When does dough become a bagel? Analyzing the remaining mistakes on ImageNet

Anonymous Author(s)
Affiliation
Address
email

Abstract

Image classification accuracy on the ImageNet dataset has been a barometer for progress in computer vision over the last decade. Several recent papers have questioned the degree to which the benchmark remains useful to the community [33, 4, 31, 43, 37], yet innovations continue to contribute gains to performance, with today's largest models achieving 90%+ top-1 accuracy. To help contextualize progress on ImageNet and provide a more meaningful evaluation for today's state-of-the-art models, we manually review and categorize every remaining mistake that a few top models make and provide insights into the long-tail of errors on one of the most benchmarked datasets in computer vision. We focus on the multi-label subset evaluation of ImageNet, where today's best models achieve upwards of 97% top-1 accuracy. Our analysis reveals that nearly half of the supposed mistakes are not mistakes at all, and we uncover new valid multi-labels, demonstrating that, without careful review, we are significantly underestimating the performance of these models. On the other hand, we also find that today's best models still make a significant number of mistakes (40%) that are obviously wrong to human reviewers. To calibrate future progress on ImageNet, we provide an updated multi-label evaluation set, and we curate **ImageNet-M**ajor: a 68-example "major error" slice of the obvious mistakes made by today's top models—a slice where models should achieve near perfection, but today are far from doing so.

1 Introduction

Computer vision models often evaluate their performance on the ImageNet classification dataset [6, 30] and many variants [29, 12, 13, 11, 39], as a signal of capability for visual understanding. As performance on the standard sets have reached diminishing returns to top-1 and top-5 accuracy, much recent work [33, 4, 29, 31, 37, 17] has focused on understanding what is left for the computer vision community to solve, and where the community should be driving toward. Prior studies of ImageNet errors have identified issues stemming from lack of multi-labels, label noise, under-specified classes, and more [33, 31, 4, 37, 18].

Label: dough; Model: bagel. When does dough become a bagel?

Label errors and label noise affect the evaluation of any model [25], and ImageNet is no exception. Many studies above have spent effort to correct and improve these labels, showing that while ImageNet performance improvements are approaching diminishing returns, the dataset can remain useful to the community, but only if we collectively continue to shepherd it. As the best models improve, however, it is becoming increasingly challenging to assess the often novel predictions these models make. For example, should we penalize models for being the first to predict that a pre-baked bagel may be a bagel, as one of the models we review in this work does?

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Machine learning models tend to make mistakes with varying severity and importance (a function of both the prediction as well as the label definitions), and previous studies [33, 4] on ImageNet have shown that non-experts find it challenging to determine the correctness of a model's prediction. Our own experience, highlighted by the doughy-bagel, is that many of the remaining mistakes these top models make are quite reasonable and probably should not be considered mistakes—understanding the severity and type of these remaining mistakes can help calibrate our barometer of progress.

Indeed, to our knowledge, there has not been an expert-review, categorization, and severity assessment of the remaining long-tailed mistakes, which becomes particularly important at these margins. Our experience working with production teams on deployed applications has suggested that manual triage and assessing individual failures provides a useful indicator of model performance that aggregate measures can fail to capture. Thus, in this work we attempt to analyze (as expert reviewers) *every remaining mistake* that a few state-of-the-art models make to better understand (a) which of the remaining mistakes remain egregious errors, (b) what error category they might fall in, and (c) what evaluations might capture the most important long-tail failures that remain.

In this paper we analyze the ImageNet multi-label validation subsets [31], in which expert labelers were used to assess the correctness of model predictions through the year 2020, and on which a 1000-image human-evaluated subset provides a direct comparison to expert human performance. By analyzing the mistakes of two large 2022-era ImageNet models, we found that:

- **Nearly half of each model's mistakes were deemed correct** under a careful, expert multi-label re-evaluation, halving the error rate. Had we not analyzed the models' mistakes, we would be severely underestimating the models' actual performance.
- **Approximately 40% of the remaining mistakes can be classified as 'major' errors**: errors that most humans would likely not make, suggesting that many of the long-tailed mistakes aren't simply label noise, but legitimate mistakes that leave room for improvement.

What do these lessons portend for the future of ImageNet evaluation? Top-1 will become increasingly noisy as our best models get better (though we have not yet completely saturated top-1). Our work shows that multi-label accuracy, while better at capturing "true" errors compared to top-1, suffers from a lack of a comprehensive, accurately labeled large evaluation set, which is expensive to procure and challenging to maintain.

We therefore propose ImageNet-M, a 68-example evaluation split of the multi-label evaluation set composed of "major" mistakes that several top-performing models make; we believe this subset is one that future image classification models should achieve near perfect accuracy on, and provides three clear benefits: (1) we attempt to comprehensively-label all examples for multi-label annotations to prevent the need to review novel correct predictions, (2) we endeavor to maintain and provide a way for the public to add new correct predictions; (3) the evaluation set is small enough to encourage completeness and allow the community to inspect their own errors.

