Exploring Dataset-Scale Indicators of Data Quality

Anonymous Author(s)
Affiliation
Address
email

Abstract

Modern computer vision foundation models are trained on massive amounts of data, incurring large economic and environmental costs. Recent research has suggested that improving data quality can significantly reduce the need for data quantity. But what constitutes data quality in computer vision? We posit that the quality of a given dataset can be decomposed into distinct sample-level and dataset-level constituents, and that the former have been more extensively studied than the latter. We ablate the effects of two important dataset-level constituents: label set design, and class balance. By monitoring these constituents using key indicators we provide, researchers and practitioners can better anticipate model performance, measured in terms of its accuracy and robustness to distribution shifts.

1 Introduction

1.1 Motivation

In recent years, the field of computer vision has seen remarkable progress in a range of sub-disciplines [Li et al., 2023; Mildenhall et al., 2021; Rombach et al., 2021]. Much of this progress has derived from *foundation models* [Yuan et al., 2021; Radford et al., 2021]. Modern computer vision foundation models are trained on massive amounts of data, and the general trend has been to achieved improved performance by scaling up dataset size.

However, increasing the size of the dataset is not the only way to improve the downstream performance of a given model. Recent work has introduced into the literature the importance of *data quality*, a phrase which is meant to encompass all facets of a dataset which impact downstream performance, aside from its size [Nguyen et al., 2022; Santurkar et al., 2022; Gadre et al., 2023].

The question of *which* facets of data quality are relevant for improved performance, however, is under-explored in the literature. Even less frequently studied is whether there exist *quantifiable, predictive indicators* of these facets. If reliable indicators of dataset quality can be found prior to training, then researchers can better estimate the impact of modifications to their datasets (or even guide dataset design), and ultimately lead to reduction of human, environmental, and economic costs of large model training [Sharir et al., 2020; Changpinyo et al., 2021; Bommasani et al., 2022].

1.2 Our Contributions

In this (preliminary) work, we present several new findings on the constituents of dataset quality. Unlike most previous works, our aim is to discover indicators that are properties at the dataset-scale, rather than measures of image-level quality. Via a series of controlled ablation studies, we explore the downstream performance impact of certain constituents, and show that these can be used as predictive indicators of the quality of datasets prior to model training. In brief:

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

1. We conduct thorough ablation studies on two important dataset properties — *label set size*, and *class imbalance* — and analyze the (sometimes) complex effects they have on downstream metrics in image classification tasks.
2. For each of the above two constituents, we provide a list of *key indicators* which are predictive of model performance. Our indicators are inexpensive to compute, scalable to very large dataset sizes, and can be identified prior to model training.

Our results can be viewed as building blocks towards a systematic taxonomy of the notion of "quality" at the dataset scale, which may enable improved design choices for datasets in computer vision.

2 Related Work

The question of how best to curate a dataset from a raw, large set of images is an important problem in computer vision. The common approach is to use full supervision; classes are chosen in advance, raw samples are algorithmically/manually filtered, and then manually labeled by annotators, and (sometimes) class balance is enforced; the classic example is ImageNet [Deng et al., 2009]. Such datasets tend to produce very strong baselines [Kornblith et al., 2019]. Other approaches take a more relaxed approach to filtration, assigning unfiltered web-scraped images to human labelers and applying a wide range of possible class labels [Kuznetsova et al., 2018].

As datasets scale from millions to billions of samples, human labeling becomes impractical. While early approaches such as Thomee et al. [2016] initially curated datasets without any supervision, *weak supervision* strategies have now become popular. With no human in the loop, proxy measures of quality become essential. Sample-level quality indicators include encoding format, size, aspect ratio, and offensive content [Sharma et al., 2018]. Labeling strategies for these images sometimes rely on image tags from large social image-sharing sites [Sun et al., 2017; Mahajan et al., 2018].

One challenge in studying dataset curation in computer vision derives from the fact that many of the largest image datasets cited in the literature are not publicly avaialable [Sun et al., 2017; Mahajan et al., 2018; Nguyen et al., 2022] A notable recent exception to this is the work of Schuhmann et al. [2021]. Subsequent works such as [Ilharco et al., 2021; Gadre et al., 2023; Feuer et al., 2023] have taken advantage of these developments to train new models and design data-centric challenges.

