A Data-Centric Image Classification Benchmark

Lars Schmarje *
Multimedia Information Processing Group
University of Kiel
las@informatik.uni-kiel.de

Yuan-Hong Liao
University of Toronto, Vector Institute
andrew@cs.toronto.edu

Reinhard Koch
Multimedia Information Processing Group
University of Kiel
rk@informatik.uni-kiel.de

Abstract

High-quality labeled datasets are critical to the advances in machine learning and tend to benefit all kinds of model-centric algorithms, such as novel architectures and loss functions. The labeling process is usually label-intensive and time-consuming since it includes many turns of data selection, data cleaning, and data analysis. There are tons of work that aim to solve each specific step, but it lacks an understanding of how to combine them and, most importantly, a standard testbed for different dataset improving techniques. We, therefore, present the concept of a multi-domain benchmark for *acquiring consistent labels* with limited budgets. In contrast to most benchmarks that encourage novel model-centric algorithms, our multi-domain data-centric benchmark encourages algorithms to improve the provided dataset. The proposed benchmark consists of different resolutions, class distributions and domains ranging from biological to medical domains.

1 Introduction

Various benchmarks are used to gauge the advances in machine learning. Most benchmarks usually introduce a set of new tasks and a corresponding fixed dataset, such as ImageNet [11]. Machine learning researchers are encouraged to compete by devising novel algorithms on model architectures, loss functions, optimization techniques, etc., while discouraged to improve the data, such as relabeling the existing dataset or utilizing other in-domain datasets. Given that label errors are prevalent in machine learning datasets [25], this constraint can lead to overfitting the label errors in the train set. Recent work [3, 43] echo the importance of the data quality by showing that almost every model architecture improves after cleaning the train set labels.

There are many ways to improve datasets. Many data cleaning algorithms [1, 7, 14, 6] have been translated into tools and possibly repair errors. Removing biased [40, 23, 23, 39] or sprious correlations [38] from train set is particularly important in safety-critical applications, such as health care system. Active learning aims to sample new labeled data from the unlabeled pool [28, 12, 36] The advances in semi-supervised learning show that adding in-domain unlabeled data increases model performances [37] and model robustness [5]. While there are many different ways to improve datasets, it still lacks a systematic way to combine them, and most importantly, a standard testbed for different dataset improving techniques.

Most dataset improving techniques acquire additional human annotations, hampering direct comparisons between different approaches. To provide a reproducible and reliable benchmark for different

*Corresponding Author

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

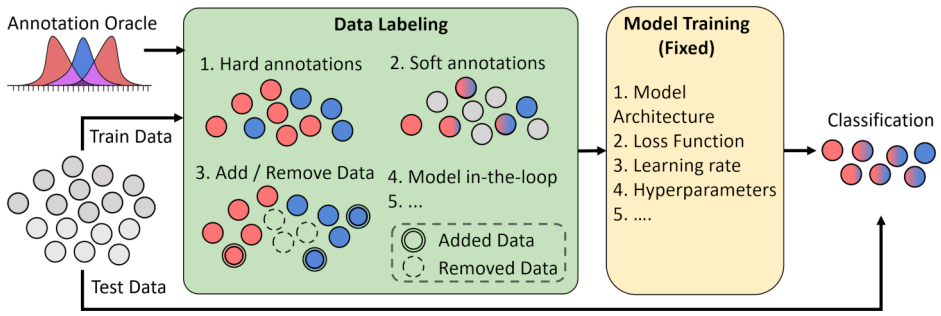


Figure 1: Illustration of data-centric Benchmark – The input for the benchmark a set of unlabeled images (grey dots) which are separated into training and test data. The unlabeled training data is labeled (green block) and a classifier is trained on the labeled data (yellow block). The target challenge is a classification of the test data with the trained model. Common benchmarks [18, 8, 17] focus on the model training and expect the labeled data to be given. In our benchmark, we fix the used model and focus on better data labeling e.g. by roughly labeling all samples, by detailed labeling of few samples or removing (dashed circles) or adding training data (double stroked circles).

dataset improving approaches, we need to include the worker simulations. Worker simulation has been widely discussed in many different domains, such as emergency and evacuation simulations [2], competitive crowd marketplaces [29], team performance [30], etc. In this work, we take the first step by focusing on image classification.