2 Related Work

Multi-label annotations on ImageNet. As models continue to improve ImageNet top-1 accuracy, there has been an increased interest in evaluating ImageNet multi-label accuracy [33, 4, 31, 43, 37]. Stock et al. [33] use non-expert human studies and explanations (e.g., model criticisms) on predictions of a then-SOTA model on ImageNet, finding that machine accuracy is underestimated and advocating for multi-label evaluation. Beyer et. al. [4] introduce a set of Reassessed Labels (ReaL) for the ImageNet validation set containing multi-label annotations. The researchers first collected proposal labels from model predictions using a testbed of 19 models, and then, in order to reduce the number of predictions to review, they narrowed the set of models down to the 6 models that had the highest precision and recall above 97% on a set of 256 images reviewed by 5 experts. The top predictions from these 6 models were then reviewed by human annotators from a crowdsourcing platform. In a similar vein, Shankar et. al. [31] provide a multi-label annotation dataset for 20,000 of the 50,000 ImageNet validation images and find that roughly 20% of images have more than one valid label. To generate the annotations, the researchers first collected predictions from a testbed of 70 models for each image and then reviewed the unique model predictions using a panel of three human experts. Additionally, human accuracy was evaluated on a subset of 1,000 images and a panel of 5 human experts reviewed human and model predictions on this subset. Tsipras et al. [37] also find that

20% of images in the validation set contain objects from multiple classes, and identify sources of ambiguous label classes. Hooker et. al [14] use human evaluations to label a subset of examples from ImageNet, finding that they contained multiple labels 40–60% of the time. More recently, Yun et al. [43] obtained pixel-level multi-label ground truths for the ImageNet *training set* using a machine annotator, and found that training a model with the dense annotations leads to small improvements in both top-1 and multi-label ImageNet accuracy.

Our paper focuses on (and adds to) the multi-label evaluation of ImageNet using expert labelers, updating the dataset to collapse classes that are overlapping or subset relationships to better capture the remaining real errors. We adopt the labeling methodology of Shankar et al. [31] and use a panel of human experts to determine the validity of novel predictions. In contrast to prior work, we re-visit all remaining model mistakes using human expert review, and categorize them by type and severity.

ImageNet Label Error. Multi-label annotation datasets for ImageNet (including our updated annotations) identify a set of images that have no correct ground truth label (i.e. label errors). Van Horn et. al [38] used experts from the Cornell Lab of Ornithology to estimate that at least 4% of the bird images are misclassified, and more recently Northcutt et. al [25] used MTURK workers to review algorithmically identified potential errors and found a label error rate of 5.83% in ImageNet. Lee et. al [18] sample 400 ImageNet mistakes and perform a categorization, similarly finding significant label error and label ambiguity.

Mistake analysis. Recent work has sought to understand or analyze model mistakes on ImageNet, mainly focusing on similarities or differences between mistake sets of independently trained classifiers. For example, Mania et al. [20] found that predictions of different models on ImageNet are more similar than one would expect if the models were making mistakes independently. Geirhos et al. [9] studied a 16-class ImageNet classification task and similarly found that CNNs make remarkably consistent mistakes with one another but CNNs and humans have an error consistency that is only slightly above what can be expected from chance. In follow up work, Mania et al. [21] studied the *dominance probabilities* for pairs of ImageNet models, which capture the probability that a higher accuracy model will make a mistake on a particular image that a lower accuracy model correctly classifies. Their empirical analysis of dominance probabilities on ImageNet implies that the mistakes of higher accuracy models are typically subsets of the mistakes of lower accuracy models. However, Gontijo-Lopes et al. [10] and Andreassen et al. [1] found that training on larger and more diverse datasets as well as zero-shot evaluation can lead to models with more mistake diversity, which in turn can be used to build more accurate ensembles. Similarly, Nguyen et al. [23] found systematic differences in the errors between wide and deep ResNets on ImageNet. Ciregan et al. [5] demonstrated that powerful models on MNIST allowed for remaining error analysis in the single-label context, highlighting the ambiguity in many of those remaining errors.

In contrast to this prior work, we take a step towards fully calibrating SOTA model evaluations by exhaustively and visually reviewing every remaining model mistake using manual expert review in the *multi-label context*; we believe the remaining errors identified to be legitimately incorrect and tempered by severity and category ratings that we hope prove useful to future evaluations.

3 Mistake analysis method and taxonomy

To obtain an initial set of mistakes remaining on ImageNet, we used a standard ViT [7] model scaled to 3B parameters (ViT-3B) that was pre-trained on JFT-3B [34] and fine-tuned on ImageNet-1K [6], achieving a top-1 accuracy of 89.5% (details in Appendix A). We also later review mistakes made by the Greedy Soups model [40]. Using the `imagenet2012_multilabel` dataset [31], we measured the initial *multi-label accuracy* (MLA) of the ViT-3B model to be 96.3%. The model made a total of 676 apparent mistakes, which we sought to manually review in detail. Exhaustively examining every mistake has been made more convenient and practical due to the quality of today's top models as well as the smaller subset of 20k validation images present in the multi-label set.

