Identifying Statistical Bias in Dataset Replication

Logan Engstrom [*1] **Andrew Ilyas** [*1] **Shibani Santurkar** [1] **Dimitris Tsipras** [1] **Jacob Steinhardt** [2]
Aleksander Mądry [1]

Abstract

Dataset replication is a useful tool for assessing whether improvements in test accuracy on a specific benchmark correspond to improvements in models' ability to generalize reliably. In this work, we present unintuitive yet significant ways in which standard approaches to dataset replication introduce statistical bias, skewing the resulting observations. We study ImageNet-v2, a replication of the ImageNet dataset on which models exhibit a significant (11-14%) drop in accuracy, even after controlling for *selection frequency*, a human-in-the-loop measure of data quality. We show that after remeasuring selection frequencies and correcting for statistical bias, only an estimated $3.6\% \pm 1.5\%$ of the original $11.7\% \pm 1.0\%$ accuracy drop remains unaccounted for. We conclude with concrete recommendations for recognizing and avoiding bias in dataset replication. Code for our study is publicly available[1].

1. Introduction

The primary objective of supervised learning is to develop models that generalize robustly to unseen data. Benchmark test sets provide a proxy for out-of-sample performance, but can outlive their usefulness. For example, evaluating on benchmarks alone may steer us towards models that adaptively overfit (Reunanen, 2003; Rao et al., 2008; Dwork et al., 2015) to the test set and do not generalize. Alternatively, we might select for models that are sensitive to insignificant aspects of the dataset creation process and thus do not generalize robustly (e.g., models that are sensitive to the exact humans who annotated the test set).

To diagnose these issues, recent work has generated new, previously "unseen" testbeds for standard datasets through a process known as dataset replication. Though not yet widespread in machine learning, dataset replication is a natural analogue to experimental replication studies in the natural sciences (cf. (Bell, 1973)). These studies play an important role in verifying empirical findings, and ensure that results are neither affected by adaptive data analysis, nor overly sensitive to experimental artifacts.

Recent dataset replication studies (Recht et al., 2019b;a; Yadav & Bottou, 2019) have generally found little evidence of adaptive overfitting: progress on the original benchmark translates to roughly the same amount (or more) of progress on newly constructed test sets. On the other hand, model performance on the replicated test set tends to drop significantly from the original one.

One of the most striking instances of this accuracy drop is observed by Recht et al. (2019b), who performed a careful replication of the ImageNet dataset and observe an 11-14% gap between model accuracies on ImageNet and their new test set, ImageNet-v2. The magnitude of this gap presents an empirical mystery, and motivates us to understand what factors cause such a large drop in accuracy.

In this paper, we identify a mechanism through which the dataset replication process itself might lead to such a drop: noisy readings during data collection can introduce statistical bias. We show that re-calibrating the ImageNet-v2 dataset while correcting for this bias results in an accuracy gap of $3.6\% \pm 1.5\%$, compared to the original $11.7\% \pm 1.0\%$ drop between ImageNet and ImageNet-v2.

Our explanation revolves around what we refer to as the "statistic matching" step of dataset replication. Statistic matching ensures that model performance on the original test set and its replication are comparable by controlling for variables that are known to (or hypothesized to) impact model performance.[2] Drawing a parallel to medicine, suppose we wanted to replicate a study about the effect of a certain drug on an age-linked disease. After gathering subjects, we have to reweight or filter them so that the age distribution matches that of the original study—otherwise, the results of the studies are incomparable. This filter-

[*] Equal contribution [1] MIT [2] UC Berkeley. Correspondence to: Logan Engstrom <engstrom@mit.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

[1] `https://git.io/data-rep-analysis`

[2] In causal inference terms, statistic matching is an instance of covariate balancing (Stuart, 2010; Imai & Ratkovic, 2013).

ing/reweighting step is analogous to statistic matching in our context, with participant age as the relevant statistic.

To construct ImageNet-v2, Recht et al. (2019b) perform statistic matching based on the "selection frequency" statistic, which for a given image-label pair measures the rate at which crowdsourced annotators select the pair as correctly labeled. As we discuss in the next section, selection frequency is a well-motivated choice of matching statistic, since (a) Deng et al. (2009) use a similar metric to gather ImageNet images in the first place (Deng et al., 2009), and (b) Recht et al. (2019b) have found that selection frequency is highly predictive of model accuracy.

Why does a significant drop in accuracy persist even after matching selection frequencies? In this paper, we show that (inevitable) mean-zero noise in selection frequency readings leads to bias in the selection frequencies of the replicated dataset, which translates to a drop in model accuracies. Finite-sample reuse makes this bias difficult to detect.

The bias-inducing mechanism that we identify applies whenever statistic matching is performed using noisy estimates. We characterize the mechanism theoretically in Section 2. In Section 3, we remeasure selection frequencies using Mechanical Turk and observe that as our mechanism predicts, ImageNet-v2 images indeed have lower selection frequency on average. After presenting a framework for studying the effect of statistical bias on model accuracy (Section 4), we use de-biasing techniques to estimate a bias-corrected accuracy for ImageNet-v2 (Section 5) using the remeasured selection frequencies. In Section 7, we discuss the implications of the identified mechanism for ImageNet-based computer vision models specifically, and for data replication studies more generally.

2. Identifying Sources of Reproduction Bias

The goal of dataset replication is to create a new dataset by reconstructing the pipeline that generated the original test set as closely as possible. We expect (and intend) for this process to introduce a distribution shift, partly by varying parameters that should be irrelevant to model performance (e.g. the exact identity of the annotators used to filter the dataset). To ensure that results are comparable with original test sets, however, dataset replication studies must control for distribution shifts in variables that impact task performance. This is accomplished by subsampling or reweighting the data so that each relevant variable's distributions under the replicated dataset and the original dataset match. We refer to this process as *statistic matching*.

Our key observation is that standard approaches to statistic matching can lead to bias in the final replicated dataset: we illustrate this phenomenon in the context of the ImageNet-v2 (`v2`) dataset replication (Recht et al., 2019b). Before we identify the source of this bias in ImageNet-v2 construction, we review the data collection process for both ImageNet and ImageNet-v2.

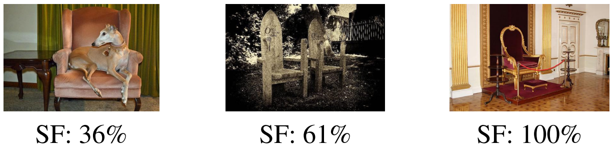

SF: 36% SF: 61% SF: 100%

Figure 1. The smallest, median, and largest selection frequency images from `v1` corresponding to the "throne" class (description: *the chair of state for a monarch, bishop, etc.; "the king sat on his throne"*—the "throne" class was randomly chosen). The images become easier to identify as the labeled class as selection frequency increases; for additional context, we give a random sampling of selection frequency/image pairs in Appendix B.