There have been many proposed approaches to predict the behavior of models for a given test set, coalescing in a NeurIPS competition in December 2020 [Jiang et al., 2021]. Broadly, submitted methods fell into three meta-categories: (i) generalization measures derived from theoretically motivated generalization bounds; (ii) data augmentation methods, which estimate the generalization error of a trained model by computing its accuracy on synthetic data, and (iii) geometry and statistics of intermediate representations. The most successful approach was that of [Natekar & Sharma, 2020], which used a combination of (ii) and (iii).

An important dimension in which foundation models have been found to outperform smaller computer vision models — and which we use as a key metric in our experiments — is *distributional robustness*, a test-time paradigm which aims to estimate model robustness to distribution shifts [Recht et al., 2019]. A *distribution shift* is defined as evaluation data which differs from the data on which a model was trained due to natural factors. Real world image classifiers require predictable model behavior under such shifts. Models trained on large, heterogeneous datasets tend to provide greater distributional robustness than their counterparts trained on less data [Feuer et al., 2023].

3 Experimental Setup

Measurable indicators of dataset quality. As discussed above in Sec. 2, most existing works which attempt to evaluate data quality do so at the *sample level*; they focus on properties pertaining to individual samples, such as image resolution and image fidelity. Our focus instead is on exploring holistic *dataset-level* properties, which are typically determined by the dataset's creators and (if that dataset is used as a benchmark) are typically treated as immutable.

In this short paper, we focus on understanding the predictive power of two such properties: (i) number of classes (label set size), and (ii) number of samples per class (individual class size). We chose these properties since they are ubiquitous, concretely measurable, and tend to have large impacts on the performance of the corresponding trained models.

Data scaling. We define *horizontal scaling* (H-scaling) as scaling up the label set size in the dataset while holding individual class size constant. We define *vertical scaling* (V-scaling) as scaling up individual class size while holding the label set size constant.

Architecture. Our baseline architecture against which all variations are compared is a modified ResNet-50 with a 1000-class linear classification head [He et al., 2015]. The specifics of the modifications are described in [Ilharco et al., 2021]. Whenever we use an architecture other than our standard baseline, we refer to it by its name in the timm library from Wightman [2019].

Pretraining datasets. The experiments described in this paper were primarily conducted using the JANuS dataset, introduced by [Feuer et al., 2023]. JANuS is a composite image-caption dataset sourced from four different datasets of origin. Every sample in JANuS contains one or more labels (for training conventional computer vision models) and one or more captions (for training vision-language models) with varying degrees of supervision. These properties make it an ideal dataset for conducting controlled experiments for H-scaling and V-scaling. In addition to JANuS, we make use of LAION, ImageNet and OpenImages, from Deng et al. [2009]; Kuznetsova et al. [2018]; Schuhmann et al. [2021], respectively; further details are provided in the experiments below.

Training details. In our experiments, we train with mixed precision, at a batch size of 256, and do not use gradient clipping. We use the AMP library to implement the training process. Learning rate is chosen once, via grid search, for each new architecture / dataset pair. Models are typically distributed across a single node with 4 NVIDIA A100 GPUs. All models are trained for 256 epochs. Following Santurkar et al. [2022], we use SimCLR augmentations (resize, crop, flip, jitter, blur, grayscale) rather than CLIP augmentations (resize and crop) for model training. A few of our models are not trained from scratch, but are instead evaluated zero-shot using weights sourced from Wightman [2019]; we note this whenever it is the case.

Labels. We evaluate on a single, broad-scope label set of 100 classes corresponding to ImageNet-100 (IN100), which is the first constituent dataset of JANuS. In our tables, we refer to the validation set for IN100 as IN100-Val, and the average shift accuracy as IN100-Avg. Rob. OI100 refers to OpenImages-100, the second constitutent dataset of JANuS.

Distribution Shifts. Folliwing the literature, whenever we measure robustness, we report the average of four shifts on ImageNet. *ImageNet-V2* was designed to duplicate, as closely as possible, the original ImageNet test set [Recht et al., 2019]. *Imagenet-Sketch* is a distribution shift covering sketches, paintings, drawings and illustrations [Wang et al., 2019]. *Imagenet-R* is a 200-class subset of ImageNet-2012 focused on renditions of everyday objects [Hendrycks et al., 2021a]. *Imagenet-A* is a 200-class subset of ImageNet-2012 which was algorithmically selected [Hendrycks et al., 2021b].