3.1 Panel Review

To exhaustively and accurately assess every remaining mistake, we formed a panel of five reviewers and followed a process similar to [31] to evaluate the predictions made by this model on the 676 mistakes — we avoided using non-expert crowd-source platforms specifically because the remaining

mistakes are often difficult to assess by non-expert annotators [37, 4]. For every mistake, the panel determined: (1) Did the model make a mistake? (2) Was the original ground truth annotation correct? (3) If the model made a mistake, what is the category, type, and severity of the mistake? The `imagenet2012_multilabel` dataset contains a field for every image indicating which classes a large previous suite of models predicted that were determined to be incorrect (wrong_multi_labels). Of these 676 initial mistakes, 221 were novel: they were not reviewed in the original multi-label annotation process since none of the models evaluated made the same prediction. Each member of the panel reviewed all 221 novel mistakes.

Similar to [31], we built a review tool that allowed each panelist to see a) the predicted class, b) the predicted top softmax score, c) the set of ground-truth labels, d) the set of previously incorrect labels, and e) the image. We also employed the labeling guide produced by the authors of [31] when investigating the definition of a class, and a tool to iterate through the images of every ImageNet validation example for that class, using the validation images to help define the class boundaries (rather than the gloss or lay definition of the class). See Appendix B for screenshots of the review tools. In addition, we collapsed a small number of classes for which previous work has identified exhibited extreme overlap [24], such as 'missile' and 'projectile missile'. In Appendix C we provide these new collapsed class mappings.

We also used Google Image Search to help provide context to some assessments; in one interesting but not isolated case, a prediction of a taxi cab (with no obvious taxi cab indicators beyond yellow color) was present in the image; we determined the prediction to be correctly a taxi cab and not just a standard vehicle by identifying a landmark bridge in the background in order to localize the city, and a subsequent image search for taxis in that city yielded the images of the same taxi model and license plate design, validating the model's actually correct prediction.

Each panelist rated whether these novel mistakes had a mislabeled ground truth, or whether the prediction should be added to the set of correct, unclear, or wrong multi-labels. As a group, we reviewed any image where there was no unanimous agreement, allowing those in the minority to make their case and change minds, or highlight potential oversights. The use of a panel and discussion was important: in a non-trivial number of cases, a single panelist found unique evidence for or against a prediction that no other panelist saw that led to a different outcome. After locking in final votes, we took the majority assessment, or used 'unclear' for a tie. After this re-assessment, 140 of the novel predictions were deemed correct (or the ground truth deemed incorrect), leaving 536 remaining mistakes to assess.

3.2 Mistake severity and category

The remaining mistakes then comprised images that had either previously been deemed wrong by the panel in Shankar et al. [31], or deemed wrong by the current panel. We then began a review of the mistakes for the category and severity of mistakes. During this second phase review, we re-reviewed all images in detail, potentially overturning decisions made by the previous panel, where they missed the presence of an object that was in an example or the decision was inconsistent with the labeling guide or validation examples we relied on.[1]

Severity: We assessed each mistake's severity with the assumption that not all mistakes are equal: some mistakes are extremely borderline, particularly because the ImageNet class definitions are imprecise or because the image itself provides ambiguous or incomplete information. For any remaining mistake, we broke down the severity levels into **major** and **minor** mistakes. **A major mistake** is a model prediction that a human who understands the class definitions would find obviously incorrect. For example Figure 1(a) shows an example image where the prediction is jigsaw puzzle but the label is dough; although the pieces are somewhat jigsaw-puzzle-shaped, untrained humans are more likely to classify this as dough than jigsaw puzzle. **A minor mistake** is a model prediction that a human who understands the class definitions would probably find to be incorrect, but in a more subtle way than a major mistake. Some minor mistakes are so subtle that even expert-trained humans might debate their correctness.

[1]The labeling guide in Shankar et al. [31] was constructed after the initial panel review but prior to the human accuracy assessment, resulting in some validation labels that would be inconsistent with the labeling guide we used.

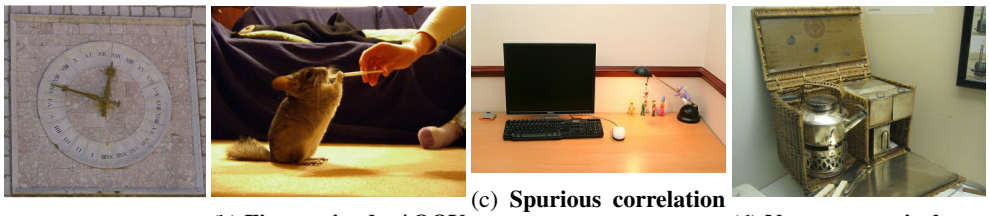
(a) **Major mistake**
Label: dough
Model: jigsaw puzzle

(b) **Minor mistake**
Label: kuvasz
Model: Great Pyrenees

(c) **New multilabel**
Label: tape player
Model: cassette

(d) **Problematic**
Label: bee
Model: fly

Figure 1: **Mistake Severity.** Examples of the two mistake severities (a-b), a correct model prediction where the model identifies a previously missing multi-label (c); and a problematic example (d) where the label (bee) is incorrect (object is a bee-fly, which is a type of fly).