ImageNet and selection frequency. ImageNet (Deng et al., 2009; Russakovsky et al., 2015) (which we also refer to as ImageNet-v1 or `v1`) is one of the most widely used datasets in computer vision. To construct ImageNet, Deng et al. (2009) first amassed a large candidate pool of image-label pairs using image search engines such as Flickr. The authors then asked annotators on Amazon Mechanical Turk (MTurk) to select the candidate images that were correctly labeled. Each image is shown to multiple annotators, and an image's *selection frequency* [3] is then defined as the fraction of annotators that selected it.

Intuitively, images with low selection frequency are likely either confusing or incorrectly labeled, while images with high selection frequency are "easy" for humans to identify as the proposed label (we show examples of selection frequencies in Figure 1; further examples are in Appendix 8). Therefore, Deng et al. (2009) include only images with high selection frequency in the final ImageNet dataset[4].

ImageNet-v2. ImageNet-v2 is a replication of ImageNet-`v1` that controls for selection frequency via statistic matching. Following the protocol of Deng et al. (2009), Recht et al. (2019b) collected a pool of candidate image-label pairs, and estimated their selection frequencies via MTurk, along with a subset of the `v1` validation set. Recht et al. (2019b) then estimated the distribution of ImageNet-v1 selection frequencies for each class. They subsampled 10 images of each class from the candidate pool according to the estimated class-specific distributions of `v1`.

For example, suppose 40% of "goldfish" images in

[3] Note that the term "selection frequency" was in fact coined by Recht et al. (2019b), but it is also useful for describing the initial setup of Russakovsky et al. (2015), who instead referred to their process as "majority voting."

[4] An image is included in the ImageNet test set if a "convincing majority" (Russakovsky et al., 2015) of annotators select it.

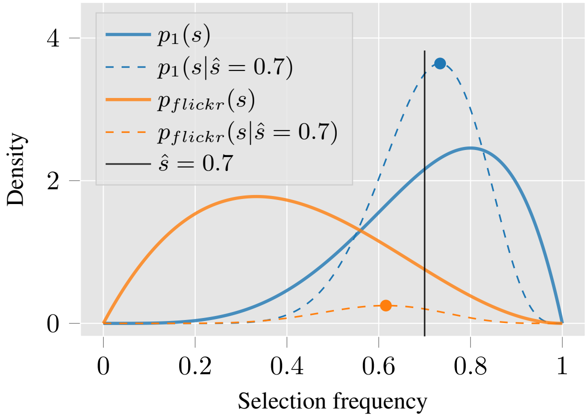


Figure 2. For an image x, the *selection frequency* $s(x) \in [0, 1]$ of an image, described in Section 2, captures how recognizable it is to humans. A distribution over images induces a one-dimensional distribution over selection frequencies $p_i(s)$, shown in solid orange and blue for the Flickr and ImageNet-v1 data distributions respectively. We consider a case where we are given, for a specific image x, a *noisy* version of $s(x)$ ($\hat{s}(x)$). We visualize the corresponding distribution of the true selection frequency $s(x)$ given this noisy $\hat{s}(x) = 0.7$. As discussed in Section 2, even though $\hat{s}(x)$ is an unbiased estimate of $s(x)$, the most likely value of $s(x)$ for a given noisy reading of $\hat{s}(x)$ depends on the distribution from which x is drawn. This is the driving phenomenon behind the observed bias between ImageNet and ImageNet-v2.

ImageNet-v1 have selection frequency in the histogram bucket $[0.6, 0.8]$—when constructing ImageNet-v2, Recht et al. (2019b) would in turn sample 4 "goldfish" images from the same histogram bucket in the candidate images.

Statistic matching should ensure that v1 and v2 are balanced in terms of selection frequency, and partly justifies the expectation that models perform similarly on both.

Sources of bias. We identify two places where the matching strategy of Recht et al. (2019b) might introduce statistical bias. One potential source of bias could arise from binning the images into histograms—since there are relatively few bins, within each bin the ImageNet images might have different selection frequencies from the corresponding Flickr images. (For example, the ImageNet-v1 images in the $s(x) \in [0., 0.2]$ bucket might actually have selection frequency 0.15, whereas the Flickr images in the same bucket might have $s(x) = 0.1$.) However, this source of error appears to have not had a pronounced effect (at least on average), as Recht et al. (2019b) report that the average selection frequency of the ImageNet-v2 images actually matches that of the ImageNet-v1 test set.

Our analysis revolves around a second and more subtle source of bias, however. This bias stems from the fact that for any given image x, the selection frequency $s(x)$ is never measured exactly. Instead, we are only able to measure $\hat{s}(x)$, a finite-sample estimate of the statistic, attained by averaging over a relatively small number of annotators.

To model the impact of this seemingly innocuous detail, suppose that the selection frequencies $s(x)$ of ImageNet and Flickr images are distributed according to $p_1(s(x))$ and $p_{flickr}(s(x))$ respectively (or more briefly, $p_1(s)$ and $p_{flickr}(s)$)—see Figure 2 for a visualization. Now, suppose that for an image x, we get an unbiased noisy measurement $\hat{s}(x) = 0.75$ of the selection frequency via crowdsourcing. Then, even if $\hat{s}(x)$ is an unbiased estimate of $s(x)$, the most likely value of $s(x)$ for the image is not $\hat{s}(x)$, but in fact depends on the distribution from which x was drawn. Indeed, for the (hypothetical) distributions shown in Figure 2, if x is a Flickr image then it is more likely that $s(x) < 0.75$ and $\hat{s}$ is an overestimate, since a priori an image is likely to have a low selection frequency (i.e., there is more $p_{flickr}(s)$ mass below 0.75) and the noise is unbiased. Conversely, if x is an ImageNet test set image in this same setting, it is more likely that $s(x) > 0.75$. Therefore, if we use a Flickr image with a noisy selection frequency 0.75 to "match" an ImageNet image with the same noisy selection frequency, the true selection frequency of the ImageNet image is actually likely to be higher. We can make this explicit by writing down the likelihood of s given $\hat{s} = 0.75$ (also plotted in Figure 2):

$$p_i(s|\hat{s} = 0.75) = \frac{p_i(s) \cdot p(\hat{s} = 0.75|s)}{p_i(\hat{s} = 0.75)} \; \forall\, i \in \{1,\, flickr\},$$

which depends on the prior $p_i(\cdot)$ and therefore is not equal for both values of i.