Data filtration. We define *data filtration* as any strategy which sub-selects from a larger pool of possible samples. A simple example of a filtration strategy would be conducting a web search for the target classes in a dataset, and selecting the first k samples in the search.

Metrics for distributional robustness. Our primary metric is *average robustness* (abbv: Avg. Rob.), which is the average test-set accuracy of a model on a set of distribution shifts. Although this measure is easy to interpret, it can conceal substantial performance differences between shifts.

4 Results

The fact that both label set size and class (im)balance impact image classification models should be folklore to computer vision practitioners. However, to the best of our knowledge, these two properties have not been directly contrasted, given an overall data budget. We address the following questions:

1. **For a given overall data budget, is it better to scale up individual class sizes (V-scaling), or to scale the number of classes (H-scaling)?** See Sec. 4.1.
2. **How does class imbalance impact accuracy and robustness?** See Sec. 4.2.

4.1 Label Set Size

Scaling up dataset size has become the *de facto* driving force for improving the accuracy (and robustness) of image classification models. But what is the *right* way to scale up datasets: should we just scale up samples per class? Or are there benefits if the model is trained on a larger set of classes?

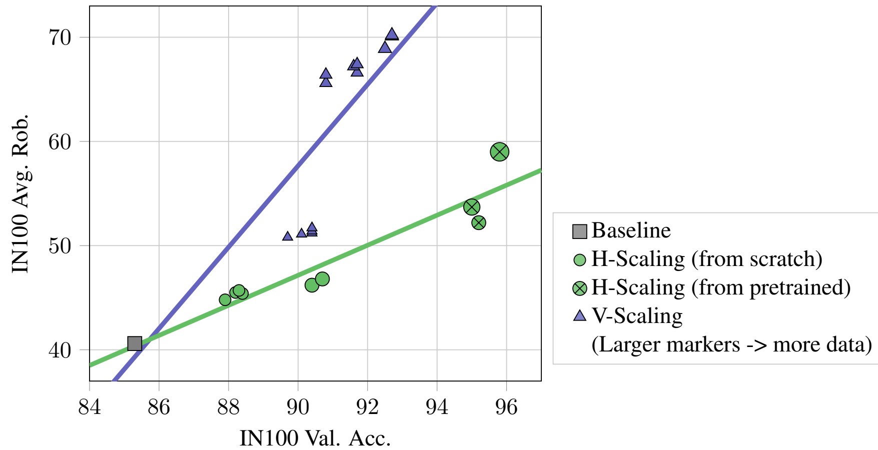

Figure 1: ***Models benefit from scaling data horizontally, as well as vertically.*** *Vertical scaling (increasing the number of samples for in-distribution classes) is generally understood to improve accuracy and robustness. We show that horizontal scaling, increasing the number of out-of-distribution samples, also helps. Dot size represents size of training dataset (vertical scaling is ID, horizontal is OOD). Darker points represent models with more parameters. Image best viewed in color.*

To fairly compare these two choices, we keep the test label set constant. The net effect of the latter case is that the model sees image examples that are OOD with respect to the test set, and we zero out the logits of the OOD classes at test time.

We enumerate results on two distinct sets of large scale classification models. All vertically scaled models are trained by us. For horizontal scaling, we employ a mix of pretrained and from-scratch (with the larger models generally evaluated from pretrained checkpoints). We distinguish between them in Fig. 1 with distinct labels.

4.1.1 Key Indicators

A natural indicator for vertical scaling is per-class dataset size (though this approach demands separate attention to class balance; see Sec. 4.2). Training label set size is the indicator for horizontal scaling.