(a) **Fine-grained**
Label: wall clock
Model: sundial

(b) **Fine-grained w/ OOV**
Label: syringe
Model: hamster

(c) **Spurious correlation**
Label: mouse, desk, monitor, screen
Model: desktop computer

(d) **Non-prototypical**
Label: stove
Model: hamper

Figure 2: **Mistake Category.** Examples of the four mistake categories. In the fine-grained with OOV example, the animal is a chinchilla, which is not an ImageNet class but is visually similar to a hamster, which is an ImageNet class. In the spurious correlation example, the scene contains relevant context for desktop computer, but there is no such object in the image.

We recognize that these severities are subject to the influences of the worldview of the panelists (and the web) and should be judged accordingly (see Section 5.2). For transparency, we provide the panel assessments of the severity for all the mistakes, and in Figure 1 we provide some examples of the severities of various mistakes made by this model for the reader's calibration. In Appendix D we provide many additional examples of every severity and category.

Category: After reviewing many mistakes, we formulated four mistake types (Figure 2).

(1) Fine-grained errors are where the model makes a mistake between two similar types of organisms or objects, one of which is a groundtruth label. These mistakes often occur when the two confused classes are already similar (e.g. two dog breeds), or when either of them is very broad or ill-scoped (e.g. the "bake shop" class includes any baked good or bakery).

(2) Fine-grained with out-of-vocabulary: there is an object in the image that is not in the ImageNet class hierarchy but is similar to a predicted class that is in ImageNet. We separate out this category because it highlights the possible benefits of training models to expect new classes to appear at test time, and the importance of model uncertainty and calibration in the face of this 'open world'.

(3) Spurious correlations: Either (a) the predicted object is not plausibly in the image but surrounding cues may have been used, or (b) the predicted object does not match the groundtruth. In extreme cases, there is no clear indication of the predicted object; in more subtle cases, it is more clear the model is trying to predict the class of an object in the image but the predicted object would not be considered either semantically or visually similar to the groundtruth class.

(4) Non-prototypical labels: The predicted label is not present but the groundtruth object is a non-prototypical example of the class that bears resemblance to the predicted label. Non-prototypical mistakes are relatively rare, and capture the 'long tail' of examples for each class. These mistakes highlight the internal diversity of each class, and the difficulty of modeling the long within-class tail.

4 Analyzing the remaining mistakes

After review of all original 676 mistakes (comprising both novel predictions and previously reviewed mistakes), we found that a total of 298 were either correct or unclear, or determined the original groundtruth incorrect or problematic. Our evaluation of the ViT-3B model on this re-labeled dataset is shown in Table 1, with the model making a total of 378 mistakes on the dataset. In other words, approximately 44% of the initial mistakes made by this model were determined to be correct!

	All		Organisms		Objects	
	MLA	MLA Re-labeled	MLA	MLA Re-labeled	MLA	MLA Re-labeled
ViT-3B	96.3%	97.9%	96.3%	97.8%	96.4%	98.0%

Table 1: Multi-label accuracy (MLA) of ViT-3B model before and after our re-labeling.

4.1 Mistake category and severity

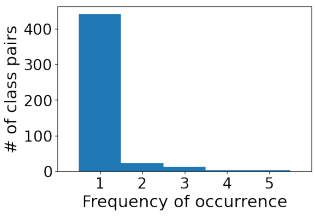
Each of the 378 remaining mistakes was assigned both a mistake category and severity (Table 2) by the expert panel.

Model	Dataset	Categories				Severities	
		Fine-grained (FG)	FG w/ OOV	Spurious	Non-prototypical	Major	Minor
ViT-3B	ImageNet	64.0%	14.3%	13.8%	7.9%	40.7%	59.3%
ViT-3B	ImageNetV2	66.0%	9.4%	15.1%	9.4%	41.5%	58.5%
Greedy Soups	ImageNet	69.1%	10.4%	12.9%	7.6%	35.7%	64.3%

Table 2: **Mistake Category and Severity:** We classified the majority of the ViT-3B mistakes as fine-grained or a variant of fine-grained; many of the mistakes were considered "major" mistakes; these distributions held on ImageNetV2 as well as for the Greedy Soups model.