The distribution of candidate Flickr images is likely skewed to have lower selection frequencies than v1—after all, Deng et al. (2009) narrowed down Imagenet-v1 from a large set of candidates based on quality. Therefore, one would expect the underlying true selection frequencies of the v1 images to be higher than (and in general, not equal to) their matched ImageNet-v2 counterparts.

A simple model of the bias. To better understand the source of the bias, consider a simple model in which the ImageNet-v2 selection process is cast as a rejection sampling procedure. Here, the densities $p_1(\hat{s}(x))$ and $p_{flickr}(\hat{s}(x))$ are estimated from samples (analogous to the histograms of Recht et al. (2019b))—then, for a given Flickr image x, we "accept" x into the v2 dataset with probability proportional to $p_1(\hat{s}(x))/p_{flickr}(\hat{s}(x))$ (analogous to the bin-wise sampling of Recht et al. (2019b)). If selection frequency readings were not noisy, i.e. if $\hat{s}(x) = s(x)$, then the resulting density of selection frequencies in the v2 dataset would be given by

$$p_{flickr}(s(x)) \cdot \frac{p_1(s(x))}{p_{flickr}(s(x))} = p_1(s(x)),$$

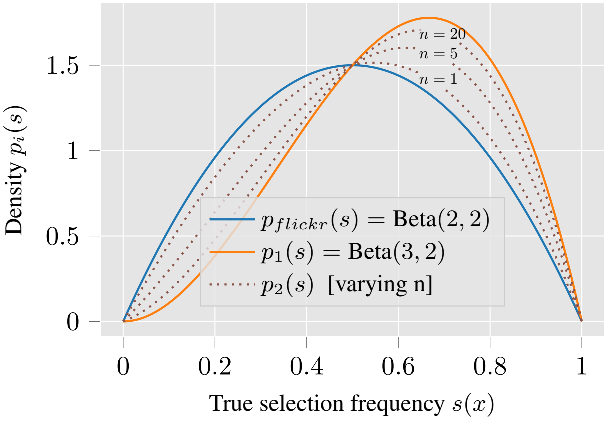


Figure 3. Illustrations accompanying the simple theoretical model, where we assume $p_1(s(x))$ and $p_{flickr}(s(x))$ are $\text{Beta}(\alpha+1, \beta)$ and $\text{Beta}(\alpha, \beta)$ (here $\alpha = \beta = 2$). As more samples are used to estimate $s(x)$ for each image, the resulting ImageNet-v2 distribution tends towards the v1 distribution, but does not match it for any finite number of samples per image.

and the selection frequencies of v2 would be distributed in the same way as those of v1, as intended. However, the inevitable noisiness of the selection frequencies means that in reality, the density of selection frequencies for v2 is

$$p_{flickr}(s(x)) \cdot \int_{\hat{s}} p(\hat{s}|s) \frac{p_1(\hat{s}(x))}{p_{flickr}(\hat{s}(x))}.$$

Now as a toy example, suppose $p_{flickr}(s)$ and $p_1(s)$ are given by beta distributions $\text{Beta}(\alpha, \beta)$ and $\text{Beta}(\alpha+1, \beta)$ respectively (c.f. Figure 2). Furthermore, suppose that $\hat{s}(x)$ is given by an average of n Bernoulli draws with success probability $s(x)$. Then, a series of calculations (shown in Appendix C) reveals that the resulting v2 selection frequency distribution is given by:

$$\frac{n}{n+\beta+\alpha} \cdot p_1(s) + \frac{\alpha+\beta}{n+\alpha+\beta} \cdot p_{flickr}(s). \qquad (1)$$

Note that as $n \to 0$ (no filtering is done at all), the above expression evaluates to exactly $p_{flickr}(s)$, as expected. Then, as the number of workers n tends to infinity (i.e. $\hat{s}$ becomes less noisy), the distribution of ImageNet-v2 selection frequencies converges to the desired $p_1(s)$. For any finite n, however, the resulting v2 distribution will be a non-degenerate mixture between $p_{flickr}(s)$ and $p_1(s)$, and therefore does not match the distribution of selection frequencies $p_1(s)$ exactly. The results of this toy model (depicted in Figure 2) capture the bias that could be incurred by the data replication pipeline of Recht et al. (2019b).

3. Remeasuring Selection Frequencies

In this section, we measure the effect of the described noise-induced bias on the true and observed selection frequencies of images in ǎnd v2. Using an annotation task closely resembling those of the ImageNet-v2 and ImageNet MTurk experiments, we collect new selection frequency estimates for all of ImageNet-v2 and for a subset of ImageNet. In these tasks, MTurk annotators were shown grids of 48 images at a time, each corresponding to an ImageNet class. Each grid contained a mixture of ImageNet, Flickr, and in our case, ImageNet-v2 images of the corresponding class (since ImageNet-v2 was not yet realized at the time of the other experiments), as well as control images from other classes. We describe the setup in more detail in Appendix B.1. Annotators were tasked with selecting all the images in the grid containing an object from the class in question. Each image was seen by 40 distinct annotators, and assigned an observed selection frequency equal to the fraction of these workers that selected it.

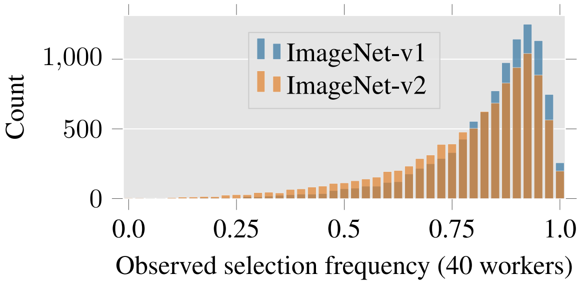


Figure 4. Selection frequency histograms for v1 and v2 based on our selection frequency re-measurement experiment. Results indicate that v2 seems to have lower selection frequency.

Histograms of observed selection frequencies for v1 and v2 are shown in Figure 4. We find that the average selection frequencies of the v1 and v2 images were 85.2% ± 0.1% and 80.7% ± 0.1% respectively compared to 71% and 73% reported by Recht et al. (2019b) [5]. Thus, the initial 2% gain in selection frequency measured by Recht et al. (2019b) turns into a 5% drop[6]. Our model of dataset replication bias predicts this discrepancy: once observed selection frequencies are used for matching, they no longer provide an unbiased estimate of true selection frequency.

