From ImageNet to Image Classification: Contextualizing Progress on Benchmarks

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

Abstract

Building rich machine learning datasets in a scalable manner often necessitates a crowd-sourced data collection pipeline. In this work, we use human studies to investigate the consequences of employing such a pipeline, focusing on the popular ImageNet dataset. We study how specific design choices in the ImageNet creation process impact the fidelity of the resulting dataset—including the introduction of biases that state-of-the-art models exploit. Our analysis pinpoints how a noisy data collection pipeline can lead to a systematic misalignment between the resulting benchmark and the real-world task it serves as a proxy for. Finally, our findings emphasize the need to augment our current model training and evaluation toolkit to take such misalignments into account. [1]

1. Introduction

Large-scale vision datasets and their associated benchmarks (Everingham et al., 2010; Deng et al., 2009; Lin et al., 2014; Russakovsky et al., 2015) have been instrumental in guiding the development of machine learning models (Krizhevsky et al., 2012; Szegedy et al., 2016; He et al., 2016). At the same time, while the progress made on these benchmarks is undeniable, they are only proxies for real-world tasks that we actually care about—e.g., object recognition in the wild. Thus, it is natural to wonder:

How aligned are existing benchmarks with their motivating real-world tasks?

On one hand, significant design effort goes into ensuring that these benchmarks accurately model real-world challenges (Ponce et al., 2006; Torralba & Efros, 2011; Everingham et al., 2010; Russakovsky et al., 2015). On the other hand, the sheer size of machine learning datasets makes meticulous data curation impossible. Dataset creators thus resort to scalable methods such as automated data retrieval and crowd-sourced annotation (Everingham et al., 2010; Russakovsky et al., 2015; Lin et al., 2014; Zhou et al., 2017), often at the cost of faithfulness to the task being modeled. As a result, the dataset annotations obtained can sometimes be ambiguous, incorrect, or otherwise misaligned with the ground truth (cf. Figure 1). Still, despite our awareness of these issues (Russakovsky et al., 2015; Recht et al., 2019; Stock & Cisse, 2018; Hooker et al., 2019; Northcutt et al., 2019), we lack a precise characterization of their pervasiveness and impact, even for widely-used datasets.

(a) missile

(b) stage

(c) monastery

Figure 1: Judging the correctness of ImageNet labels may not be straightforward. While the labels shown above appear valid, *none* of them match the ImageNet labels (respectively "projectile", "acoustic guitar", and "church").

Our contributions

We develop a methodology for obtaining fine-grained data annotations via large-scale human studies. These annotations allow us to precisely quantify ways in which typical object recognition benchmarks fall short of capturing the underlying ground truth. We then study how such *benchmark-task misalignment* impacts state-of-the-art models—after all, models are often developed by treating existing datasets as the ground truth. We focus our exploration on the ImageNet dataset (Deng et al., 2009) (specifically, the ILSVRC2012 object recognition task (Russakovsky et al., 2015)), one of the most widely used benchmarks in computer vision.

Quantifying benchmark-task alignment. We find that systematic annotation issues pervade ImageNet, and can

[*]Equal contribution [1]EECS, MIT. Correspondence to: DT <tsipras@mit.edu>, SS <shibani@mit.edu>, LE <engstrom@mit.edu>, AI <ailyas@mit.edu>, AM <madry@mit.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

[1]We release our refined ImageNet annotations at `https://github.com/MadryLab/ImageNetMultiLabel`.

often be attributed to design choices in the dataset collection pipeline itself. For example, during the ImageNet labeling process, annotators were not asked to classify images, but rather to validate an automatically-obtained candidate label without knowledge of other dataset classes. This leads to:

- *Multi-object images* (Section 4.1): While each image is associated with a single label, we find that more than one fifth of ImageNet images contain objects from *multiple* classes. In fact, the dataset label often does not even correspond to what humans deem the "main object". Nevertheless, models still achieve significantly-better-than-chance prediction performance on these images, indicating that they must exploit idiosyncrasies of the dataset that humans are oblivious to.
- *Bias in label validation* (Section 4.2): Even when there is only one object in an image, collectively, annotators often end up validating several mutually exclusive labels. These correspond, for example, to images that are ambiguous or to classes with synonymous labels. The ImageNet label in these cases is determined not by annotators themselves, but rather by the fidelity of the automated image retrieval process. In general, given that annotators are ineffective at filtering out errors under this setup, the automated component of the data pipeline may have a disproportionate impact on the quality of ImageNet annotations.

More broadly, these issues point to an inherent tension between the goal of building large-scale benchmarks that are realistic and the scalable data collection pipelines needed to achieve this goal. Hence, for benchmarks created using such pipelines, the current standard for model evaluation—accuracy with respect to a single, fixed dataset label—may be insufficient to correctly judge model performance.

Human-based performance evaluation. In light of this, we use our annotation pipeline to directly measure human-model alignment. We find that more accurate ImageNet models also make predictions that annotators are more likely to agree with. In fact, we find that models have reached a level where non-expert annotators are largely unable to distinguish between predicted labels and ImageNet labels (i.e., even model predictions that don't match the dataset labels are often judged valid by annotators). While reassuring, this finding highlights a different challenge: non-expert annotations may no longer suffice to tell apart further progress from overfitting to idiosyncrasies of the ImageNet distribution.

2. A Closer Look at the ImageNet Dataset

We start by briefly describing the original ImageNet collection and annotation process. As it turns out, several details of this process significantly impact the resulting dataset.