Category: 78.3% of errors were assessed to be fine-grained in nature (either in-vocabulary or OOV), and 13.8% categorized as a spurious correlation. To measure whether these categories are meaningful, we measured the "hierarchy distance" between the groundtruth label and the model's prediction using the WordNet hierarchy [22]. For example, a hierarchy distance of 1 means that the groundtruth and model prediction share the same parent; a distance of 2 means they share the same grandparent. We found that 80.9% of errors with a hierarchy distance of 1 were assessed as fine-grained. In contrast, 54.3% of errors with a hierarchy distance of 3 were fine-grained, matching intuition that predictions close in the hierarchy are very likely to be fine-grained errors, and predictions far in the hierarchy are more likely to be spurious correlations, fine-grained with out-of-vocabulary, or non-prototypical examples. We note that WordNet does not provide a perfect (automated) category function: "goblet" and "vase" are a hierarchy distance of 4 apart, and we encountered one model mistake for this pair that we nonetheless assessed as fine-grained.

Analyzing class confusions: Given that most failures were fine-grained, we tried to identify any patterns present in the class confusions made by the model, but found no consistent pattern. The inline figure shows the frequency of occurrence for confused class pairs. The distribution is long-tailed in nature—the majority of class pairs occur exactly once or twice and only a handful of class pairs occur three or more times. The long-tailed nature of the class confusions suggests that we will not be able to resolve a large fraction of model mistakes by focusing on cleaning up or adding additional data to only a small number of classes.



Severity: We determined that around 40% of errors were assessed to be "major" errors, indicating that this model still appears to make mistakes that a human familiar with the class definitions would not make, despite the fact that the model on average performs better than an expert human. We return to 'major' errors later in Section 5.1, as we believe that a subset of these errors can be a useful evaluation slice for future ImageNet benchmarking.

4.2 Out-of-distribution generalization

We evaluated the ViT-3B model on the ImageNetV2 multi-label subset which produced over 900 unreviewed errors. To assess what aspects of our analysis generalize to other datasets, we sampled

100 of these errors using the same panel review system employed for ImageNetV1. We discovered that 47 of the 100 ImageNetV2 predictions were either correct or had problematic labels, leaving 53 mistakes that we reviewed for category and severity. For ImageNetV1, we had previously found that 44% (296/676) of mistakes were either problematic or correct and we found no statistically significant difference between the two datasets in this regard ($\chi^2(1, N = 776) = 0.36$, $p = .55$). These proportions suggest that large models are frequently uncovering new correct multi-labels, suggesting that mistake analysis and label correction needs to be part of the lifecycle and maintenance of benchmark development of long-tailed errors to properly assess performance as a benchmark saturates. Table 2 compares the category and severity breakdowns between the two datasets—overall the model is making similar types and severities of mistakes on both datasets. A chi-square test of independence shows that there is no significant difference between either the mistake categorization $\chi^2(3, N = 434) = 0.97$, $p = .80$. or mistake severity $\chi^2(1,\ N = 434) = 0.01$, $p = .92$.

4.3 Generalization to new models

As models produce higher top-1 accuracy, how do the types of mistakes they make and improve upon change? We use the Greedy Soups model that obtains 90.9% top-1 accuracy on ImageNet validation [40], measuring its MLA after our initial re-labeling at 98.1%, and yielding 341 total remaining (and partially unreviewed) errors.

The Soups model corrected 209 mistakes that the ViT-3B model made, while the model made 170 mistakes where the ViT-3B model was correct, yielding an overall accuracy improvement; 28 mistake examples were common with ViT-3B but with a different prediction. In total there were 198 novel predictions made by this model that needed to be reviewed; upon review using the same panel method, we found 46.5% (92/198) were problematic (10) or actually correct (82), showing with a second model that model predictions on mistakes need to be reviewed, and that the single label expected by top-1 is often insufficient. The Soups model in the end made only 249 errors, for an MLA of 98.6%, a 0.5% absolute increase compared to unreviewed mistakes. The categorization and severity of these errors is shown as the last line in Table 2. Again, a chi-square test of independence shows that there is no significant difference between either the mistake categorization ($\chi^2(3, N = 629) = 2.41$, $p = .49$) or mistake severity ($\chi^2(1,\ N = 629) = 1.61$, $p = .20$).

4.4 Comparison to humans

We compare the performance of the ViT-3B and Greedy Soups models to the best human labeler from Shankar et al. [31][2] by evaluating on the subset of 1,000 ImageNet val images used to compute human accuracy in the prior work. To fairly compare to the models, we re-compute the human accuracy scores using the original human predictions and our updated label set. Overall, the re-labeling did not significantly change human accuracy; the best human labeler achieved 97.3% MLA on the original multi-labels and 97.4% MLA on the updated multi-labels.

Table 3 compares the re-labeled MLA of the ViT-3B and Greedy Soup model to the best human for all ImageNet classes as well as the subset of ImageNet classes corresponding to objects and organisms. Similar to Shankar et al. [31], we also find that the performance of the models is more uniform across the object and organism classes, but humans do substantially better on the object classes than the organism classes. However, unlike prior work, current models outperform the best human when evaluated on all ImageNet classes (though humans still achieve slightly better performance on the object classes).