4.1.2 Findings

Somewhat surprisingly, we find that models achieve more robust and accurate overall representations via horizontal scaling. It is well known that models benefit from seeing more ID samples during

Model	Dataset Size	Training Stages	IN100-Val / Avg. Rob.
resnetv2_50x1_bit	15.4 Mn	PT 21800 FT 1000 ZS 100	95.8% / 59.0%
resnetv2_50	12.6 Mn	PT 1000 FT 12000 ZS 100	95.0% / 53.7%
resnetv2_50	1.2 Mn	PT 1000 ZS 100	95.2% / 52.2%
swin_base_patch4_window7	14.2 Mn	PT 21800 ZS 100	86.7% / 43.4%
flexivit_base_1000ep	14.2 Mn	PT 21800 ZS 100	68.9% / 18.8%
convnext_base	14.2 Mn	PT 21800 ZS 100	66.4% / 36.2%
resnetv2_50x1_bit	14.2 Mn	PT 21800 ZS 100	26.7% / 5.2%

Table 1: ***Horizontal scaling with large label sets.*** *We find that horizontal scaling with large label sets works much better when training occurs in multiple stages. Surprisingly, the process works almost as well whether the larger label set comes first or second; label set size is not an inherent obstacle to the success of horizontal scaling. Model names correspond to those found in the timm library [Wightman, 2019]. PT (pretraining) refers to the first stage of model training. FT (fine-tuning) refers to subsequent stages of model training. ZS (zero-shot) refers to the evaluation process described in Sec. 4.1.2. Parameter count is rounded to the nearest million; label set size to the nearest hundred.*

training, and that contrastive models require positive as well as large batch negative samples to learn [Shalev-Shwartz & Ben-David, 2014; Radford et al., 2021]. Less well-studied is whether non-contrastive models also benefit from scaling batch negatives (OOD classes, in this case). We find that increasing the number of OOD classes from 0 to 900 leads to reliable gains in both accuracy and robustness.

OOD Data Source	Dataset Size	IN100-Val / Avg. Rob.
OpenImages	1.2 Mn	91.5% / 50.2%
ImageNet	1.3 Mn	90.7% / 46.8%
FractalDB	1 Mn	85.4% / 40.2%

Table 2: ***Horizontal scaling with different out-of-distribution approaches.*** *Horizontal scaling works as well or better when the out-of-distribution classes are not from ImageNet. Horizontal scaling on synthetic out-of-distribution classes underperforms scaling on natural images. Dataset size is rounded to the nearest 100,000.*

When we pretrain and fine-tune models, even very large label sets can benefit from horizontal scaling. As the label set grows extremely large, our naive approach to horizontal scaling fails; we examine a range of timm model checkpoints pretrained on the entirety of ImageNet (which contains almost 22,000 classes), even larger and better-performing architectures fail to match much smaller models fine-tuned on IN1000 (see Tab. 1).

There are many potential explanations for this phenomenon. Candidates include the lack of a validation set against which to optimize, or the extreme class imbalance. Another possible explanation is that horizontal scaling cannot succeed when the class space is very large. We show that this is not the case – multi-stage training, even on very large label sets, (see Tab. 1, lines 1, 2) outperforms naive horizontal scaling on 1000 classes (Tab. 1, line 3).

The fact that pretraining and fine-tuning seems to work in *either direction* is particularly surprising to us. Starting with fewer classes and fine-tuning on many more works almost as well as the more intuitive approach of starting with a model pretrained on many classes and fine-tuning on a much smaller number of classes. We leave further exploration of these dynamics to future work.

Horizontal scaling works with non-ImageNet images. In (Tab. 2), we describe the results of horizontal scaling on both ImageNet and non-ImageNet out of distribution classes. Specifically, we experiment with adding classes from OpenImages which do not overlap with ImageNet-1000 classes, and synthetic classes from the FractalDB dataset [Kataoka et al., 2020]. We find that the OpenImages classes actually *out-perform* the ImageNet classes substantially, despite the dataset being slightly smaller. This finding suggests that dataset blending, combined with horizontal scaling, is a promising approach for training more performant computer vision models.

4.2 Class (Im)Balance

It has long been observed that class-imbalanced datasets underperform compared to balanced ones [Goodfellow et al., 2016]. In line with those observations, we confirm that class imbalance has a substantial negative effect on both validation accuracy and average robustness for IN100 classification. Less frequently explored is the question of *why* imbalanced datasets underperform.