Detecting bias using the original data. Our MTurk task measures a significant selection frequency gap betweeen v1 and v2 (~5%), but also measures average selection frequencies for both datasets to be significantly higher than reported by Recht et al. (2019b), suggesting differences in experimental setup. Indeed, while the tasks themselves were identical, we did make a few changes to the deployment setup of Recht et al. (2019b) to improve data qual-

[5] 95% bootstrapped CI.

[6] Our model in Section 2 predicts a distributional difference in selection frequencies between v1 and v2; a gap between means is sufficient but not necessary evidence for this difference.

ity. These changes are outlined in Appendix B.2: examples include introducing worker screening qualifications[7], and using different proportions of images per grid. Since the task interface remained constant and workers are not able to distinguish between ImageNet-v1 and ImageNet-v2 images while labeling, we believe that the changes made improve data quality across both datasets while negligibly affecting the selection frequency gap between them.

Still, we can fully control for experimental differences by analyzing the raw data of Recht et al. (2019b) directly, taking care to avoid bias from observed selection frequency reuse. We defer the exact data analysis to Appendix D. Although there are insufficient samples to properly estimate the bias-adjusted accuracy with the original data, we show that the observed results are consistent with a large accuracy correction (i.e., our 3.6% gap estimate is plausible). These results suggest that statistic matching bias affects the `v2` dataset, even fully controlling for experimental setup. In the coming sections, we quantify the effects of this bias on model accuracies.

4. Understanding the Accuracy Gap

Our findings so far have suggested that statistic matching bias results in a downwards bias in ImageNet-v2 true selection frequencies. In this section, we quantify the impact on this bias on ImageNet-v2 accuracy.

4.1. Notation and terminology

Here we overview the notation and terminology useful in discussing the bias in ImageNet-v2 accuracy.

Selection frequencies. In Section 2 we defined the true selection frequency $s(x)$ for an image x as the (population) rate at which crowd annotators select the image as correctly labeled. The true selection frequency of an image is unobservable, and often approximated by the observed selection frequency, $\hat{s}_n(x) \sim \frac{1}{n}\text{Binom}\left(n, s(x)\right)$, which can be estimated from an n-annotator MTurk experiment. When n is clear from context, we will omit it and write $\hat{s}(x)$.

Distributions. We use $\mathcal{D}_1$ and $\mathcal{D}_2$ to denote the distributions of `v1` and `v2` images respectively, and $\mathcal{S}_1$ and $\mathcal{S}_2$ for the corresponding finite test sets. As in Section 2, we denote by $p_i(s)$ the probability density of true selection frequencies for images drawn from $\mathcal{D}_i$. Similarly, we use $p_i(\hat{s}_n(x))$ to denote the probability mass function of the observed selection frequency for dataset i.

We let $\mathcal{D}_2|s_1$ be the distribution of `v2` images reweighted to have the same selection frequency distribution as `v1`. Formally, $\mathcal{D}_2|s_1$ is the compound distribution $(x_2 \sim \mathcal{D}_2|s(x_2) \sim p_1(s))$. Sampling from $\mathcal{D}_2|s_1$ corresponds to first sampling a `v1` image x_1, then sampling an image x_2 from `v2`, conditioned on $s(x_2) = s(x_1)$.

Accuracies. For a classifier c, let $f_c(x)$ be an indicator variable of whether c correctly classifies x. Since our analysis applies to any fixed classifier c, we omit it and use $f(x)$. We then define $\mathcal{A}_X$ as classifier accuracy on distribution or test set X—for example, accuracy on `v1` is given by $\mathcal{A}_{\mathcal{D}_1} = \mathbb{P}_{x_1 \sim \mathcal{D}_1}\left(f(x_1) = 1\right) = \mathbb{E}_{x_1 \sim \mathcal{D}_1}\left[f(x_1)\right]$.

4.2. Breaking down the accuracy gap

The accuracy gap between the `v1` and `v2` test sets is given by $\mathcal{A}_{\mathcal{S}_1} - \mathcal{A}_{\mathcal{S}_2}$. What fraction of this gap can be attributed to bias in selection frequency? To answer this, we decompose this accuracy gap into three elements whose contribution can be studied separately:

$$\underbrace{\left(\mathcal{A}_{\mathcal{S}_1} - \mathcal{A}_{\mathcal{D}_2|s_1}\right)}_{\text{bias-corrected accuracy gap}} + \underbrace{\left(\mathcal{A}_{\mathcal{D}_2|s_1} - \mathcal{A}_{\mathcal{D}_2}\right)}_{\text{selection gap}} + \underbrace{\left(\mathcal{A}_{\mathcal{D}_2} - \mathcal{A}_{\mathcal{S}_2}\right)}_{\text{finite sample gap} \approx 0}. \quad (2)$$

Bias-corrected accuracy gap. The first term of (2), called the bias-corrected accuracy gap, captures the portion of the `v1`-`v2` accuracy drop that *cannot* be explained by a difference in selection frequency, and instead might be explained by benign distribution shift or adaptive overfitting.

Selection gap. The second term of (2) is accuracy gap that can *only* be attributed to selection frequency, since it compares accuracy on $\mathcal{D}_2$ to accuracy on a reweighted version of $\mathcal{D}_2$. If there was no bias, and the distribution of selection frequencies for `v1` and `v2` matched exactly, then this term would equal zero ($\mathcal{D}_2|s_1$ would equal $\mathcal{D}_2$). Thus, the selection gap translates the effect of discrepancy in true selection frequency between `v1` and `v2` into a discrepancy in accuracy. Since we measured `v1` as having higher true selection frequency, we expect the selection gap to be positive and thus explain a portion of the accuracy gap that was previously attributed to distribution shift.

Finite-sample error. The final term refers to the finite-sample error from using $10,000$ images as a proxy for distributional accuracy. We believe that this term is negligible, since (a) 95% bootstrapped confidence intervals for the classifiers we evaluate are all at most 0.1%, and (b) there can be no adaptive overfitting on $\mathcal{S}_2$ with respect to $\mathcal{D}_2$. Thus, we drop this term from consideration and instead use $\mathcal{A}_{\mathcal{D}_2}$ and $\mathcal{A}_{\mathcal{S}_2}$ interchangeably.