The ImageNet creation pipeline. ImageNet is a prototypical example of a large-scale dataset (1000 classes and more than million images) created through automated data collection and crowd-sourced filtering. At a high level, this creation process comprised two stages (Deng et al., 2009):

1. *Image and label collection:* The ImageNet creators first selected a set of classes using the WordNet hierarchy (Miller, 1995). Then, for each class, they sourced images by querying several search engines with the WordNet synonyms of the class in several languages. Note that for each of the retrieved images, the *proposed label*—that will be assigned to this image should it be included in the dataset, is already determined. That is, the label is simply given by the WordNet node that was used for the corresponding search query.
2. *Image validation via the* CONTAINS *task.* To validate the image-label pairs retrieved in the previous stage, the ImageNet creators employed annotators via the Mechanical Turk (MTurk) crowd-sourcing platform. Specifically, for every class, annotators were presented with its description (along with links to the relevant Wikipedia pages) and a grid of candidate images. Their task was then to select all images in that grid that contained an object of that class (with *explicit* instructions to ignore clutter and occlusions). The grids were shown to multiple annotators and only images that received a "convincing majority" of votes (based on per-class thresholds estimated using a small pool of annotators) were included in ImageNet. In what follows, we will refer to this filtering procedure as the CONTAINS task.

Revisiting the ImageNet labels

The process described above is a natural method for creating a large-scale dataset, especially if it involves a wide range of classes. However, even putting aside occasional annotator errors, the resulting dataset might not accurately capture the ground truth. Indeed, as we discuss below, this pipeline design itself can lead to certain *systematic* errors in the dataset. The root cause for many of these is that the image validation stage (i.e., the CONTAINS task) only asks annotators to verify if a *specific* label (i.e., WordNet node for which the image was retrieved), shown in isolation, is valid for a given image. Crucially, annotators are never asked to choose among different possible labels for the image and, in fact, have no knowledge of what the other classes even are. This can introduce discrepancies in the dataset in two ways:

Images with multiple objects. Annotators are instructed to *ignore* the presence of other objects when validating a particular ImageNet label for an image. However, these objects could themselves correspond to other ImageNet classes. This can lead to the selection of images with multiple valid

labels or even to images where the dataset label does not correspond to the most prominent object in the image.

Biases in image filtering. Since annotators have no knowledge other classes, they do not have a sense of the granularity of image features they should pay attention to (e.g., the labels in Figure 1 appear reasonable until one becomes aware of the other ImageNet classes). Moreover, the task itself does not necessary account for their expertise. Indeed, one cannot reasonably expect non-experts to distinguish, e.g., between all the 24 terrier breeds that are present in ImageNet. As a result, if annotators are shown images containing objects of a different, yet similar class, they are likely to select them as valid. This implies that potential errors in the collection process (e.g., automated search retrieving images that do not match the query label) are unlikely to be corrected during validation (via the CONTAINS task) and thus can propagate to the final dataset.

In the light of the above, it is clear that eliciting ground truth information using the ImageNet creation pipeline may not be straightforward. In Section 3, we present a framework for improving this elicitation and then, in Section 4, we use that framework to investigate the discrepancies highlighted above and their impact on ImageNet-trained models.

3. From Label Validation to Image Classification

We begin our study by obtaining a better understanding of the ground truth for ImageNet data. To achieve this, rather than asking annotators to *validate* a single proposed label for an image (as in the original pipeline), we would like them to *classify* the image, selecting *all* the relevant labels for it. However, asking (untrained) annotators to choose from among all 1,000 ImageNet classes is infeasible.

To circumvent this difficulty, our pipeline consists of two phases, illustrated in Figure 2. First, we obtain a small set of *candidate labels* for each image (Section 3.1). Then, we present these labels to annotators and ask them to select one of them for *each* distinct object using what we call the CLASSIFY task (Section 3.2). These phases are described in detail below, with additional information in Appendix B.

For our analysis, we use 10,000 images from the ImageNet validation set—i.e., 10 randomly selected images per class. Note that since the ImageNet training and validation sets were created using the same procedure, analyzing the latter is sufficient to understand systematic issues in that dataset.

3.1. Obtaining candidate labels

To ensure that the per-image annotation task is manageable, we narrow down the candidate labels to a small set. To this end, we first obtain *potential labels* for each image by simply combining the top-5 predictions of 10 models from different parts of the accuracy spectrum with the existing ImageNet label (yields approximately 14 labels per image; cf. Appendix Figure 11). Then, to prune this set further, we reuse the ImageNet CONTAINS task—asking annotators whether an image contains a particular class (Section 2)—but for *all* potential labels. The outcome of this experiment is a *selection frequency* for each image-label pair, i.e., the fraction of annotators that selected the image as containing the corresponding label.[2] We find that, although images were often selected as valid for many labels, relatively few of these labels had high selection frequency (typically less than five per image). Thus, restricting potential labels to this smaller set of *candidate labels* allows us to hone in on the most likely ones, while ensuring that the resulting annotation task is still cognitively tractable.