Inspired by recent work investigating Zipfian distributions Chan et al. [2022], we posit two hypotheses. Zipf's law stipulates that the frequency of an event is inversely proportional to its rank in a frequency table. Zipfian distributions occur in natural language, where a small number of words (like "the" and "and") occur very frequently, while the majority of words occur rarely. We interrogate each of these hypotheses: first, that the underperformance is due to the existence of a few overrepresented classes, a property which we call **left-skewedness**; second, that the underperformance is due to the existence of long-tail classes with very few samples in them, which we call **long-tailedness**.

4.2.1 Key indicators

We propose two potential key indicators to better compare these hypotheses. Our proposed indicator for left-skewedness is the percentage of samples in the dataset which are members of the most common k% of classes. For the problems described in this paper, we heuristically set $k = 5$. In a perfectly balanced dataset, then, left-skewedness will be 5%. In OI100 without rebalancing, it is 64%.

Our proposed indicator of long-tailedness is the percentage of classes with fewer than k samples in them (%<k Classes). We heuristically explore two choices: $k = 500$ and $k = 100$.

4.2.2 Findings

Data Source	Dataset Size	Left-skew	Long-tail @ 500 / Long-tail @ 100	IN100-Val / Avg. Rob.
in100 (100%)	125,000	5%	0% / 0%	85.3% / 40.6%
in100 (62%)	130,000	5%	0% / 0%	82.5% / 43.1%
oi100 (71%)	190,000	45%	0% / 0%	82.2% / 44.3%
oi100 (60%)	101,000	13%	0% / 0%	79.3% / 41.3%
oi100 (88%)	90,000	25%	0% / 0%	76.6% / 38.8%
oi100 (67%)	135,000	31%	9% / 9%	73.9% / 40.7%
oi100 (57%)	105,000	12%	9% / 9%	73.4% / 39.1%
oi100 (100%)	135,000	64%	64% / 9%	67.7% / 37.2%
oi100 (100%)	53,000	18%	67% / 9%	58.2% / 31.1%

Table 3: **Under-represented classes trigger performance declines.** *We ablate class imbalance by blending samples from ImageNet and OpenImages, which share a data source. Surprisingly, class imbalance alone does not cause model performance to degrade (Left-skew, the percentage of samples which are in one of the 5 most common classes, is not predictive of performance). Rather, it is the existence of long-tail classes containing very few samples (Long-tail @ k). The 9 longest tailed classes in OI100 ($k = 100$) account for the majority of performance decline. Sample sizes are rounded to the nearest 1,000 samples. Percentiles are rounded to the nearest percentile.*

Our main results on class imbalance can be found in Tab. 3, where we report data source, dataset size, our key indicators, validation accuracy, and average accuracy under shift.

Surprisingly, we find that class imbalance alone does not trigger a performance decline. When we preserve the imbalance in the largest classes of OpenImages, but rebalance small classes, performance *improves* slightly (see Tab. 3 line 3) compared to the baseline (Tab. 3 line 8). Furthermore, when we decrease the degree of imbalance in OpenImages by truncating the largest classes, performance degrades (see Tab. 3 line 9). We conclude that models improve when training on more samples of a given class, even when classes are imbalanced.

Rather, the performance declines are attributable the presence of underrepresented classes. In Tab. 3, our long-tailedness indicator is in perfect rank-order agreement with IN100 validation accuracy. Dataset size and left-skewedness exhibit only weak agreement.

Furthermore, the very smallest classes account for most of the decline. We ablate this by rebalancing all classes with more than $k = 100$ samples (using this approach, 91 of 100 classes are balanced). We find that so doing accounts for only 39% of the overall accuracy gains. The majority of accuracy loss from class imbalance is attributable to the very small classes.

5 Conclusions and Future Work

In this paper, we outline useful indicators of data quality which are inexpensive to compute and can be used during pretraining. In Sec. 4.1, we showed that horizontal scaling benefits models trained from scratch. Based on this, for small datasets, as an alternative to pretraining and fine-tuning, we suggest co-training on target classes and classes drawn from existing image datasets in order to ensure that the overall label set size is large. In Sec. 4.2, we showed that underrepresented classes, rather than class imbalance, harms model performance. In a setting with limited resources, we recommend rebalancing the classes with the fewest samples first.

In future work, we hope to introduce additional important dataset-level factors that influence pretraining and provide indicators which make use of the image space as well as the caption space.