Computing selection-adjusted accuracy. We have shown

[7] Worker qualification is a service provided by MTurk that only allows "high-reputation" annotators (typically measured by historical annotation quality on the platform) to complete a given task. Qualifications have been shown to significantly impact data quality: in (Peer et al., 2013), using qualifications lowered the number of inattentive workers from 33% to less than 1%.

how to decompose the v1-v2 accuracy gap into a component explained by selection frequency (selection gap), and a component unexplained by selection frequency (bias-corrected accuracy gap). The challenge in computing this decomposition is estimating $\mathcal{A}_{\mathcal{D}_2|s_1}$, the selection-adjusted v2 accuracy. While the closed form of $\mathcal{A}_{\mathcal{D}_2|s_1}$ is

$$\int_s \mathbb{E}_{\mathcal{D}_2}[f(x)|s(x)=s]\cdot p_1(s)\,ds,$$

we have no access to $p_i(s)$ for any value of i (we do not even have direct access to $s(x)$ for any image x). In the next section, we explore methods for estimating $\mathcal{A}_{\mathcal{D}_2|s_1}$ using only the observed selection frequencies that we collected.

5. Quantifying the Bias

In the previous sections, we showed that statistic matching based on noisy observed selection frequencies may lead ImageNet-v2 images to have lower true selection frequencies than expected. In Section 4 we related this discrepancy in selection frequency to a corresponding discrepancy in model accuracy between v1 and v2, which we called the "selection gap." In this section, we explore a series of methods for estimating this gap—we estimate that the selection gap accounts for 8.1% of the 11.7% v1-v2 accuracy drop.

5.1. Naïve approach

We have introduced the selection-adjusted v2 accuracy,

$$\mathcal{A}_{\mathcal{D}_2|s_1} = \int_s \mathbb{E}_{\mathcal{D}_2}[f(x)|s(x)=s]\cdot p_1(s)\,ds, \qquad (3)$$

which captures model accuracy on a version of ImageNet-v2 reweighted to have the same true selection frequency distribution of ImageNet-v1. Since we do not observe true selection frequencies, we cannot evaluate $\mathcal{A}_{\mathcal{D}_2|s_1}$, and are instead forced to estimate it. A natural way to do so is to use observed selection frequencies in place of true ones, leading to the following "naïve estimator:"

$$\hat{\mathcal{A}}^n_{\mathcal{D}_2|s_1} = \sum_{k=0}^{n} \mathbb{E}_2\left[f(x_2)|\hat{s}_n(x_2)=\frac{k}{n}\right]\cdot p_1\left(\hat{s}_n(x_1)=\frac{k}{n}\right). \qquad (4)$$

The naïve estimator is a computable[8] but biased estimator of the selection-adjusted accuracy. This follows from our analysis in Section 2, since $\hat{\mathcal{A}}^n_{\mathcal{D}_2|s_1}$ is just a mechanism for statistic matching between ImageNet-v1 and ImageNet-v2 using observed selection frequencies in place of true selection frequencies. Thus, the selection-adjusted v2 accuracy computed by the naïve estimator is likely to still underestimate the true selection-adjusted accuracy $\mathcal{A}_{\mathcal{D}_2|s_1}$.

We can verify this bias empirically by varying the number of annotators n used to calculate $\hat{s}_n(x)$ for each image, and visualizing the resulting trends in $p_i(\hat{s}_n(x))$ (Figure 5a), $p_i(f(x)=1|s(x))$ (Figure 5b), and $\hat{\mathcal{A}}^n_{\mathcal{D}_2|s_1}$ (Figure 5c). The results corroborate our analysis in Section 2 and our findings from Section 3. Specifically, Figure 5 plots each term in the definition of the naïve estimator,

$$\underbrace{\hat{\mathcal{A}}^n_{\mathcal{D}_2|s_1}}_{\text{Fig. 5c}} = \sum_{k=0}^{n} \underbrace{\mathbb{E}_2\left[f(x_2)|\hat{s}_n(x_2)=\frac{k}{n}\right]}_{\text{Fig. 5b}}\cdot \underbrace{p_1\left(\hat{s}_n(x_1)=\frac{k}{n}\right)}_{\text{Fig. 5a}},$$

and allows us to draw the following conclusions:

- Figure 5a shows that the distribution of observed v1 selection frequencies $p_1(\hat{s}_n(x))$ becomes increasingly skewed as more annotators are used to estimate selection frequencies (i.e. as bias decreases).
- Figure 5b plots selection frequency-conditinoed classifier accuracy, $\mathbb{E}_{x_2\sim\mathcal{D}_2}\left[f(x_2)|\hat{s}_n(x_2)=\frac{k}{n}\right]$ as a function of n. The plot indicates that when we use observed selection frequency in place of true selection frequency, we overestimate model accuracy on images with low selection frequency and underestimate accuracy on images with high selection frequency.
- Combining these two sources of bias, Figure 5c shows that as we reduce bias by increasing n, the selection-adjusted v2 accuracy increases for every classifier.

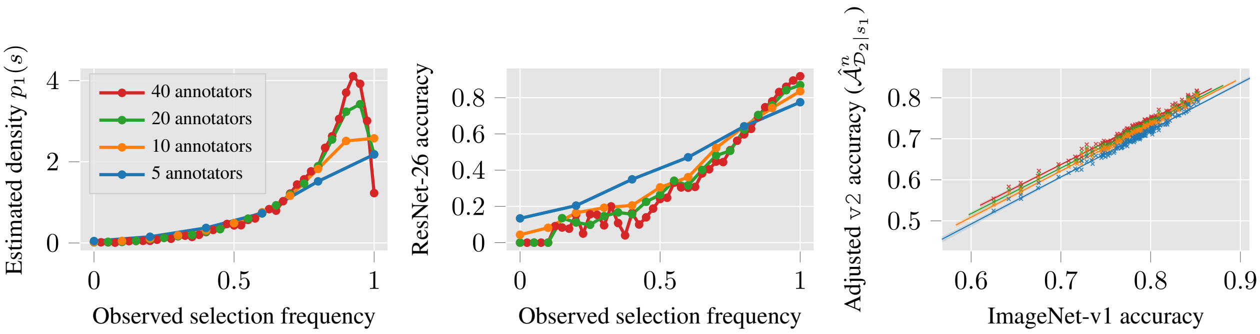
It turns out that computing (4) using the 40 annotators per image that we collected in Section 3 already produces selection-adjusted v2 accuracies that are on average 6.0% higher than the initially observed v2 accuracy. Thus, despite still suffering from matching bias, the naïve reduces the v1-v2 accuracy drop to 5.7%. In the following sections, we explore two different techniques for debiasing the naïve estimator and explaining more of the accuracy gap.