3.2. Image classification via the CLASSIFY task

Once we have identified a small set of candidate labels for each image, we present them to annotators to obtain fine-grained label information. Specifically, we ask annotators to identify: (a) *all* labels that correspond to objects in the image, and (b) the label for the *main* object (according to their judgment). Crucially, we explicitly instruct annotators to select only *one label per distinct object*—i.e., in case they are confused about the correct label for a specific object, to pick the one they consider most likely. Moreover, since ImageNet contains classes that could describe parts or attributes of a single physical entity (e.g., "car" and "car wheel"), we ask annotators to treat these as distinct objects, since they are not mutually exclusive. We present each image to multiple annotators and then aggregate their responses (per-image) as described below. In the rest of the paper, we refer to this annotation setup as the CLASSIFY task.

Identifying the main label and number of objects. From each annotator's response, we learn what they consider to be the label of the main object, as well as how many objects they think are present in the image. By aggregating these two quantities based on a majority vote over annotators, we can get an estimate of the number of objects in the image, as well as of the main label for that image.

Partitioning labels into objects. Different annotators may choose different labels for the same object and thus we need to map their selections to a single set of distinct objects. To illustrate this, consider an image of a soccer ball and a terrier, where one annotator has selected "Scotch

[2]Note that this notion of selection frequency (introduced by Recht et al. (2019)) essentially mimics the majority voting process used to create ImageNet (cf. Section 2), but using a fixed number of annotators per grid instead of an adaptive process.

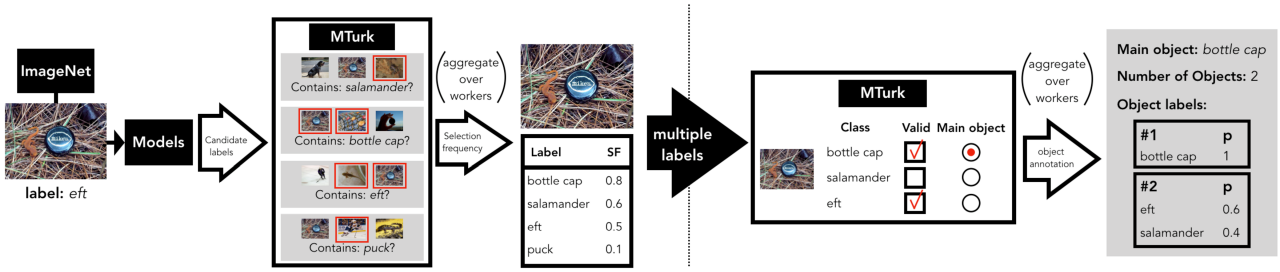


Figure 2: Overview of our data annotation pipeline. First, we collect potential labels for each image using models prediction (Section 3.1). Then, we ask annotators to gauge the validity of each label (in isolation) using the CONTAINS task (described in Section 2). Next, we present all highly selected labels for each image to a new set of annotators, asking them to select *one* label for every distinct object, as well as a label for the main object according to their judgement, i.e., the CLASSIFY task (Section 3.2). Finally, we aggregate their responses to obtain fine-grained image annotations (Section 3.2).

terrier" and "soccer ball" and another "Toy terrier" and "soccer ball". We would like to partition these selections into objects as ["soccer ball"] and ["Scotch terrier" or "Toy terrier"] since both responses indicate that there are two objects in the image and that the soccer ball and the terrier are distinct objects. More generally, we would like to partition selections in a way that avoids grouping labels together if annotators identified them as distinct objects. To this end, we employ exhaustive search to find a partition that optimizes for this criterion (since there exist only a few possible partitions to begin with). Finally, we label each distinct object with its most frequently selected label.

The resulting annotations characterize the content of an image in a more fine-grained manner compared to the original ImageNet labels. Note that these annotations may still not perfectly match the ground truth. After all, we also employ untrained, non-expert annotators, that make occasional errors, and moreover, some images are inherently ambiguous without context (e.g., Figure 1c). Nevertheless, as we will see, these annotations are already sufficient for our study.

4. Quantifying the Benchmark-Task Alignment of ImageNet

Our goal in this section is two-fold. First, we want to use our refined image annotations (cf. Section 3) to examine potential sources of discrepancy between ImageNet and the motivating object recognition task.[3] Next, we want to assess the impact of these deviations on models developed using ImageNet benchmark. To this end, we will measure how the accuracy of a diverse set of models (see Appendix A.2 for a list) is affected when they are evaluated on different subpopulations of images our annotations enable us to identify.

[3]Since we primarily focus on systematic issues in ImageNet, we defer discussing clear mislabelings to Appendix C.3.

4.1. Multi-object images

We start by taking a closer look at images which contain objects from more than one ImageNet class—how often these additional objects appear and how salient they are. (Recall that if two labels are both simultaneously valid for an image, that is they are not mutually exclusive—e.g., "car" and "car wheel"—we refer to them as two objects.) We find that more than a fifth of the images contain *at least two* objects, and, in fact, multiple class pairs *consistently co-occur* (see Appendix Figure 15). This indicates that multi-object images in ImageNet are not caused solely by irrelevant clutter, but also arise from the dataset class selection. Even though, in principle, ImageNet classes correspond to distinct objects, some of these objects can overlap greatly in terms of how they occur in the real world.

Model accuracy on multi-object images. Model performance is typically measured using (top-1 or top-5) accuracy with respect to a *single* ImageNet label, treating it as the ground truth. However, it is not clear what the right notion of ground truth annotation even is when classifying multi-object images. Indeed, we find that models perform significantly worse on multi-label images based on top-1 accuracy (measured w.r.t. ImageNet labels): accuracy drops by more than 10% across all models—see Figure 3.