A major reason for label errors in image classification is the intra- and interobserver variability [27, 9, 26, 16, 31, 4, 10, 32, 15]. Following previous work [21], we simulate the worker responses from the underlying label distribution. Specifically, we model the underlying labels by assuming that every image has an unknown soft probability distribution $l \in [0, 1]^k$ for classification task over k classes. The assumption is motivated by two reasons. Firstly, the mislabeled images exist due to annotators' subjective opinions, e.g., the grade of an illness. A hard label $l \in \{0, 1\}^k$ could not model such a difference over the complete annotator population. Secondly, if we look for examples of biological processes, there are images of intermediate transition stages between two classes, e.g., the degeneration of living plankton to dead biomass.

The literature describes a variety of method to solve the classification task of images. Many of these algorithms are non-data-centric and thus either expect high quality hard labels [37] or consider some image as falsely classified [20]. These algorithms do not consider the unknown soft label distribution and thus ignore valuable information in the data. A consensus processes can be used to get more consistent labels but this requires many annotations and thus is often not viable on a large scale. Data-centric algorithms like [31, 32] consider the soft label distribution to create proposals for more consistent annotations e.g. by clustering visual similar images and therefore can solve the task most likely with better and fewer labels. However, no general testbed exists to compare these algorithms which hinders the research in this area.

As shown in Figure 1, we present the concept of a multi-domain benchmark for *acquiring consistent labels* with limited budgets. In contrast to most benchmarks, our benchmark includes the worker annotations simulations, providing participants noisy annotations at a fixed cost. The benchmark expects a fixed number of images and labels from the data-centric algorithms. The performance is determined by the test set performance of the neural network trained on the provided labels.

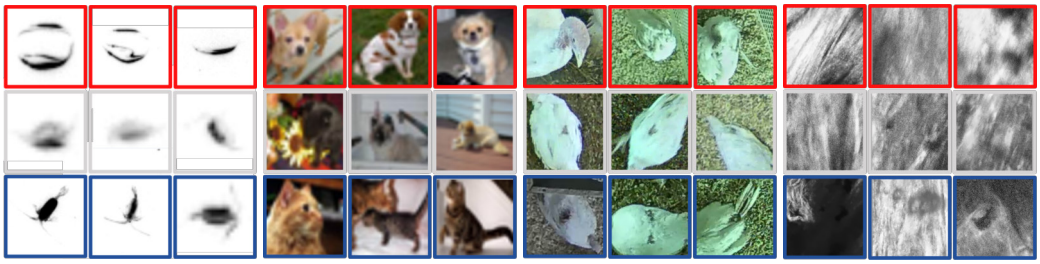
2 Benchmark

The main goal of our benchmark is acquiring consistent labels with as few annotations as possible. The benchmark can be used to compare a wide variety of data-centric methods. With the amount of annotation being fixed, the participants need to decide which image to annotate and how much they trust each annotation.

In general, the described task is a two dimensional problem with regard to the quality labels and the amount of annotations.A two-dimensional optimization is difficult to evaluate across different

Table 1: Overview of possible usable dataset – # is an abbreviation for number. The class imbalance is given as the percentage of the smallest and largest class with regard to the complete dataset. n is the average of annotations per image

Name	# classes	Input size $[px]$	# Images	Class Imbalance $[\%]$		n
				Smallest	Largest	
Plankton [31]	10	96x96	12280	4.16	30.37	24
Turkey [41]	2	96x96	8040	9.66	90.33	3
Mice Bone [34]	3	224x224	724	10.81	63.98	3
CIFAR10-H [27]	10	32x32	10000	9.88	10.16	51



(a) Plankton [31] (b) CIFAR10-H [27] (c) Turkey [41] (d) Mice Bones [33]

Figure 2: Example images for possible datasets – The red and blue rows are examples of two different with a high agreement between the annotators. The grey center row show images with a high annotation variability between the red and blue class. Images taken from [32] with authors permission.

methods and thus we decided to simplify the task by introducing an annotation budget b. An annotation budget $b \in \mathbb{R}_{>0}$ describes how many annotations can be used relative to the size of the dataset.For example a budget of 0.5 for a dataset with 1000 images means that we can use 500 annotations. However, we can decide how we distribute the annotations between the images e.g. 5 images with 100 annotations each or 100 images with 5 annotations each. An annotation $a \in \{0,1\}^k$ is a hard approximation of l from a human or an oracle in an automatic environment. The annotations can be automatically be acquired from an annotation oracle which is described in subsection 2.1.