5.2. Estimating bias with the statistical jackknife

As a first attempt at correcting for the previously identified bias, we turn to a standard tool from classical statistics. The jackknife (Quenouille, 1949; Tukey, 1958) is a nonparametric method for reducing the bias of finite-sample estimators. Here, we use it to estimate and correct for the bias in finite-sample estimates of the adjusted accuracy $\mathcal{A}_{\mathcal{D}_2|s_1}$.

Jackknifing the naïve estimator. As a first approach, we can apply the jackknife directly to the naïve estimator (cf. (4)). For the jackknife-corrected estimate to be meaningful, we have to show that the naïve estimator is a statistically consistent estimator of the true selection-adjusted

[8] This is true as long as we can reliably approximate the expectations. Here we have 10^4 images and only 41 possible values of $\hat{s}_n(x)$; also, halving the number of images negligibly affects the value of the estimator.



(a) (b) (c)

Figure 5. A series of graphs, all demonstrating bias in estimators that condition on selection frequency. **Left**: The estimated population density of selection frequencies, calculated naïvely from samples. For a given number of annotators per image n, the corresponding line in the graph has equally spaced points of the form $(k/n, \sum \mathbf{1}_{\hat{s}=k/n})$. **Middle**: Model accuracy of a ResNet-26 conditioned on selection frequency; once again, we naïvely using empirical selection frequency in place of true selection frequency for conditioning. Just as in the left-most graph, for a given n-annotator line, points at $x = k/n$ in the graph correspond to the accuracy on images with observed selection frequency k/n. **Right**: Adjusted `v1` versus `v2` accuracy plots, calculated for varying numbers of annotators per image (with adjusted accuracy computed using the naïve estimator of Section 5.1). Each point in the plot corresponds to a trained model.

accuracy (i.e., that $\lim_{n\to\infty} \hat{\mathcal{A}}^n_{\mathcal{D}_2|s_1} = \mathcal{A}_{\mathcal{D}_2|s_1}$). We prove this property in Appendix E, under the assumption that we can evaluate quantities of the form $p_i(\hat{s}_n(x) = s)$ exactly (in practice this assumption seems acceptable since the empirical variance of the estimator is small). Applying the jackknife to the naïve estimator reduces the adjusted accuracy gap further, from **5.7%** to **4.6%**.

Considerations and limitations. For the jackknife to perform reliably, we must have that (a) the leave-one-out estimators have low enough variance, and (b) the bias is an analytic function in $1/n$ that is dominated by the $\Theta(\frac{1}{n})$ term in its power series expansion. We address the first of these concerns by plotting jackknife confidence intervals (c.f. (Efron & Tibshirani, 1994)) for our estimates. Consideration (b) carries a bit more weight: as shown in App. E.1, the n-sample naïve estimator has a roughly linear relationship in $1/n$, but not a perfect one—in particular, the estimator seems to increase at a rate slightly faster than $1/n$, suggesting that as a result, the jackknife still provides an underestimate of the selection-adjusted accuracy. Another potential source of error is finite-sample error in measuring the expectations $\mathbb{E}_{x_2\sim\mathcal{D}_2}\left[f(x_2)|\hat{s}_n(x_2)\right]$, but as previously mentioned this is likely negligible due to the dataset size and the invariance of the results to the number of images.

In the next section, we present another approach to estimating the selection-adjusted accuracy that relies on a different set of assumptions: parametric modeling.

5.3. Estimating bias with a parametric model

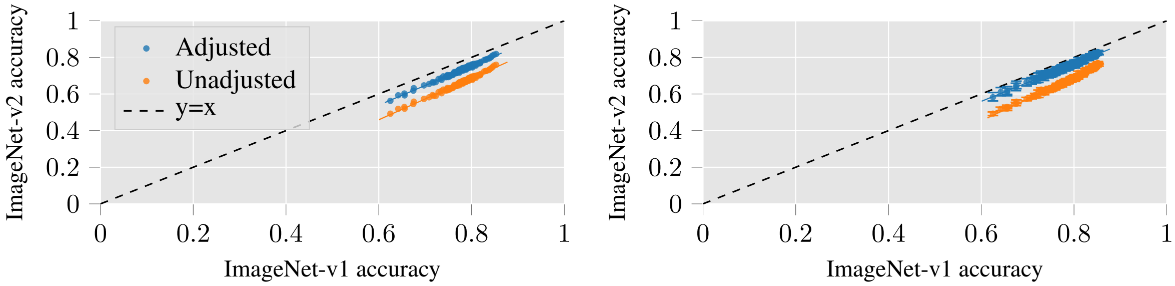
We now explore a more fine-grained approach to estimating the selection-adjusted accuracy of ImageNet-v2, namely explicit parametric modeling. Recall that the adjusted accuracy captures accuracy on ImageNet-v2 reweighted to match ImageNet-v1 in terms of true selection frequency distribution, and is given by:

$$\mathcal{A}_{\mathcal{D}_2|s_1} = \int_{s\in[0,1]} p_2\left(f(x_2)|s(x_2) = s\right) \cdot p_1(s)\, ds \quad (5)$$

In Sections 5.1 we computed a biased estimate of $\mathcal{A}_{\mathcal{D}_2|s_1}$ using observed selection frequencies $\hat{s}$ in place of true selection frequencies. Then, in 5.2 we corrected for bias in the naïve estimator post-hoc using the statistical jackknife.

In constrast, the model-based approach tries to circumvent this bias altogether: we parameterize functions of the true selection frequency directly (i.e., $p_1(s)$ and $p_2(f(x) = 1|s(x) = s)$), then fit parameters that maximize the likelihood of the observed data while taking into account the noise model. For example, since the distribution of $\hat{s}_n(x)$ given $s(x])$ is the binomial distribution, we can write (and optimize) a closed-form expression for the likelihood of observing a given set of selection frequencies based on a parameterized true selection frequency distribution $p_1(s;\theta)$. We estimate selection-adjusted accuracy in two steps. First, we fit models for the true selection frequency distributions $p_1(s)$ and $p_2(s)$. Then, we use our estimate of $p_2(s)$ in conjunction with observed data to fit models for $p_2(f(x)|s(x) = s)$. Finally, we recover estimates for $\mathcal{A}_{\mathcal{D}_2|s_1}$ by numerically computing the integral in $\mathcal{A}_{\mathcal{D}_2|s_1}$ (c.f. (5)), plugging in the learned parametric estimates.

