error-prone process. Each form of data manipulation represents an entire sub-area in the literature and has many nuances. For example, even simple data augmentation techniques like vertical flip can have corner cases (e.g., "6" vs. "9") and may only apply to specific domains (e.g., in medical imaging, "a vertical flip of a mass would still result in a realistic mass [Hussain et al., 2017]"). Using multiple techniques quickly becomes an open-ended problem with a combinatorial set of solutions. The challenge is creating a pipeline of steps that expands a limited dataset into one representative of the real world. This type of benchmark aims to encourage our understanding of how to transform data into a valuable training set.

Measures: Training dataset benchmarks measure a novel training dataset by training various models and measuring the resulting accuracy. In most machine learning competitions, you are asked to build a high-performance model given a fixed dataset. Here we invert the traditional format and ask you to improve a dataset given a fixed model. Submitters are provided with a dataset to improve by applying data-centric techniques such as adding examples representing edge cases and using data augmentation.

Metric: The accuracy of a fixed model(s) trained on the submitted dataset and evaluated on a hidden test set.

2.2.2 Test dataset

Purpose: Conceptually, a Test Dataset benchmark measures a novel test dataset, or adversarial test data for inclusion in such a dataset, by evaluating if it is (1) labeled incorrectly by a variety of models, (2) labeled correctly by a humans, and (3) novel relative to other existing test data. The purpose of this type of benchmark is to foster innovation in the way we sample data for test sets and build an understanding of how data properties influence ML performance with respect to accuracy, reliability, fairness, diversity, and reproducibility.

Measures: Test dataset benchmarks measure a set of "adversarial" test data examples. Each participant is given a dataset of candidate data and asked to select a fixed number of "adversarial" examples. Adversarial examples are those for which a standard set of models fail to predict the label correctly. Participants will be allowed to employ any creative strategy, from human-curation to algorithmic data selection/prediction. Initially, only data from the target dataset will be allowed for submission and only for the target list of labels provided. New data, augmented data, or generative approaches will not be allowed.

Metric: The Test Dataset benchmark metric is the sum across the set of examples where human raters verify the label and one or more models fail to identify the label of each such example. The value is the fraction of models that fail to identify the label correctly. In the case of a challenge based on the benchmark, the value is divided by the number of submissions that include the particular example to reward rarer instances.

2.2.3 Selection algorithm

Purpose: Collecting large amounts of data has become straightforward, but turning that data into useful training sets has not. The unprecedented amount of available data in practice holds a wealth of information about the world around us. However, unlocking that value can be cumbersome and resource-hungry. Naively processing the data wastes valuable computational and labeling resources because the data is often redundant and heavily skewed. If we shift the focus to the long-tails, we can save valuable resources while maintaining quality. The challenge is algorithmically identifying and selecting the most informative examples from a dataset to use for training.

Measures: The selection algorithm benchmark evaluates the quality of algorithmic methods for curating datasets (e.g., active learning or core-set selection techniques). Selection algorithms identify the most informative examples from a large dataset. Each algorithm will select, target class, and submit the best possible training data of maximum size (e.g., 1,000 examples) from a dataset for a fixed set of target classes (e.g., 200 examples per class).

Metric: The submitted data will be evaluated by (1) training a fixed set of models, (2) evaluating the resulting models on held-out data, and (3) aggregating the quality scores across models and held-out datasets. The models are fixed and fine-tune pre-computed embeddings to keep the benchmark affordable and focused on data as the primary degree of freedom.

2.2.4 Data Debugging algorithm

Purpose: Training datasets often contain data errors. On one hand, these can be missing values (e.g. NULL values, or values that break functional dependencies such as City->ZIP Code). On the other hand, they can be wrong values (e.g. wrong labels, or noisy features undetectable by simple data validators). Repairing data errors is usually a costly process that often involves human labor. It is therefore useful to manage the tradeoff between the cost of repair and expected benefits. Given a fixed budget B representing the number (or proportion) of data examples that are allowed to be repaired, the challenge is to select a subset of training examples that will give the biggest improvement in the performance of the trained model.

Measures: DataPerf challenges provide a collection of problems, each defined by a set of attributes (e.g. source dataset, type of data error, error distribution across examples, model type etc). For each problem, we provide a labeled training dataset containing artificially injected data errors (following some combination of a random strategy, or rule based, or maybe even manual). If the error type

DRAFT

is missing values, for each missing value (designated by NULL), we provide a corresponding set of viable "candidate values", one of which is the "true" one. Also, this "erroneous dataset" is accompanied by a "repaired dataset" which is used for evaluation but should be hidden from the algorithm. Finally, a labeled validation dataset (used for tuning the algorithm) and a test dataset (used for the final evaluation; should be hidden from the algorithm) are provided.