	All Classes	Organisms	Objects
ViT-3B	97.8%	97.4%	98.1%
Greedy Soup	98.6%	98.4%	98.8%
Best Human [31]	97.4%	95.4%	99.0%

Table 3: **Multi-label accuracy compared to humans**. Both the ViT-3B and Greedy Soup model achieve better MLA on all ImageNet classes than the best human labeler from Shankar et al. [31]. However, on the object classes, where the class boundaries are substantially less ambiguous for humans, the best human labeler still outperforms the models.

[2] Best human is chosen based on MLA on all classes (as opposed to object or organism subset) and corresponds to Human E in the original work.

4.5 Analyzing the Training Data

Finally, we investigate how much we can understand model validation set errors through the lens of the training data. To do so, we inspect the K = 10 nearest neighbor training images using JFT-pretrained (before ImageNet fine tuning) embeddings for the ViT-3B model. Doing this, we rediscover (originally documented in Sun et al. [34] and Kolesnikov et al. [16]) that 797 (1.59%) ImageNet validation images exist in the training set as **exact duplicates**. Interestingly, we are the first to notice that **every single leaked image has a different label in the training set than the validation set**. In Appendix G.1 we show removing these images has relatively little impact on the model's performance, and detail a more pernicious leakage pattern of "near duplicates" in Appendix G.2 that is hard to fully quantify. Finally, we show how the training data sometimes explain spurious correlations in Appendix G.3.

5 Recommendations and Discussion

In this section we provide recommendations for future evaluation, starting with ImageNet-M: our curated multi-label evaluation set of "major mistakes" that we suggest should be reported on in addition to metrics such as top-1 and multi-label accuracy.

5.1 ImageNet-M: A "major mistakes" evaluation split

Studies over the last several years seeking to understand whether ImageNet remains an informative benchmark have typically concluded that aspects of ImageNet remain useful but top-1 accuracy less so. These works encourage alternative related metrics for ImageNet such as multi-label accuracy [4, 31] or object vs. organism breakdowns [31]. In many of these cases, stronger, more generalizable models often continue to incrementally but inevitably improve on these metrics.

With an emphasis on understanding which long-tail errors are unambiguously errors, we suggest that a benchmark focusing on the most egregious long-tail errors could provide a useful additional signal about whether the improvements are meaningful. In particular, we desire a long-tailed benchmark where we believe that 100% accuracy is achievable. Because the minor mistakes are more subject to interpretation and discussion than the major mistakes, we believe a benchmark focused on the latter will help the community judge what is meaningful improvement on ImageNet.

To that end, we leverage our expert-reviewed analysis to produce a small slice of the ImageNet multi-label set where (1) today's *best* top-1 models are still more wrong than right, and (2) the mistakes are largely unambiguous to a human given a reasonable understanding of the ImageNet label set. We call this evaluation slice **ImageNet-M**ajor.

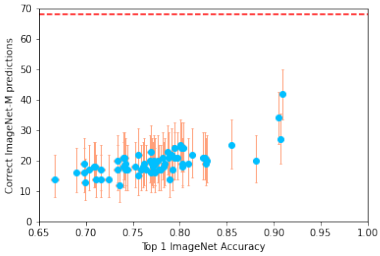
ImageNet-M example selection method. The ViT-3B model made 155 "major" mistakes, for which we analyzed whether each example was labeled correctly for three additional models: (1) the Greedy Soups model, (2) a model pre-trained on Instagram data but fine-tuned on ImageNet that achieves 85.4% top-1 [19], and (3) A zero-shot evaluation [28, 15, 27] using a CoCa [42] model pretrained on JFT and noisy image-text data. In order to maximize prediction diversity, we purposefully selected models with varying pre-training data and training methodologies, including a zero-shot model that does not see ImageNet image-label associations directly [10].

From this suite of four models, we assembled a subset where three or more of the models make a major mistake, yielding 68 such major mistakes. This process is similar in spirit to ImageNet-A [13], except we use four high-performing but diverse models, and we restrict the set to ImageNet images and the corresponding model prediction that were rated as "major" errors. We analyzed the predictions of all models on these examples (including any novel predictions made by these additional models) and verified that none of them were correct new multi-labels, and that any model's mistakes were major mistakes. In addition, we attempted to comprehensively manually label additional labels that no model has yet predicted but would be correct, in an attempt to reduce the likelihood that future models are penalized for making novel but unreviewed correct predictions.

We design the ImageNet-M 68-example subset as an additional split of the validation set that we believe has the following properties: (a) many top performing models trained in different ways make mistakes on this set; (b) the mistakes are all major mistakes as determined by expert-reviewers; (c) the example set is small enough to permit manual inspection by model evaluators; (d) strives to

be comprehensively labeled with respect to the ImageNet label set; (e) in theory provides a subset that *future models could achieve perfect accuracy on* without worrying about underspecified class definitions. We anticipate stronger models may yet make correct unreviewed predictions on this slice, so we will endeavor to update the set of multi-labels as they are reported to us.