Fitting a model to $p_i(s(x))$. We model the $p_i(s(x))$ as members of a parameterized family of distribution $p_i(s(x);\theta)$ with true parameters $\theta_i^\star$. Then, for each dataset i, we model the observed selection frequencies as sampled



(a) Adjustment via jackknife

(b) Adjustment via mixture model

Figure 6. Accuracy on `v1` versus `v2` adjusted using the two techniques discussed in this section. On the left (respectively, right) we use the jackknife (parametric model) of Section 5.2 (5.3) to estimate adjusted accuracies for `v2`. The graphs confirm that the "true" gap in accuracy between `v1` and `v2` is indeed much smaller than the initially observed gap. Confidence intervals on the left are based on the jackknife standard error, and confidence intervals on the right are based on 400-sample 95% bootstrap confidence intervals

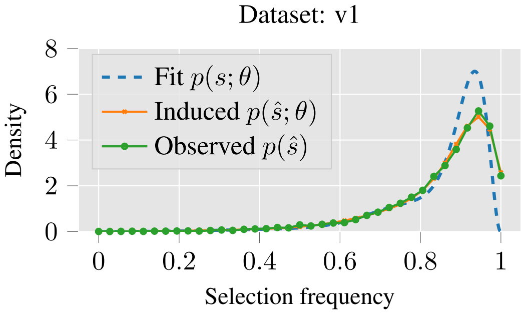


Figure 7. Our fit beta mixture models $p_i(s;\hat{\theta})$ for "true" selection frequency, the noisy selection frequency distribution they induce $p_i(\hat{s};\hat{\theta}$, and the observed selection frequency $p_i(\hat{s})$. The fit $p_i(s;\hat{\theta})$ distributions place more density on higher selection frequencies than naïvely estimating $p_i(s)$ from the observed $p_i(\hat{s})$.

from a compound distribution, in which one first samples $s \sim p_i(\cdot;\theta_i^\star)$, then observes $\hat{s} \sim \text{Binom}\,(n,s)$ (where n is the number of MTurk annotators).

To infer each $\theta_i^\star$, we use maximum likelihood estimation on the observed samples over the compound distribution. We opt to use mixtures of beta distributions as the family $p_i(\cdot;\theta)$ over which to optimize[9], and we use Expectation-Maximization to find, for each dataset, the maximum likelihood mixture of $k = 3$ beta-binomial distributions for the observed $\hat{s}_n(x)$. We provide further detail (including pseudocode) on the fitting process for $p_i(s(x);\theta)$ in Appendix F. We plot the resulting fitted distributions $p_i(s(x))$ in Figure 7. Our estimated $p_i(\hat{s};\hat{\theta}_i)$ distributions continue the trend previously seen in Figure 5a, and show the extent to which our naïve 40-sample empirical estimates of $p_i(s(x))$ exhibit bias.

Fitting a model to $p_2(f(x) = 1|s(x) = s)$**.** Next, we consider accuracy conditioned on selection frequency:

$$g(s) = p_2(f(x_2) = 1|s(x_2) = s).$$

While introducing the naïve estimator (Section 5.1), we found that that estimating $g(s)$ using observed selection frequencies instead of true selection frequencies results in bias (Figure 5b). Under the parametric approach, we instead model $g(s)$ as a member of a parametric class (i.e., $g(s) = g(s;\omega)$), then account for noise in observed selection frequencies via the following identity:

$$p_2(f(x_2) = 1, \hat{s}_n(x_2) = \hat{s}) = \int_{s\in[0,1]} g(s)\cdot p(\hat{s}|s)\cdot p_2(s)\,ds.$$

We parameterize g as a cubic spline, and estimate the parameter by minimizing the squared error between the left and right sides above using a quadratic program solver.

Results. Once we have estimated probability distributions $p_i(s(x);\theta_i)$ and the conditional classification function $g(s(x);\omega)$, we can compute an estimate of $\mathcal{A}_{\mathcal{D}_2|s_1}$ using Equation (5) and numerical integration. Figure 6b depicts various models' `v1` and `v2` accuracies both with and without the adjustment for selection frequency. Our estimate for the frequency-adjusted gap in accuracy averaged over all models is $3.6\% \pm 1.5\%$, around 30% of the original $11.7\% \pm 1.0\%$ gap in accuracy.

Beyond accuracy gap, Recht et al. (2019b) also studied the linear relationship between `v2` accuracy and `v1` accuracy while varying the classifier used—this is plotted by the blue dots in Figure 6b. This relationship is linear for our adjusted accuracies as well (cf. Figure 6b), however the slope we find is 1.01 ± 0.09 instead of 1.13 ± 0.05.

Considerations and limitations. Error in parametric modeling generally stems from two sources: finite-sample error and model misspecification. These sources of error affect

[9] A beta distribution composed with a binomial is a *beta-binomial* distribution—the basis of beta-binomial regression.

all parametric models, but we take various precautions to mitigate their impact on our estimates. To assess our finite-sample error, we give 95% bootstrapped confidence intervals (details are in Appendix F), which are displayed as error bars in Figure 6b. We also ensure that our results are not sensitive to the number of annotators used to fit the parametric models (cf. Appendix F). As with any modeling decision, our choices of model classes might not fully capture the ground-truth, and thus may be a source of error. We account for this as much as possible by demonstrating the robustness of our results to varying the number of free parameters (cf. Appendix F).

6. Related Work

Rapid improvements on standard datasets (e.g. (LeCun, 1998; Krizhevsky, 2009; Russakovsky et al., 2015; Zhou et al., 2017) in computer vision) has drawn interest to verifying and testing the robustness of progress thus far. Previous work has characterized cross-dataset generalization(Torralba & Efros, 2011), and explored the impact of synthetic perturbations on generalization, such as adversarial examples (Kurakin et al., 2016; Tsipras et al., 2019; Ilyas et al., 2019; Su et al., 2018) or various other corruption robustness measures (Hendrycks & Dietterich, 2019; Kang et al., 2019). Recently, a number of works have emerged around evaluating performance on newly reproduced test sets, including works focusing on ImageNet (Recht et al., 2019b) and CIFAR (Recht et al., 2019a). In our work, we study a source of statistical bias that may affect such dataset replication. Similar phenomena have been noted in the context of the natural sciences (e.g., ecology (Greig-Smith, 1983)) and causal inference (Stipak & Hensler, 1982).

7. Discussion and Conclusions

Dataset replication pipelines can introduce unforeseen, often unintuitive statistical biases. In the case of ImageNet-v2, even using unbiased estimates of image selection frequency in the data generation pipeline results in a significant statistical bias, and ultimately turns out to account for a large portion of the observed accuracy drop. Our findings give rise to the following considerations.