Metric: After applying the data repairs within the constraints of the budget B, a model (specified by the appropriate problem attribute) is trained and its performance (e.g. accuracy) is evaluated. The same process is repeated for the initial erroneous dataset and the fully repaired dataset, which allows to establish the min and max bars of model performance (denoted perf_err and perf_rep). Finally, the score of the data debugging algorithm is expressed as the amount of "performance gap closed", that is, (perf_alg - perf_err) / (perf_rep - perf_err).

2.2.5 Valuation algorithm

Purpose: The purpose of this type of benchmark is to foster innovation in dataset creation through data acquisition. Conceptually, there is a data market, with data acquirers and data providers making transactions at a dataset granularity. Datasets can be labeled or unlabeled. For this challenge, we focus on an unlabeled dataset. This challenge aims to help data acquirers estimate the incremental value of an unlabeled dataset, prior to acquisition and labeling, against an existing labeled dataset.

We assume the data acquirer already has a local dataset (D_A), which it uses to train its local model. However, this model is not good enough and the acquirer wants to acquire more data examples to improve their model. There is a data provider with an unlabeled dataset (D_B). The data acquirer needs a lightweight algorithm to estimate the value of the entire dataset before acquisition and labeling. We assume data acquirers and providers are connected through a trusted, auditable computation infrastructure, which has access to providers' and acquirers' data.

Measures: The DataPerf Valuation benchmark measures the quality of an algorithm that estimates the relative value of using a new dataset D_B along with an existing dataset D_A with respect to a test set D_{test} and $Model_M$, where D_B does not have labels at the time of the estimate. Participants will be provided the labeled dataset D_A and unlabeled dataset D_B, a $Model_M$, and a test set D_{test} to evaluate their algorithm. Participants will be allowed to employ any creative strategy possible, from algorithmic labeling of the unlabeled dataset to unsupervised learning. However, introduction of new data, trained models, or manual labeling of the dataset is not allowed.

Metric: The base metric is the absolute difference between estimated accuracy $Acc'(D_A + D_B)$ to the true accuracy $Acc(D_A + D_B)$ of a model trained on the union of the two datasets. The true accuracy will be calculated with the true labels during the evaluation of the participants' algorithms. For more thorough evaluation, the full metric is the root mean squared error across a range of such problems.

2.2.6 Slice discovery algorithm

Purpose: Machine learning models that achieve high overall accuracy often make systematic errors on important subgroups (or slices) of data. For instance, models trained to detect collapsed lungs in chest X-rays have been shown to make predictions based on the presence of chest drains, a device typically used during treatment. As a result, these models frequently make prediction errors on cases without chest drains, a critical data slice where false negative predictions could be life-threatening. Identifying underperforming slices is challenging when working with high-dimensional inputs (e.g. images, audio) where data slices are often unlabeled. This benchmark evaluates automated slice discovery algorithms that mine unstructured data for underperforming slices.

Measures: The DataPerf Slice Discovery benchmark consists of a large set of slice discovery problems. Each problem includes (1) a labeled dataset, (2) a model trained on that dataset, and (3) ground truth slice annotations for one or more slices on which the model makes systematic errors. For example, one slice discovery problem in the benchmark includes a dataset of images with binary labels for the class "vehicle", a model trained to detect vehicles, and labels for a critical data slice, "mopeds", on which the model underperforms. Participants are asked to submit algorithms that given the (1) dataset and (2) model can automatically identify the underperforming (3) slice. An algorithm may output up to five slices, each specified as ranking over the examples in the dataset.

DRAFT

Metric: When using slice discovery algorithms, practitioners typically inspect the top-k examples in a predicted slice to understand what concept the slice corresponds. Our primary metric, the maximum precision-at-k between a predicted slice and the ground truth slice, aligns well with this use case. Maximum precision-at-k is computed for each problem in the benchmark. Each slice discovery algorithm gets a single score for comparisons and leaderboards: the average maximum precision-at-k across the problems.

2.3 Suite: Benchmark Types x Tasks

We define the DataPerf Benchmark Suite as a cross-product of benchmark types and specific ML tasks to measure the datasets or algorithms under test. Specifically, the initial version of DataPerf uses the following tasks:

Image Classification: The Image Classification task involves labeling images from Open Images. Open Images was chosen because of its large size and permissive license.