References

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, July 2022. URL `http://arxiv.org/abs/2108.07258`. arXiv:2108.07258 [cs].

Stephanie C. Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya Singh, Pierre H. Richemond, Jay McClelland, and Felix Hill. Data Distributional Properties Drive Emergent In-Context Learning in Transformers, November 2022. URL `http://arxiv.org/abs/2205.05055`. arXiv:2205.05055 [cs].

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Benjamin Feuer, Ameya Joshi, Minh Pham, and Chinmay Hegde. Distributionally Robust Classification on a Data Budget. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=D5Z2E8CNsD`.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. DataComp: In search of the next generation of multimodal datasets, July 2023. URL `http://arxiv.org/abs/2304.14108`. arXiv:2304.14108 [cs].

Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. `http://www.deeplearningbook.org`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL `http://arxiv.org/abs/1512.03385`.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021b.

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. URL https://doi.org/10.5281/zenodo.5143773.

Yiding Jiang, Parth Natekar, Manik Sharma, Sumukh K. Aithal, Dhruva Kashyap, Natarajan Subramanyam, Carlos Lassance, Daniel M. Roy, Gintare Karolina Dziugaite, Suriya Gunasekar, Isabelle Guyon, Pierre Foret, Scott Yak, Hossein Mobahi, Behnam Neyshabur, and Samy Bengio. Methods and analysis of the first competition in predicting generalization of deep learning. In Hugo Jair Escalante and Katja Hofmann (eds.), *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pp. 170–190. PMLR, 06–12 Dec 2021. URL https://proceedings.mlr.press/v133/jiang21a.html.

Hirokatsu Kataoka, Kazushige Okayasu, Asato Matsumoto, Eisuke Yamagata, Ryosuke Yamada, Nakamasa Inoue, Akio Nakamura, and Yutaka Satoh. Pre-training without natural images. In *Asian Conference on Computer Vision (ACCV)*, 2020.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do Better ImageNet Models Transfer Better? In *CVPR*. arXiv, June 2019. doi: 10.48550/arXiv.1805.08974. URL http://arxiv.org/abs/1805.08974. arXiv:1805.08974 [cs, stat].

Alina Kuznetsova, Hassan Rom, Neil Gordon Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4. *International Journal of Computer Vision*, 2018.

Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-SAM: Segment and Recognize Anything at Any Granularity, July 2023. URL http://arxiv.org/abs/2307.04767. arXiv:2307.04767 [cs].

Dhruv Mahajan, Ross B. Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II*, volume 11206 of *Lecture Notes in Computer Science*, pp. 185–201. Springer, 2018. doi: 10.1007/978-3-030-01216-8_12. URL https://doi.org/10.1007/978-3-030-01216-8_12.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, December 2021. ISSN 0001-0782. doi: 10.1145/3503250. URL https://dl.acm.org/doi/10.1145/3503250.

Parth Natekar and Manik Sharma. Representation based complexity measures for predicting generalization in deep learning, 2020.

Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of clip. *NeurIPS*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, and Tatsunori Hashimoto. Is a caption worth a thousand images? a controlled study for representation learning. *ICLR*, 2022.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *Data Centric AI NeurIPS Workshop*, 2021.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014. ISBN 1107057132.

Or Sharir, Barak Peleg, and Yoav Shoham. The Cost of Training NLP Models: A Concise Overview, April 2020. URL `http://arxiv.org/abs/2004.08900`. arXiv:2004.08900 [cs].

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL `https://aclanthology.org/P18-1238`.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 843–852, October 2017. doi: 10.1109/ICCV.2017.97. ISSN: 2380-7504.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl S. Ni, Douglas N. Poland, Damian Borth, and Li-Jia Li. Yfcc100m: the new data in multimedia research. *Commun. ACM*, 59:64–73, 2016.

Haohan Wang, Songwei Ge, Eric P. Xing, and Zachary Chase Lipton. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019.

Ross Wightman. Pytorch image models. *GitHub repository*, 2019. doi: 10.5281/zenodo.4414861.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A New Foundation Model for Computer Vision, November 2021. URL `http://arxiv.org/abs/2111.11432`. arXiv:2111.11432 [cs].