In light of this, a more natural notion of accuracy for multi-object images would be to consider a model prediction to be correct if it matches the label of *any* object in the image. On this metric, we find that the aforementioned performance drop essentially disappears—models perform similarly on single- and multi-object images (see Figure 4). This indicates that standard accuracy, measured w.r.t. single label, can be overly pessimistic. Note that while top-5 accuracy also accounts for many multi-object confusions, which was after all its original motivation (Russakovsky et al., 2015), it tends to inflate accuracy on *single object* images, by treating

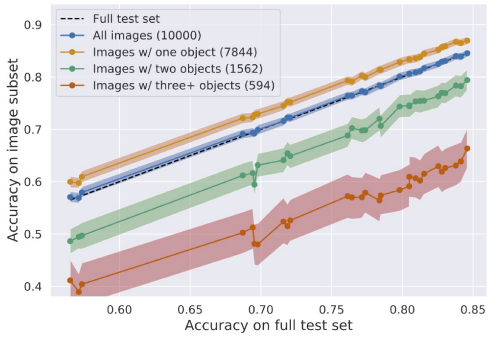


Figure 3: Top-1 model accuracy on multi-object images (as a function of overall test accuracy). Accuracy drops across all models. Confidence intervals: 95% via bootstrap.

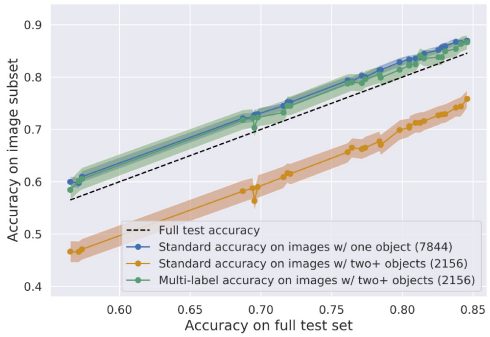


Figure 4: Evaluating multi-label accuracy on ImageNet: the fraction of images where the model predicts the label of *any* object in the image. Based on this metric, the performance gap between single- and multi-object images virtually vanishes. Confidence intervals: 95% via bootstrap.

several erroneous predictions as valid (cf. Appendix C.1).

Human-label disagreement. Although models suffer a sizeable accuracy drop on multi-object images, they are still relatively good at predicting the ImageNet label—much better than choosing the label of one object at random. This bias could be justified whenever there is a distinct main object in the image and that object corresponds to the ImageNet label. However, we find that for nearly a third of the multi-object images, the ImageNet label does *not* denote the most likely main object as judged by human annotators. Nevertheless, model accuracy (w.r.t. the ImageNet label) on these images is still high—see Figure 5.

On these samples, models must thus base their predictions on features that humans do not consider salient. For instance, we find that these disagreements often arise when the ImageNet label corresponds to a very distinctive object (e.g., "pickelhaube"), but the image also contain another more prominent object (e.g., "military uniform")—see Figure 6. To do well on these classes, the model likely picks

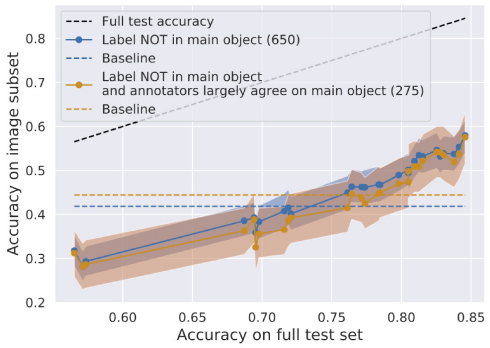


Figure 5: Model accuracy on images where the annotator-selected main object does not match the ImageNet label (650 out of 2156 multi-object images, see examples in Appendix Figure 18 for additional samples.

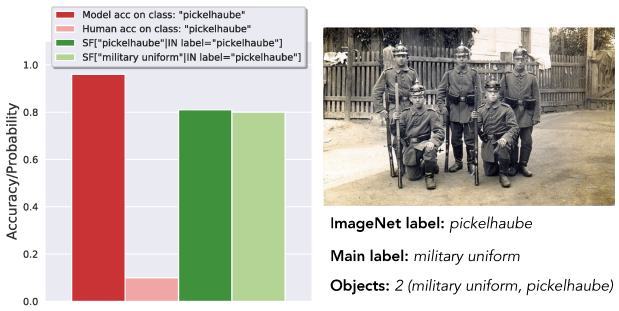


Figure 6: Example of a class where humans disagree with the label as to the main object, yet models still predict the ImageNet label. Here, for images of that class, we plot both model and annotator accuracy as well as the selection frequency (SF) of both labels.

up on ImageNet-specific biases, e.g., detecting that military uniforms often occur in images of class "pickelhaube" but not vice versa. While this might be a valid strategy for improving ImageNet accuracy, it causes models to rely on features that may not generalize to the real world.

4.2. Bias in label validation

We now turn our attention to assessing the effectiveness of the ImageNet filtering process (see Section 2). Our goal is to understand how likely annotators are to detect mislabeled images during the label validation stage (i.e., the CONTAINS task). Recall that annotators are asked a somewhat leading question, of whether a specific (pre-determined) label is valid for a given image. This could make them prone to answer positively even for images of a different, yet similar, class. In order to understand how likely annotators are to validate such mislabeled images, we need to repeat the CONTAINS task multiple times per image, each time presenting it to a new group of annotators under a different label.

Since this exact process is part of our annotation pipeline Section 3.1, we can analyze these results directly.