Roman Numeral OCR: The Roman Numeral OCR task identifies hand-written roman numerals. Roman numerals were chosen because the limited domain makes it easy to generate novel data, and there is no major existing dataset that could be submitted to the training set benchmark.

Keyword Identification: The keyword identification task involves labeling short speech utterances with the spoken work.

NLP: The Natural Language Processing task encompasses five subtasks from the Dynabench Platform [Kiela et al., 2021], including hate speech detection [Vidgen et al., 2021], natural language inference [Nie et al., 2020], question answering [Bartolo et al., 2020], sentiment analysis [Potts et al., 2021], and visual question answering [Sheng et al., 2021].

We use this benchmark matrix for three reasons. First, each column embodies a data ratchet for a specific problem in the form of a training set benchmark and a test set benchmark. Second, the same algorithm can be submitted to all benchmarks in the same row for algorithmic benchmarks to demonstrate generality. Third, pragmatically, rules and infrastructure developed to support one benchmark may be leveraged for other challenges. Fourth, the Dynabench platform has made initial strides towards improving leaderboard culture, with leaderboard infrastructure that can be dynamically expanded or adjusted.

DRAFT

2.4 Competition: Leaderboards + Challenges

We plan leaderboards and challenges based on the DataPerf benchmarks to encourage constructive competition, identify best-of-breed ideas, and inspire the next generation of concepts for building and optimizing datasets. A leaderboard is a public summary of benchmark results. Leaderboards help to identify state-of-the-art approaches quickly. A challenge is a public contest to achieve the best result(s) on a leaderboard in a fixed period of time. Challenges motivate rapid progress in developing new approaches through recognition, awards, and/or prizes. We are interested in benchmarks related to the quality of the datasets and the algorithms for working with datasets. We will host the leaderboard and challenges on an augmented Dynabench system supported by the MLCommons Association.

3 Members of DataPerf

The remainder of this section describes how DataPerf builds upon prior benchmarks, challenges, tracks, workshops, and organizations in more detail. DataPerf adopts inspiration and several of the best practices from these prior works. We welcome members of the community to join us!

3.1 CATS4ML

The Crowdsourcing Adverse Test Sets for Machine Learning (CATS4ML) Data Challenge [Aroyo et al.] aims to raise the bar for ML evaluation sets and to find as many examples as possible that are confusing or otherwise problematic for algorithms to process, starting with image classification. Many evaluation datasets contain items that are easy to evaluate, e.g., photos with a subject that is easy to identify. Thus they miss the natural ambiguity of real-world context. The absence of

ambiguous real-world examples in evaluation undermines the ability to test machine learning performance reliably, which makes ML models prone to developing "weak spots." CATS4ML relies on human skills and intuition to spot new data examples about which machine learning is confident but misclassified. An open CATS4ML challenge that asks participants to submit misclassified samples from the Google Open Images dataset was unveiled at HCOMP 2020[3].

3.2 DCAI competition

The Data-Centric AI competition inverts the traditional format of most machine learning competitions. You are asked to build a high-performance model given a fixed dataset and instead ask you to improve a dataset given a fixed model. Submitters are provided with a dataset to improve by applying data-centric techniques such as fixing incorrect labels, adding examples representing edge cases, using data augmentation, etc. The competition is inspired by MNIST and focuses on the classification of Roman numeral digits. Before launching the contest, 15 machine learning engineers attempted to achieve a high score in under 1 hour. Their best submission increased the baseline accuracy from 65% to 69%. The first competition received over 1,000 submissions, and the best result was 85% accuracy. The DCAI benchmark forms the basis for our data selection algorithm in DataPerf.

3.3 DCBench

DCBench[4] is a benchmark for algorithms used to construct and analyze machine learning datasets. It comprises a diverse set of tasks, each corresponding to one step in a broader machine learning pipeline. For example, machine learning practitioners will commonly spend some time cleaning input features, so DCBench includes a task for algorithms that select training data points for cleaning. Each task is instantiated as a collection of problems with varied specifications (e.g., dataset, model architecture). A problem consists of artifacts from steps upstream in the pipeline and specifications for running downstream steps. These artifacts and specifications are shared across evaluations, ensuring that algorithms are evaluated independently of the rest of the machine learning pipeline. The tasks in DCBench are supported by a standard Python API that facilitates downloading problems, running evaluations, and comparing methods. DCBench provides the basis for the Debugging Algorithm and Slicing Algorithm benchmarks used in DataPerf.