The benchmark challenge is to relabel the data with a given budget b so that an given model Φ could be trained more successfully. The usage of one or few annotations will be insufficient to approximate the unknown soft probability distribution l in all cases. Thus the participants of the benchmark have to decide how to use their budget. Possible decisions could be the following: labeling many images roughly, labeling few images in detail, add or remove data points. A main goal is to identify and relabel poor approximation of l which are misleading during the training of Φ and thus improve the final classification [32, 43]. We can evaluate the consistency of the generated labels with the κ metric [22] and the prediction accuracy of Φ on the test set. We will describe suitable datasets the benchmark in subsection 2.2.

2.1 Annotation Oracle

For the systematic evaluation of our benchmark, we need ideally the soft label distribution l for every image and humans which provide new annotations during the relabeling. However, the soft label distribution is unknown and including humans in an automatic benchmark is not feasible. Thus, we need to approximate l and provide an annotation oracle based on the approximation of l.

If we have n annotations $a_1, ..., a_n \in \{0,1\}^k$ per image, we can approximate l with the average of the annotations $\hat{l} = \frac{1}{n} \sum_{i \in \{1,..,n\}} a_i$. We assume that for $n \longrightarrow \infty$ the approximation $\hat{l}$ is identical to l because it represent the consensus of different annotations. The annotation oracle models the human annotations by sampling from the approximated underlying label distributions $\hat{l}$. We might

need add some simple thresholds to model elements like confirmation bias if the annotation should also consider a given proposal.

We are aware that this annotation oracle is a simplification of the reality and discuss possible benefits and issues in section 3.

2.2 Datasets

The main requirement for using a dataset in the benchmark is that we have n annotations per image with n as high as possible to approximate l better. We focus on 2D image classification with hundreds to thousands of images because otherwise the required annotations would not be obtainable. Preferable we would use already existing published [27] or unpublished [32, 4, 41, 24, 35] datasets with multiple annotations. An non-extensive overview about possible datasets is given in Table 1 and some example images are given in Figure 2. Overall, we aim at 5 to 10 different datasets with a wide variety. The variety should include the domains, resolutions, image modalities, number of classes, number of approximations and class distributions.

3 Discussion

In this part, we will discuss the beneficial effects, limitations and open questions of the our benchmark.

Evaluation – The general image classification task is very flexible and can be a proxy task for many use cases depending on the evaluation metric. For example we can use the Kullback-Leiber divergence [19] to measure the similarity between the predicted and the approximated ground-truth distribution $\hat{l}$ instead of classification score like accuracy or F1-Score . This score would focus more on understanding the underlying distribution than to focus on possibly overconfident predictions [13] and could improve even the classification performance [42].

Simulation – A limitation of this benchmark is the simulation of human annotations. We can only approximate the unknown distribution l with an average over multiple annotations. The annotation oracle uses random samples from our approximated distribution $\hat{l}$ as annotations and thus do not consider the difference between annotators. These simplifications could lead to a simulation of humans which is incorrect and thus would lead to ill-defined evaluations. However, this assumption is often used in the literature [32, 31] which makes us confident that the oracle can simulate the reality sufficiently.

Dataset Size – Moreover, we can not consider datasets with million of images because multiple annotations for approximated distribution $\hat{l}$ would not be feasible.

Some open questions remain which need to be experimentally verified or be discussed in the community. It is currently unclear how much improvement can be gained from only tuning the data with a fixed model for training. It might be of interest to allow more recent models for training over time but this would limit the comparability. If we evaluate on fixed test set we need to ensure perfect labels. However, as we argued above we can only approximate the unknown ground-truth distribution and this will make mistakes which leads to misleading results.

Conclusion – Overall, we believe the benchmark to beneficial for data-centric research as it considers the complete label distribution and does not ignore several aspects. Some restrictions have to be considered but the potential gain for machine learning research in general is invalueable.

Acknowledgements

We acknowledge funding of L. Schmarje by the ARTEMIS project (Grant number 01EC1908E) funded by the Federal Ministry of Education and Research (BMBF, Germany).