Evaluation. By construction, our ViT-3B model achieves 0% accuracy on ImageNet-M; the Instagram-pretrained model gets 9 of the 68 correct, while the Greedy Soups model gets 19 correct. The zero-shot model gets the best performance with 24 correct, even though the zero-shot model overall achieves lower multi-label accuracy (94.2%) than any of the other models. Because we use these four models to help choose the mistakes, these specific models comparatively will perform poorly on this benchmark. Like with ImageNet-A [13], a dataset chosen using models may bias selection in such a way that future models may easily get very high accuracy on this subset, though we try to mitigate this effect by using multiple differently-trained models in our selection criteria.



Trend of models on top-1 ImageNet vs. ImageNet-M, using Clopper-Pearson intervals. Red dashed line indicates total number of images in ImageNet-M as an upper-bound.

How do models that were not used to select this dataset perform? We evaluate the suite of 70 models from Shankar et al. [31] on this dataset, in addition to four recent top models not directly used to help filter the ImageNet-M set: a ViT-G/14 model [44] (90.5% top-1), a BASIC model [27] fine-tuned on ImageNet (90.7% top-1), an ALIGN model [15] fine-tuned on ImageNet (88.1% top-1), and a CoCa model [42] fine-tuned on ImageNet (91.0% top-1). The plot shown here shows that most models as far back as AlexNet through ResNets get between 10-25 examples correct, but recent high accuracy models such as ViT-G/14, BASIC-FT, and CoCa-FT are starting to solve more of these 'major' mistakes: CoCa-FT gets 42 of the 68 examples correct. We reviewed the mistakes made by these four models, which yielded a total of 5 novel predictions; 4 of them were verified to be wrong (and major), and 1 additional new valid prediction, for which we updated the label set accordingly. We note that future models may predict classes on these examples that are "minor" mistakes, since the definition of severity is linked to the (prediction, example) pair; should it be useful, the dataset slice can be augmented with a 'minor_wrong_multi_label' attribute to provide more fine-grained signals.

Although we did not find statistical evidence that stronger models solve major errors first, we hope that progress on image classification can be evaluated by whether improvements help reduce egregious mistakes before focusing on nebulous ones. Overall, we encourage reporting on several slices, including ImageNet-M, to give a better sense of the strengths and weaknesses of various models.

5.2 Limitations of Analysis

Much of our analysis on mistake categorization, severity, and data cleaning depends greatly on qualitative factors determined by the authors and experiment design, which we briefly discuss here.

Limitations due to mistake subset. We only reviewed multi-label annotation examples where the ViT-3B model and Greedy Soups model we chose were incorrect; while the multi-label dataset itself has undergone review of mistakes from a suite of many models, we never review any validation image whose groundtruth might be wrong, if all models evaluated also make the incorrect groundtruth prediction. We do partially review some of the mistakes made by a few other models, and build upon a dataset where mistakes made by many other models have already been reviewed, but most of the work here assesses these two specific model's predictions.

Definition of a mistake. We re-iterate that we used qualitative judgments to decide whether a prediction was a mistake, and if so, its severity and categorization. Our qualitative judgments are therefore based on a biased worldview [36, 8, 2, 26] comprising the five panelists; moreover, we are not world experts on dog or animal species, though we believe our assessments are at least as good or better than the original labeling process used for the validation set, given the research effort we made on each mistake. As a mitigation against imbuing too biased a worldview on class definitions, we relied heavily on the validation data to define the boundaries of the class, even when those examples did not match up with our personal definitions of the class. Moreover, updating (and evaluating) multi-labels potentially allows for different world-views to be expressed.

References

[1] A. Andreassen, Y. Bahri, B. Neyshabur, and R. Roelofs. The evolution of out-of-distribution robustness throughout fine-tuning, 2021. https://arxiv.org/abs/2106.15831.

[2] D. Baker. Datasets have worldviews. https://pair.withgoogle.com/explorables/dataset-worldviews. Accessed: 2022-04-28.

[3] B. Barz and J. Denzler. Do we train on test data? purging cifar of near-duplicates. *Journal of Imaging*, 6(6):41, 2020.

[4] L. Beyer, O. J. Hénaff, A. Kolesnikov, X. Zhai, and A. v. d. Oord. Are we done with imagenet? *CVPR*, 2020. https://arxiv.org/abs/2006.07159.

[5] D. Ciregan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3642–3649. IEEE, 2012.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, 2009. https://ieeexplore.ieee.org/document/5206848.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[8] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

[9] R. Geirhos, K. Meding, and F. A. Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *NeurIPS*, 2020. https://arxiv.org/abs/2006.16736.

[10] R. Gontijo-Lopes, Y. Dauphin, and E. D. Cubuk. No one representation to rule them all: Overlapping features of training methods, 2021. https://arxiv.org/abs/2110.12899.

[11] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.

[12] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

[13] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song. Natural adversarial examples. *CoRR*, abs/1907.07174, 2019.

[14] S. Hooker, A. Courville, G. Clark, Y. Dauphin, and A. Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.

[15] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[16] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby. Big transfer (bit): General visual representation learning. In *European conference on computer vision*, pages 491–507. Springer, 2020.