We find that often, under this task setup, annotators *collectively* deem multiple labels as valid for a single image—i.e., labels *other that the ImageNet label* are also validated by an independent group of annotators. In fact, for nearly 40% of the images, there exists another label that is selected as valid (in isolation) at least as often as the ImageNet label (cf. Figure 7a). Crucially, this does not only occur when multiple objects are present in the image—annotators often validate as many as 10 classes in isolation even for single-object images (cf. Figure 22a). Thus, even for images where a single ground truth label exists, the ImageNet filtering pipeline may fail to elicit this label from annotators.[4]

Moreover, we find that this confusion is not just a consequence of annotator non-expertise, but also of the task setup itself. If instead of asking annotators to judge the validity of a label in isolation, we ask them to choose among all labels simultaneously (i.e., via the CLASSIFY task), they select substantially fewer labels—see Appendix Figure 22b.

These findings highlight how sensitive annotators are to seemingly insignificant aspects of the data collection pipeline. It also indicates that ImageNet labels may not have be vetted as carefully as one might expect. Annotators might have been unable to correct errors, and consequently, ImageNet labels may be determined, to a large extent, by the fidelity (and biases) of the automated retrieval process.

Confusing class pairs. We find that there are several pairs of ImageNet classes that annotators have trouble telling apart. When asked to judge each label in isolation, independent groups of annotators deem both labels as valid for images of either ImageNet class—Appendix Figure 23. On some of these classes, models still perform quite well—likely because the search results from the automated image retrieval process do not overlap significantly. However, on others, even state-of-the-art models have poor accuracy (below 40%)—see Figure 7b. In fact, we can attribute this poor performance to model confusion *within* these pairs—*none* of the models we examined can distinguish these classes much better than chance.

The apparent performance barrier on these classes could be due to an inherent overlap in their image distributions within ImageNet. It is likely that automated image retrieval caused mixups in the images of the two classes, which went undetected during data filtering. We find that several of these classes are semantically similar—e.g., "rifle" and "assault rifle"—making annotators prone to validate incorrect images in the CONTAINS task. In some cases, we also identify errors in the annotation pipeline—overlaps in class names (e.g., "maillot" and "maillot, tank suit") and Wikipedia links (e.g., "laptop" vs. "notebook" computer).

Overall, the existence of such *ambiguous classes* highlights that choosing labels that are in principle disjoint (e.g., using WordNet) is not sufficient to ensure that the resulting dataset has non-overlapping classes—when using noisy validation pipelines, we need to factor human confusion into class selection and description. Further, given that ImageNet already contains such overlapping classes, it is natural to wonder whether accuracy on these classes can be improved (without overfitting to the test set idiosyncrasies).

5. Beyond Test Accuracy: Human-In-The-Loop Model Evaluation

Our analysis of the ImageNet dataset so far makes it clear that using top-1 accuracy as a standalone performance metric can be problematic—issues such as multi-object images and ambiguous classes make ImageNet labels an imperfect proxy for the ground truth. Taking these issues into consideration, we now turn our focus to augmenting the model evaluation toolkit with metrics that are better aligned with the underlying goal of object recognition.

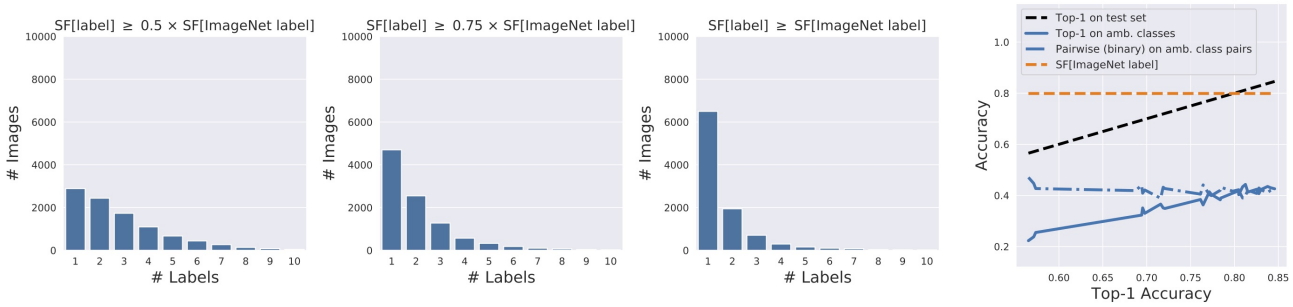
5.1. Human assessment of model predictions

To gain a broader understanding of model performance we start by *directly* employing annotators to judge model predictions. Concretely, we want to understand whether more accurate models make higher-quality predictions, i.e., if the labels they predict (including the erroneous ones) also appear more reasonable to humans. Intuitively, this would capture improvements in models that might not be reflected in accuracy alone (e.g., predicting an incorrect, yet similar animal breed), while also accounting for imperfections in ImageNet labels. Specifically, given a model prediction for a specific image, we measure:

- **Selection frequency of the prediction:** How often annotators deem the predicted label as being present in the image. We can compute these selection frequencies as we repeated the CONTAINS task using model predictions as the proposed image label. This metric accommodates for multi-object images or ambiguous classes as annotators will confirm any valid label.

- **Accuracy based on main label annotation:** How frequently the prediction matches the main label for the image, as determined using on the CLASSIFY task.