References

[1] Abedjan, Z., Chu, X., Deng, D., Fernandez, R.C., Ilyas, I.F., Ouzzani, M., Papotti, P., Stonebraker, M., Tang, N.: Detecting data errors: Where are we and what needs to be done? Proc. VLDB Endow. **9**(12), 993–1004 (Aug 2016). https://doi.org/10.14778/2994509.2994518, https://doi.org/10.14778/2994509.2994518

[2] Almeida, J.E., Rosseti, R.J.F., Coelho, A.L.: Crowd simulation modeling applied to emergency and evacuation simulations using multi-agent systems (2013)

[3] Beyer, L., Hénaff, O.J., Kolesnikov, A., Zhai, X., van den Oord, A.: Are we done with imagenet? (2020)

[4] Brünger, J., Dippel, S., Koch, R., Veit, C.: 'Tailception': using neural networks for assessing tail lesions on pictures of pig carcasses. Animal **13**(5), 1030–1036 (2019)

[5] Carmon, Y., Raghunathan, A., Schmidt, L., Liang, P., Duchi, J.C.: Unlabeled data improves adversarial robustness (2019)

[6] Chu, X., Ilyas, I., Papotti, P.: Holistic data cleaning: Putting violations into context. pp. 458–469 (04 2013). https://doi.org/10.1109/ICDE.2013.6544847

[7] Chu, X., Morcos, J., Ilyas, I.F., Ouzzani, M., Papotti, P., Tang, N., Ye, Y.: Katara: A data cleaning system powered by knowledge bases and crowdsourcing. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. p. 1247–1261. SIGMOD '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2723372.2749431, https://doi.org/10.1145/2723372.2749431

[8] Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics. pp. 215–223 (2011)

[9] Culverhouse, P., Williams, R., Reguera, B., Herry, V., González-Gil, S.: Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. Marine Ecology Progress Series **247**, 17–25 (2003)

[10] De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., Others: Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature medicine **24**(9), 1342–1350 (2018)

[11] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848

[12] Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data (2017)

[13] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning. pp. 1321–1330. PMLR (2017)

[14] Kandel, S., Paepcke, A., Hellerstein, J., Heer, J.: Wrangler: Interactive visual specification of data transformation scripts. In: ACM Human Factors in Computing Systems (CHI) (2011), http://vis.stanford.edu/papers/wrangler

[15] Karimi, D., Nir, G., Fazli, L., Black, P.C., Goldenberg, L., Salcudean, S.E.: Deep Learning-Based Gleason Grading of Prostate Cancer From Histopathology Images—Role of Multiscale Decision Aggregation and Data Augmentation. IEEE Journal of Biomedical and Health Informatics **24**(5), 1413–1426 (2020)

[16] Karimi, D., Dou, H., Warfield, S.K., Gholipour, A.: Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. Medical Image Analysis **65** (2020)

[17] Krizhevsky, A., Hinton, G., Others: Learning multiple layers of features from tiny images. Tech. rep., Citeseer (2009)

[18] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. vol. 60, pp. 1097–1105. Association for Computing Machinery (2012)

[19] Kullback, S., Leibler, R.A.: On Information and Sufficiency. Ann. Math. Statist. **22**(1), 79–86 (1951). https://doi.org/10.1214/aoms/1177729694, https://doi.org/10.1214/aoms/1177729694

[20] Li, J., Socher, R., Hoi, S.C.H.: DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In: International Conference on Learning Representations. pp. 1–14 (2020)

[21] Liao, Y.H., Kar, A., Fidler, S.: Towards good practices for efficiently annotating large-scale image classification datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4350–4359 (June 2021)

[22] McHugh, M.L.: Interrater reliability: the kappa statistic. PubMed **Biochemia**(3), 276–82 (2012), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

[23] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**(6) (Jul 2021). https://doi.org/10.1145/3457607, https://doi.org/10.1145/3457607

[24] Niemeyer, F., Galbusera, F., Tao, Y., Kienle, A., Beer, M., Wilke, H.J.: A deep learning model for the accurate and reliable classification of disc degeneration based on mri data. Investigative Radiology **56**(2), 78–85 (2021)

[25] Northcutt, C.G., Jiang, L., Chuang, I.L.: Confident learning: Estimating uncertainty in dataset labels (2021)