[17] S. Kornblith, J. Shlens, and Q. V. Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[18] H. S. Lee, A. A. Agarwal, and J. Kim. Why do deep neural networks still not recognize these images?: A qualitative analysis on failure cases of imagenet classification. *arXiv preprint arXiv:1709.03439*, 2017.

[19] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.

[20] H. Mania, J. Miller, L. Schmidt, M. Hardt, and B. Recht. Model similarity mitigates test set overuse. *NeurIPS*, 2019. `https://arxiv.org/abs/1905.12580`.

[21] H. Mania and S. Sra. Why do classifier accuracies show linear trends under distribution shift? *arXiv preprint arXiv:2012.15483*, 2020. `https://arxiv.org/abs/2012.15483`.

[22] G. A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41, 1995.

[23] T. Nguyen, M. Raghu, and S. Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. In *International Conference on Learning Representations (ICLR)*, 2021.

[24] C. Northcutt, L. Jiang, and I. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

[25] C. G. Northcutt, A. Athalye, and J. Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021. `https://arxiv.org/abs/2103.14749`.

[26] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021.

[27] H. Pham, Z. Dai, G. Ghiasi, H. Liu, A. W. Yu, M.-T. Luong, M. Tan, and Q. V. Le. Combined scaling for zero-shot transfer learning. *arXiv preprint arXiv:2111.10050*, 2021.

[28] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[29] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400, 2019.

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and F.-F. Li. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015. `https://arxiv.org/abs/1409.0575`.

[31] V. Shankar, R. Roelofs, H. Mania, A. Fang, B. Recht, and L. Schmidt. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning (ICML)*, 2020. `http://proceedings.mlr.press/v119/shankar20c/shankar20c.pdf`.

[32] J. Shen, P. Nguyen, Y. Wu, Z. Chen, M. X. Chen, Y. Jia, A. Kannan, T. Sainath, Y. Cao, C.-C. Chiu, et al. Lingvo: a modular and scalable framework for sequence-to-sequence modeling. *arXiv preprint arXiv:1902.08295*, 2019.

[33] P. Stock and M. Cissé. Convnets and imagenet beyond accuracy: Explanations, bias detection, adversarial examples and model criticism. *CoRR*, abs/1711.11443, 2017.

[34] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *International Conference on Computer Vision (ICCV)*, 2017. `https://arxiv.org/abs/1707.02968`.

[35] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020.

[36] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[37] D. Tsipras, S. Santurkar, L. Engstrom, A. Ilyas, and A. Madry. From imagenet to image classification: Contextualizing progress on benchmarks. *CoRR*, abs/2005.11295, 2020.

[38] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *CVPR*, 2015. `https://vision.cornell.edu/se3/wp-content/uploads/2015/05/Horn_Building_a_Bird_2015_CVPR_paper.pdf`.

[39] H. Wang, S. Ge, Z. Lipton, and E. P. Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019.

[40] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *arXiv preprint arXiv:2203.05482*, 2022.

[41] Y. Xu, H. Lee, D. Chen, B. Hechtman, Y. Huang, R. Joshi, M. Krikun, D. Lepikhin, A. Ly, M. Maggioni, et al. Gspmd: general and scalable parallelization for ml computation graphs. *arXiv preprint arXiv:2105.04663*, 2021.

[42] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[43] S. Yun, S. J. Oh, B. Heo, D. Han, J. Choe, and S. Chun. Re-labeling imagenet: from single to multi-labels, from global to localized labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2340–2350, 2021. `https://arxiv.org/abs/2101.05022`.

[44] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer. Scaling vision transformers. *arXiv preprint arXiv:2106.04560*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Yes.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We discuss the fundamental subjectivity of labels and potential biases in the Limitations section.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We have ensured that our paper conforms with the ethical review guidelines.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We detail the method by which we assessed mistakes. We have prepared a GitHub containing our mistake assessments for the two models for others to verify. We also will release the updated multi-label set for others to build upon our work.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We largely use pre-trained model outputs for most of our analysis. For the ViT-3B model, we provide training details in the Appendix, though we note that reproducing the exact model is not the contribution of our paper. We will, however, release the logits on the full ImageNet validation set for any model we analyzed, if those logits are not available already.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We use Clopper-pearson intervals for our evaluation of ImageNet-M and use significance tests for our distributional comparisons.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Our work is largely based analyzing pre-trained models; we did no training specific to this work except for the fine-tuning required for measuring the impact of de-duplication of validation leakage examples, which we provide details in the appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No] Each asset has information pertaining to the license on their respective websites.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We release our analysis metadata as well as updated multi-labels.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] The authors of the paper generated all released data / metadata and consent to its release.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] We only release metadata statistics about mistakes made on the existing dataset, or add new ImageNet labels to the existing validation set; we do not add any new potentially personally identifiable information beyond what is already present in the ImageNet validation set.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] The authors are the participants, and we provide screenshots of the review tools we used.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]