[4] Note that our selection frequency estimates for ImageNet labels may be biased (underestimates) (Engstrom et al., 2020) as these specific pairs have already been filtered during dataset creation based on their selection frequency. However, we can effectively ignore this bias since: a) our results are robust to varying the number of annotators (Appendix C.2), b) most of our results are based on the CLASSIFY task for which this bias does not apply.



(a) (b)

Figure 7: (a) Number of labels deemed valid (based on varying thresholds) by independent groups of annotators for a single image under the ImageNet filtering pipeline. For more than 70% of images, annotators collectively select another label as valid at least half as often as they select the ImageNet label (*leftmost*). (b) Model progress on ambiguous class pairs has been largely stagnant—possibly due to substantial overlap in the image distributions of these classes within ImageNet. In fact, models are unable to distinguish between these pairs better than chance (cf. pairwise accuracy).

This metric penalizes models that exploit dataset biases to predict ImageNet labels even when these do not correspond to the most prominent image object. At the same time, it only measures accuracy as non-experts perceive it—if annotators cannot distinguish between two classes (e.g., different dog breeds), models can do no better than random chance on this metric, even if their predictions actually match the ground truth.

Contextualizing model progress. We start by comparing models with varying top-1 accuracy based on these metrics. We find that models have been consistently improving on this axis (cf., Figure 8), more than what accuracy alone would explain (i.e., more predictions matching the ImageNet label). Crucially, the predictions of state-of-the-art models have, on average, gotten quite close to ImageNet labels. Annotators are almost *equally likely* to select the predicted label as valid (or as the main object) for the image as the ImageNet label. This indicates that model predictions might be closer to what non-expert annotators can recognize as the ground truth than accuracy alone suggests.

This does not imply, however, that further improvements in ImageNet accuracy will not translate into progress on the underlying task. After all, for many images, ImageNet labels could capture the ground truth and annotators simply lack the expertise to make such distinctions. However, the results in Figure 8 hint at a different issue: we may no longer by able to easily identify (e.g., using crowd-sourcing) the extent to which further gains in accuracy correspond to actual improvements, as opposed to models simply fitting to specifics of the ImageNet distribution.

Incorrect predictions. We can also use these metrics to examine model mistakes, i.e., predictions that deviate from the ImageNet label. Specifically, we can treat human assessment of these labels (w.r.t. the metrics above) as a proxy for how much these predictions deviate from the ground truth.

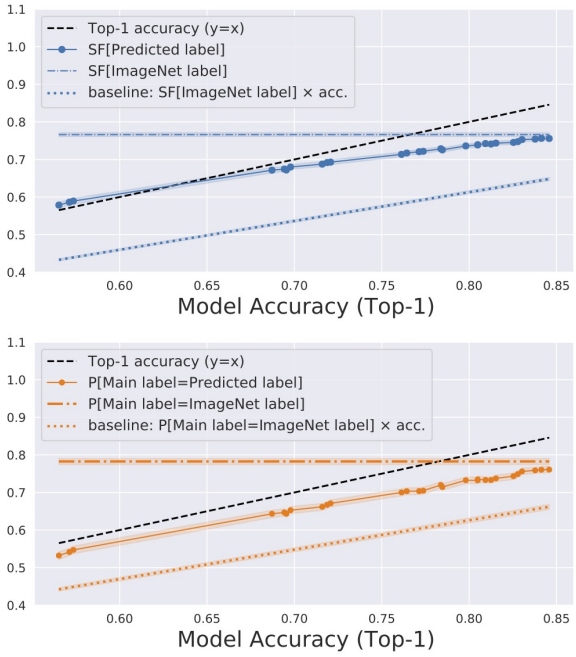


Figure 8: Annotator assessment of model predictions—we measure how often annotators select the predicted/ImageNet label: (*top*) as contained in the image (*selection frequency* [SF] from Section 3.1); (*bottom*) to denote the *main* image object (cf. Section 3.2) (shading denotes 95% confidence intervals via bootstrapping). Even though state-of-the-art models have imperfect top-1 accuracy, their predictions are, on average, almost indistinguishable according to (non-expert) annotators from the ImageNet labels themselves.

We observe that more accurate ImageNet models make progressively fewer mistakes that would be judged by humans as such (i.e., with low selection frequency)—see Figure 9. At the same time, for *all* models, a large fraction of the incorrect predictions are actually valid according to annotators, potentially corresponding to multi-object images or ambiguous classes. From a different perspective, this also highlights the pitfalls of using selection frequency as the sole filtering criterion during dataset creation. Even images which high selection frequency (w.r.t. dataset label) may be ambiguous, posing unintended challenges for models (cf. Appendix Figure 25).

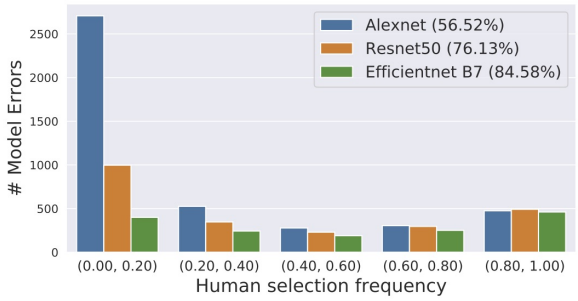


Figure 9: Distribution of annotator selection frequencies (cf. Section 3.1) for model predictions deemed incorrect w.r.t. the ImageNet label. Models that are more accurate also make fewer mistakes with low annotator selection frequency (for the corresponding image-label pair).