[26] Ooms, E., Zonderland, H., Eijkemans, M., Kriege, M., Mahdavian Delavary, B., Burger, C., Ansink, A.: Mammography: Interobserver variability in breast density assessment. The Breast **16**(6), 568–576 (dec 2007)

[27] Peterson, J., Battleday, R., Griffiths, T., Russakovsky, O.: Human uncertainty makes classification more robust. Proceedings of the IEEE International Conference on Computer Vision **2019-Octob**, 9616–9625 (2019)

[28] Riccardi, G., Hakkani-Tur, D.: Active learning: Theory and applications to automatic speech recognition. Speech and Audio Processing, IEEE Transactions on **13**, 504 – 511 (08 2005). https://doi.org/10.1109/TSA.2005.848882

[29] Saremi, R., Yang, Y., Vesonder, G., Ruhe, G., Zhang, H.: Crowdsim: A hybrid simulation model for failure prediction in crowdsourced software development (2021)

[30] Saremi, R.L., Yang, Y., Ruhe, G., Messinger, D.: Leveraging crowdsourcing for team elasticity: an empirical evaluation at topcoder. In: 2017 IEEE/ACM 39th International Conference on Software Engineering: Software Engineering in Practice Track (ICSE-SEIP). pp. 103–112 (2017). https://doi.org/10.1109/ICSE-SEIP.2017.2

[31] Schmarje, L., Brünger, J., Santarossa, M., Schröder, S.M., Kiko, R., Koch, R.: Beyond Cats and Dogs: Semi-supervised Classification of fuzzy labels with overclustering (2020)

[32] Schmarje, L., Santarossa, M., Schröder, S.M., Zelenka, C., Kiko, R., Stracke, J., Volkmann, N., Koch, R.: S2C2 - An orthogonal method for Semi-Supervised Learning on fuzzy labels (2021)

[33] Schmarje, L., Zelenka, C., Geisen, U., Glüer, C.C., Koch, R.: 2D and 3D Segmentation of Uncertain Local Collagen Fiber Orientations in SHG Microscopy. In: DAGM German Conference of Pattern Regocnition, vol. 11824 LNCS, pp. 374–386. Springer (2019)

[34] Schmarje, L., Zelenka, C., Geisen, U., Glüer, C.C., Koch, R.: Dataset of 2D and 3D Segmentation of uncertain local collagen fiber orientations in SHG microscopy (2019)

[35] Schoening, T., Bergmann, M., Ontrup, J., Taylor, J., Dannheim, J., Gutt, J., Purser, A., Nattkemper, T.W.: Semi-automated image analysis for the assessment of megafaunal densities at the Artic deep-sea observatory HAUSGARTEN. PLoS ONE **7**(6), 1–14 (2012). https://doi.org/10.1371/journal.pone.0038179

[36] Shen, Y., Yun, H., Lipton, Z., Kronrod, Y., Anandkumar, A.: Deep active learning for named entity recognition. In: Proceedings of the 2nd Workshop on Representation Learning for NLP. pp. 252–256. Association for Computational Linguistics, Vancouver, Canada (Aug 2017). https://doi.org/10.18653/v1/W17-2630, https://aclanthology.org/W17-2630

[37] Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020) (2020)

[38] Srivastava, M., Hashimoto, T., Liang, P.: Robustness to spurious correlations via human annotations (2020)

[39] Tommasi, T., Patricia, N., Caputo, B., Tuytelaars, T.: A deeper look at dataset bias (2015)

[40] Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR 2011. pp. 1521–1528 (2011). https://doi.org/10.1109/CVPR.2011.5995347

[41] Volkmann, N., Brünger, J., Stracke, J., Zelenka, C., Koch, R., Kemper, N., Spindler, B.: SO MUCH TROUBLE IN THE HERD: DETECTION OF FIRST SIGNS OF CANNIBALISM IN TURKEYS. In: Recent advances in animal welfare science VII Virtual UFAW Animal Welfare Conference. p. 82 (2020)

[42] Xue, Y., Hauskrecht, M.: Efficient learning of classification models from soft-label information by binning and ranking. In: The Thirtieth International Flairs Conference (2017)

[43] Yun, S., Oh, S.J., Heo, B., Han, D., Choe, J., Chun, S.: Re-labeling imagenet: From single to multi-labels, from global to localized labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2340–2350 (June 2021)