DRAFT

3.4 Dynabench

Dynabench[5] is a research platform for dynamic data collection and benchmarking that challenges existing ML benchmarking dogma by embracing dynamic dataset generation [Kiela et al., 2021]. Benchmarks for machine learning solutions based on static datasets have well-known issues: they saturate quickly, are susceptible to overfitting, contain exploitable annotator artifacts and have unclear or imperfect evaluation metrics. In essence, the Dynabench platform is a scientific experiment: is it possible to make faster progress if data is collected dynamically, with humans and models in the loop, rather than in the old-fashioned static way? DataPerf adopts DynaBench as the underlying platform that facilitates our leaderboards and challenges, as described previously.

3.5 NeurIPS Datasets and Benchmarks

NeurIPS 2021 launched the new Datasets and Benchmarks track [6] as a novel venue for exceptional work in creating high-quality datasets, insightful benchmarks, and discussions on how to improve dataset development and data-oriented work more broadly. Datasets and benchmarks are crucial for the development of machine learning methods but also require their own publishing and reviewing guidelines, such as proper descriptions of how the data was collected, whether they show intrinsic bias, and whether they will remain accessible. For benchmarks, reproducibility, interpretability and design are key factors that require careful consideration. One goal of DataPerf is to inspire and enable evaluation of rapidly evolving submissions to this track.

[3] https://ai.googleblog.com/2021/02/uncovering-unknown-unknowns-in-machine.html
[4] https://github.com/data-centric-ai/dcbench
[5] https://dynabench.org/
[6] https://blog.neurips.cc/2021/04/07/announcing-the-neurips-2021-datasets-and-benchmarks-track/

3.6 MLCommons and MLPerf

The MLCommons[7] association is a non-profit organization supported by 50+ member companies and academics, focused on making ML better for everyone through benchmarks, open data, and best practices. The MLCommons association hosts the MLPerf benchmark suites [Mattson et al., 2020, Reddi et al., 2020], which emerged from a collaboration between academia and industry to answer the need for fair and representative benchmarks that enable "apples to apples" comparisons of ML systems. The MLPerf benchmark suites define clear rules for measuring speed and power consumption across various systems, spanning from datacenter scale ML systems that consume megawatts of power to tiny embedded ML systems that consume only microwatts of power. The MLPerf benchmark suites jump-started a virtuous cycle of healthy and transparent competition that drove rapid improvements in ML performance. The MLCommons Association will host a working group to evolve the benchmarks and support the leaderboards and challenges.

4 Conclusion

DataPerf aims to improve ML for everyone by benchmarking datasets. Benchmarking is essential as what gets measured gets improved. It systematizes quality measurement of training and test datasets across a range of ML use cases. It also enables us to improve the quality of the algorithms used to construct such datasets. DataPerf invites the ML community to tackle practical data problems through leaderboards and challenges. It is a community-driven initiative, and it serves as a call for action to improve best practices for the creation of ML datasets. Join DataPerf to help shape the future of AI and ML.

DRAFT

References

L. Aroyo, P. Paritosh, S. Ibtasam, D. Bansal, K. Rong, and K. Wong. Adversarial test set for image classification: Lessons learned from cats4ml data challenge.

L. Aroyo, M. Lease, P. Paritosh, and M. Schaekermann. Data excellence for ai: Why should you care. *arXiv preprint arXiv:2111.10391*, 2021.

M. Bartolo, A. Roberts, J. Welbl, S. Riedel, and P. Stenetorp. Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678, 11 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00338. URL `https://doi.org/10.1162/tacl_a_00338`.

Y. Belinkov, A. Poliak, S. Shieber, B. Van Durme, and A. Rush. Don't take the premise for granted: Mitigating artifacts in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 877–891, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1084. URL `https://aclanthology.org/P19-1084`.

K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, 2008.

J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. Proceedings of Machine Learning Research, 2018.

E. Choi, H. He, M. Iyyer, M. Yatskar, W.-t. Yih, Y. Choi, P. Liang, and L. Zettlemoyer. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*, 2018.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] https://www.mlcommons.org/

E. Denton, A. Hanna, R. Amironesei, A. Smart, H. Nicole, and M. K. Scheuerman. Bringing the people back in: Contesting benchmark machine learning datasets. abs/2007.07399, 2020. URL https://arxiv.org/abs/2007.07399.