5.2. Human vs. model performance

Aside from using humans to judge the quality of model predictions, we can also directly compare the model prediction for an image to that made by humans. For instance, we can compare the confusion matrices for models and humans, treating the annotator specified main label as a human prediction. (Note that, this prediction only matches the ImageNet label on about 80% of the images, despite the fact that annotators are presented with only a few relevant labels, which include the ImageNet label, to choose from—see Figure 8.) Instead of visualizing the full 1000-by-1000 matrices (cf. Appendix C.4), we partition ImageNet classes into 11 superclasses (cf. Appendix A.1.1), allowing us to study confusions *between* and *within* superclasses separately.

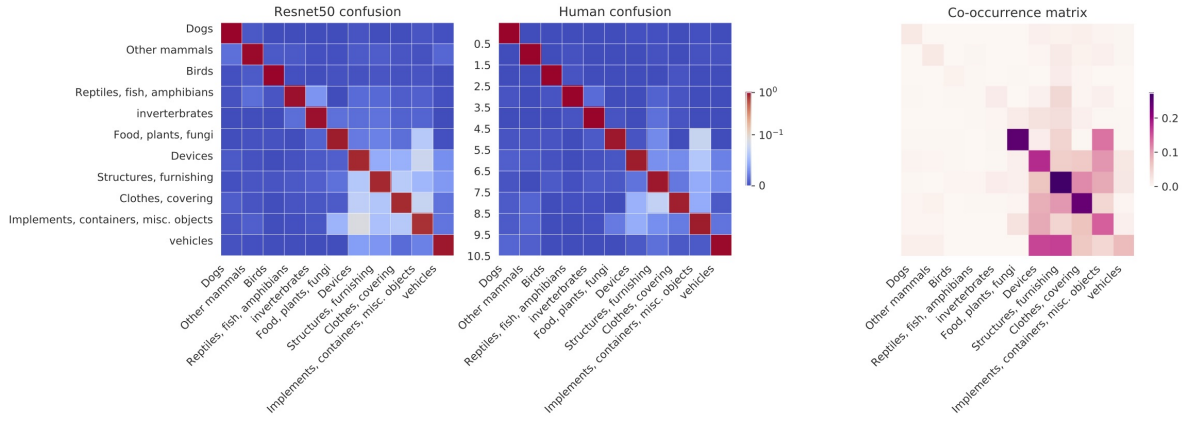
In the cross-superclass confusion matrix (cf. Figure 10a), we observe a block where both human and model confusion is high. This is particularly striking given that these superclasses are semantically quite different. To understand these confusions better, we compare the superclass confusion matrix with the superclass *co-occurrence* matrix, i.e., how often an image of superclass i (w.r.t. the ImageNet label) also contains an object of superclass j according to annotators (cf. Figure 10b). Indeed, we find that the two matrices are quite aligned—indicating that model and human confusion between superclasses might be driven largely by the existence of multi-object images. We also observe that the intra-superclass confusions are more significant for humans compared to models (cf. Appendix Figure 30), particularly on fine-grained classes (e.g., dog breeds), which could be a consequence of the issues identified in Section 4.2.

6. Related Work

Identifying ImageNet issues. Some of the ImageNet label issues we study have already been identified in prior work. Specifically, Recht et al. (2019); Northcutt et al. (2019); Hooker et al. (2019) identify class pairs that might be inherently ambiguous (similar to our findings in Section 4.2). Moreover, the existence of cluttered images was discussed by Russakovsky et al. (2015) as an indication that the dataset mirrors real-world conditions—and hence deemed desirable—and by Stock & Cisse (2018); Northcutt et al. (2019); Hooker et al. (2019) as a source of label ambiguity (similar to our findings in Section 4.1). Additionally, Stock & Cisse (2018) perform human-based model evaluation, with similar conclusions to our experiments in Section 5.1. Finally, Stock & Cisse (2018) use manual data collection and saliency maps to identify racial biases that models rely on to make their predictions. However, the focus of all these studies is not on characterizing these labeling issues in ImageNet—as such they only provide coarse estimates for their pervasiveness. In particular, none of these studies evaluate how these issues affect model performance, nor do they obtain annotations that are sufficient (in granularity and scale) to draw per-class conclusions.

Human performance on ImageNet. Studying human accuracy on the ImageNet classification task is challenging, mainly due to the fact that annotators need to be mindful of all the 1000 ImageNet classes that are potentially present. The original ImageNet challenge paper contained results for two trained annotators (Russakovsky et al., 2015), while Karpathy (2014) reports result based on evaluating themselves. An MTurk study using a subset of the ImageNet classes is presented in Dodge & Karam (2017). In contrast to these studies, we are not interested in estimating human accuracy on ImageNet but rather obtain fine-grained image annotations. Hence, we only ask annotators to select between a few labels per image.

Generalization beyond the test set. The design of large-scale vision datasets that allow generalization beyond the narrow benchmark task has long been a topic of discussion (Ponce et al., 2006; Everingham et al., 2010; Torralba & Efros, 2011; Russakovsky et al., 2015). Torralba & Efros (2011) proposed evaluating cross-dataset generalization—testing the performance of a model on a different dataset



(a) Confusion matrices

(b) Co-occurrence matrix

Figure 10: Similarity between model and human predictions (indicated by main object selection in the CONTAINS task): (a) Model (ResNet-50) and human confusion matrices on 11 ImageNet superclasses (cf. Section A.1.1). (b) Superclass co-occurrence matrix: how likely specific pairs of superclasses are to occur together (using annotations from Section 3.2).

with a similar class structure. Recht et al. (2019) focused on reproducing the ImageNet validation set process to measure potential adaptive overfitting or over-reliance on the exact dataset creation conditions. Kornblith et al. (2019) investigated the extent to which better ImageNet performance implies better feature extraction as measured by the suitability of internal model representations for transfer learning (Donahue et al., 2014; Sharif Razavian et al., 2014).