7.1. Remaining accuracy gap and unmodeled bias

Worker heterogeneity. Our study focuses on bias stemming from the fact that for a given image x one never observes $s(x)$ but rather $\hat{s}_n(x) = \text{Binom}\,(n, s(x))$. There is another source of bias due to noise that we do not model here, namely variance in the MTurk annotator population. Specifically, some annotators are more likely in general to select or reject independently of what image-label pair they are being shown. This unmodeled variance likely translates to unmodeled bias, suggesting that more of the gap might be explained by taking worker heterogeneity into account.

Task shift bias. At the time of the original ImageNet experiment, workers judged image-label pairs by some abstract set of criteria C_1. Suppose that at the time of the ImageNet-v2 experiment several years later, annotators judged image-label pairs based on an overlapping but non-identical set of criteria C_2. Ideally, we should not care about differences between C_1 and C_2—indeed, one of the goals of dataset replication is to test robustness to such benign distribution shifts. The source of the bias lies in the iterated nature of the filtering experiment. In particular, after both the original experiment and the replication, images in ImageNet-v1 now meet both C_1 and C_2. On the other hand, images in ImageNet-v2 only meet criteria C_2, and may be judged to have low selection frequency under C_1—we would thus expect models to perform better on ImageNet-v1 images due to their increased qualifications. Although this may contribute towards the remaining accuracy gap, this type of bias is difficult to study or correct for without more knowledge of both experiments.

Other sources of error. The remaining error unexplained by bias in data collection could come from one of the gap sources listed in Section 4, i.e., finite sample error, or distribution shift and adaptive overfitting. Quantifying the potential contribution of the individual terms in the remaining gap will require more experimentation and future work.

7.2. Adaptive overfitting and distribution shift

Identifying sources of distribution shift. A longstanding goal in computer vision is to develop models that are less prone to failure under small distributional shifts. A step in the journey towards this goal is precisely characterizing the kinds of distribution shifts under which models fail—examples include rotations and translations of natural images Engstrom et al. (2019), or corrupted natural images (Hendrycks & Dietterich, 2019). Our findings imply that the drop may be attributable to differences in selection frequency distribution, corroborating observations by Recht et al. (2019b) that models are sensitive to selection frequency. Differences in selection frequency distribution present another distribution shift to study in depth.

Detecting and avoiding bias in dataset replication. More broadly, our analysis identifies statistical modeling of the data collection pipeline as a useful tool for dataset replication. Indeed, characterizing the ImageNet and ImageNet-v2 generative processes and isolating them in a a simple theoretical model allowed for the discovery and correction of a source of bias in the dataset replication process.

Acknowledgements

We thank Will Fithian for discussions and advice, particularly around the spline modeling done in Section 5. We also thank Nicholas Carlini, Zachary Lipton, Kevin Foley, and Michael Yang for helpful comments on early drafts of this paper.

Work supported in part by the NSF grants CCF-1553428, CNS-1815221, the Google PhD Fellowship, the Open Phil AI Fellowship, and the Microsoft Corporation.

Research was sponsored by the United States Air Force Research Laboratory and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Bell, D. S. The experimental reproduction of amphetamine psychosis. *Archives of General Psychiatry*, 29(1):35–40, 1973.

Buhrmester, M., Kwang, T., and Gosling, S. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? In *Perspectives on Psychological Science*, 2011.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *computer vision and pattern recognition (CVPR)*, 2009.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, 2015.

Efron, B. and Tibshirani, R. *An Introduction to the Bootstrap*. 1994.

Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., and Madry, A. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning (ICML)*, 2019.

Greig-Smith, P. *Quantitative plant ecology*, volume 9. Univ of California Press, 1983.

Hendrycks, D. and Dietterich, T. G. Benchmarking neural network robustness to common corruptions and surface variations. In *International Conference on Learning Representations (ICLR)*, 2019.

Ilyas, A., Santurkar, S., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Imai, K. and Ratkovic, M. Covariate balancing propensity score. In *Journal of the Royal Statistical Society*, 2013.

Kang, D., Sun, Y., Hendrycks, D., Brown, T., and Steinhardt, J. Testing robustness against unforeseen adversaries. In *ArXiv preprint arxiv:1908.08016*, 2019.

Krizhevsky, A. Learning multiple layers of features from tiny images. In *Technical report*, 2009.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

LeCun, Y. The mnist database of handwritten digits. In *Technical report*, 1998.

Mason, W. and Watts, D. Financial incentives and the "performance of crowds". In *KDD Human Computation Workshop*, 2009.

Paolacci, G., Chandler, J., and Ipeirotis, P. Running experiments on Amazon Mechanical Turk. In *Judgement and Decision Making*, 2010.

Peer, E., Vosgerau, J., and Acquisti, A. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. In *Behavior Research Methods*, 2013.

Quenouille, M. Problems in plane sampling. In *Annals of Mathematical Statistics*, 1949.

Rao, R. B., Fung, G., and Rosales, R. On the dangers of cross-validation. an experimental evaluation. In *SIAM International Conference on Data Mining (ICDM)*, pp. 588–596, 2008.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do cifar-10 classifiers generalize to cifar-10? In *International Conference on Machine Learning (ICML)*, 2019a.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning (ICML)*, 2019b.

Reunanen, J. Overfitting in making comparisons between variable selection methods. In *Journal of Machine Learning Research*, volume 3, pp. 1371–1382, 2003.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. In *International Journal of Computer Vision (IJCV)*, 2015.

Schuman, H. and Presser, S. *Questions and Answers in Attitude Surveys*. 1981.

Stipak, B. and Hensler, C. Statistical inference in contextual analysis. In *American Journal of Political Science*, 1982.

Stuart, E. A. Matching methods for causal inference: A review and a look forward. In *Statistical Science*, 2010.

Su, D., Zhang, H., Chen, H., Yi, J., Chen, P.-Y., and Gao, Y. Is robustness the cost of accuracy? a comprehensive study on the robustness of 18 deep image classification models. In *European Conference on Computer Vision (ECCV)*, 2018.

Torralba, A. and Efros, A. A. Unbiased look at dataset bias. In *CVPR 2011*, 2011.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. In *International Conference on Learning Representations (ICLR)*, 2019.

Tukey, J. Bias and confidence in not quite large samples. In *Annals of Mathematical Statistics*, 1958.

Yadav, C. and Bottou, L. Cold case: The lost mnist digits. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2017.