M. Geva, Y. Goldberg, and J. Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv:1908.07898*, 2019.

S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. Bowman, and N. A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017. URL https://aclanthology.org/N18-2017.

Z. Hussain, F. Gimenez, D. Yi, and D. Rubin. Differential data augmentation techniques for medical imaging classification tasks. In *AMIA annual symposium proceedings*, volume 2017, page 979. American Medical Informatics Association, 2017.

D. Kiela, M. Bartolo, Y. Nie, D. Kaushik, A. Geiger, Z. Wu, B. Vidgen, G. Prasad, A. Singh, P. Ringshia, Z. Ma, T. Thrush, S. Riedel, Z. Waseem, P. Stenetorp, R. Jia, M. Bansal, C. Potts, and A. Williams. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324.

B. Koch, E. Denton, A. Hanna, and J. G. Foster. Reduced, reused and recycled: The life of a dataset in machine learning research, 2021.

T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

P. Mattson, C. Cheng, G. Diamos, C. Coleman, P. Micikevicius, D. Patterson, H. Tang, G.-Y. Wei, P. Bailis, V. Bittorf, D. Brooks, D. Chen, D. Dutta, U. Gupta, K. Hazelwood, A. Hock, X. Huang, D. Kang, D. Kanter, N. Kumar, J. Liao, D. Narayanan, T. Oguntebi, G. Pekhimenko, L. Pentecost, V. Janapa Reddi, T. Robie, T. St John, C.-J. Wu, L. Xu, C. Young, and M. Zaharia. Mlperf training benchmark. In *Proceedings of Machine Learning and Systems*, volume 2, 2020.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL https://aclanthology.org/2020.acl-main.441.

A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2021.100336. URL https://www.sciencedirect.com/science/article/pii/S2666389921001847.

A. Poliak, J. Naradowsky, A. Haldar, R. Rudinger, and B. Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2023. URL https://aclanthology.org/S18-2023.

C. Potts, Z. Wu, A. Geiger, and D. Kiela. DynaSent: A dynamic benchmark for sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long*

DRAFT

Papers), pages 2388–2404, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.186. URL https://aclanthology.org/2021.acl-long.186.

I. D. Raji, E. M. Bender, A. Paullada, E. Denton, and A. Hanna. AI and the everything in the whole wide world benchmark. abs/2111.15366, 2021. URL https://arxiv.org/abs/2111.15366.

P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

P. Rajpurkar, R. Jia, and P. Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.

V. J. Reddi, C. Cheng, D. Kanter, P. Mattson, G. Schmuelling, C.-J. Wu, B. Anderson, M. Breughe, M. Charlebois, W. Chou, R. Chukka, C. Coleman, S. Davis, P. Deng, G. Diamos, J. Duke, D. Fick, J. S. Gardner, I. Hubara, S. Idgunji, T. B. Jablin, J. Jiao, T. S. John, P. Kanwar, D. Lee, J. Liao, A. Lokhmotov, F. Massa, P. Meng, P. Micikevicius, C. Osborne, G. Pekhimenko, A. T. R. Rajan, D. Sequeira, A. Sirasao, F. Sun, H. Tang, M. Thomson, F. Wei, E. Wu, L. Xu, K. Yamada, B. Yu, G. Yuan, A. Zhong, P. Zhang, and Y. Zhou. Mlperf inference benchmark. In *Proceedings of the ACM/IEEE Annual International Symposium on Computer Architecture*, 2020.

S. Reddy, D. Chen, and C. D. Manning. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.

M. T. Ribeiro, S. Singh, and C. Guestrin. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1079. URL https://aclanthology.org/P18-1079.

N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

S. Sheng, A. Singh, V. Goswami, J. A. L. Magana, T. Thrush, W. Galuba, D. Parikh, and D. Kiela. Human-adversarial visual question answering. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

M. Tsuchiya. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1239.

B. Vidgen, T. Thrush, Z. Waseem, and D. Kiela. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.132. URL https://aclanthology.org/2021.acl-long.132.

E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1221. URL https://aclanthology.org/D19-1221.

J. Wilkinson, K. F. Arnold, E. J. Murray, M. van Smeden, K. Carr, R. Sippy, M. de Kamps, A. Beam, S. Konigorski, C. Lippert, et al. Time to reality check the promises of machine learning-powered precision medicine. *The Lancet Digital Health*, 2020.

M. Yatskar. A qualitative comparison of coqa, squad 2.0 and quac. *arXiv preprint arXiv:1809.10735*, 2018.

DRAFT