Adversarial testing. Beyond the aforementioned benign shifts in test distribution, there has been significant work on measuring model performance from a worst-case perspective. Biggio et al. (2013) and Szegedy et al. (2014) demonstrate that models are extremely brittle to imperceptible pixel-wise perturbations. Similar, yet less severe, brittleness can be observed for a range of natural perturbations, such as: worst-case spatial transformation (e.g, rotations) (Fawzi & Frossard, 2015; Engstrom et al., 2019), common image corruptions (Hendrycks & Dietterich, 2019), adversarial texture manipulation (Geirhos et al., 2019), adversarial data collection (Barbu et al., 2019; Hendrycks et al., 2019).

7. Conclusion

In this paper, we take a step towards understanding how closely widely-used vision benchmarks align with the real-world tasks they are meant to approximate, focusing on the ImageNet dataset. Our analysis uncovers systematic—and fairly pervasive—ways in which ImageNet annotations deviate from the ground truth, such as the prevalence of images with multiple valid labels, and ambiguous classes.

Crucially, we find that these deviations significantly impact what ImageNet-trained models learn, and how we perceive model progress. For instance, top-1 accuracy often underestimates the performance of models by unduly penalizing them for predicting a different, but also valid, image label. Further, current models seem to derive part of their accuracy from exploiting ImageNet-specific features that humans are oblivious to, and hence may not generalize well to the real world. Such issues make it clear that measuring accuracy alone may give us only an imperfect view of model performance on the motivating object recognition task.

Taking a step towards evaluation metrics that circumvent these issues, we design a framework that enables us to utilize crowdsourced annotation to directly judge the correctness of model predictions. On the positive side, we find that models that are more accurate on ImageNet also tend to be more human-aligned in their errors. In fact, on average, annotators turn out to be unable to distinguish the (potentially incorrect) predictions of state-of-the-art models from the ImageNet labels. While this might be reassuring, it also indicates that we are at a point where we cannot easily gauge (e.g., via simple crowd-sourcing) whether further progress on the ImageNet benchmark is meaningful, or is simply a result of overfitting to this benchmarks' idiosyncrasies.

More broadly, our findings highlight an inherent conflict between the goal of building large and diverse datasets that capture complexities of the real world and the need for their annotation process to be scalable. Indeed, in the context of ImageNet, we found that some of the very reasons that make the collection pipeline scalable (e.g., the CONTAINS task, crowdsourced annotation) were at the core of systematic annotation issues. We believe that developing annotation pipelines that better capture the ground truth while remaining scalable is an important avenue for future research.

Acknowledgements

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

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases (ECML-KDD)*, 2013.

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.

Dodge, S. and Karam, L. A study and comparison of human and deep learning recognition performance under visual distortions. In *International conference on computer communication and networks (ICCCN)*, 2017.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning (ICML)*, 2014.

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.

Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Steinhardt, J., and Madry, A. Identifying statistical bias in dataset replication. In *ArXiv preprint arXiv:2005.09619*, 2020.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. The pascal visual object classes (voc) challenge. In *International Journal of Computer Vision*, 2010.

Fawzi, A. and Frossard, P. Manitest: Are classifiers really invariant? In *British Machine Vision Conference (BMVC)*, 2015.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., and Li, M. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 558–567, 2019.

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

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019.

Hooker, S., Courville, A., Dauphin, Y., and Frome, A. Selective brain damage: Measuring the disparate impact of model pruning. *arXiv preprint arXiv:1911.05248*, 2019.

Hu, J., Shen, L., and Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

Karpathy, A. What I learned from competing against a ConvNet on ImageNet, 2014. Accessed: 2018-09-23.

Kornblith, S., Shlens, J., and Le, Q. V. Do better imagenet models transfer better? In *computer vision and pattern recognition (CVPR)*, 2019.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision (ECCV)*, 2014.

Liu, C., Zoph, B., Neumann, M., Shlens, J., Hua, W., Li, L.-J., Fei-Fei, L., Yuille, A., Huang, J., and Murphy, K. Progressive neural architecture search. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 19–34, 2018.

Miller, G. A. Wordnet: a lexical database for english. *Communications of the ACM*, 1995.

Northcutt, C. G., Jiang, L., and Chuang, I. L. Confident learning: Estimating uncertainty in dataset labels. *arXiv preprint arXiv:1911.00068*, 2019.

Ponce, J., Berg, T. L., Everingham, M., Forsyth, D. A., Hebert, M., Lazebnik, S., Marszalek, M., Schmid, C., Russell, B. C., Torralba, A., et al. Dataset issues in object recognition. In *Toward category-level object recognition*, 2006.

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

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.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Sharif Razavian, A., Azizpour, H., Sullivan, J., and Carlsson, S. Cnn features off-the-shelf: an astounding baseline for recognition. In *conference on computer vision and pattern recognition (CVPR) workshops*, 2014.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

Stock, P. and Cisse, M. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *European Conference on Computer Vision (ECCV)*, 2018.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

Tan, M. and Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

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

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

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.

Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8697–8710, 2018.