Evaluating Machine Accuracy on ImageNet

Vaishaal Shankar [*1] **Rebecca Roelofs** [*2] **Horia Mania** [1] **Alex Fang** [1] **Benjamin Recht** [1] **Ludwig Schmidt** [1]

Abstract

We evaluate a wide range of ImageNet models with five trained human labelers. In our year-long experiment, trained humans first annotated 40,000 images from the ImageNet and ImageNetV2 test sets with multi-class labels to enable a semantically coherent evaluation. Then we measured the classification accuracy of the five trained humans on the full task with 1,000 classes. Only the latest models from 2020 are on par with our best human labeler, and human accuracy on the 590 object classes is still 4% and 11% higher than the best model on ImageNet and ImageNetV2, respectively. Moreover, humans achieve the same accuracy on ImageNet and ImageNetV2, while all models see a consistent accuracy drop. Overall, our results show that there is still substantial room for improvement on ImageNet and direct accuracy comparisons between humans and machines may overstate machine performance.

1. Introduction

ImageNet, the most influential data set in machine learning, has helped to shape the landscape of machine learning research since its release in 2009 (Deng et al., 2009; Russakovsky et al., 2015). Methods live or die by their "performance" on this benchmark, measured by how frequently images are assigned the correct label out of 1,000 possible classes. This task is inherently an odd one: seldom do we reduce scene analysis and visual comprehension to a single scalar number. Though models now can nearly perform at 90% accuracy on the ImageNet (Xie et al., 2019b), we do not have much context for what such performance means: what kinds of errors do these models make? Are current models nearing a fundamental Bayes error or is there still room for improvement? Are the models overly sensitive to labeling biases as suggested in recent work (Recht et al., 2019)?

[*]Equal contribution [1]University of California, Berkeley [2]Google Brain. Correspondence to: Vaishaal Shankar <vaishaal@berkeley.edu>.

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

In this paper, we contextualize progress on ImageNet by comparing a wide range of ImageNet models to five trained human labelers. Our year-long experiment consists of two parts: first, three labelers thoroughly re-annotated 40,000 test images in order to create a testbed with minimal annotation artifacts. The images are drawn from both the original ImageNet validation set and the ImageNetV2 replication study of Recht et al. (2019). Second, we measured the classification accuracy of the five trained labelers on the full 1,000-class ImageNet task. We again utilized images from both the original and the ImageNetV2 test sets. This experiment led to the following contributions:

Multi-label annotations. Our expert labels quantify multiple issues with the widely used `top-1` and `top-5` metrics on ImageNet. For instance, about 20% of images have more then one valid label, which makes `top-1` numbers overly pessimistic. To ensure a consistent annotation of all 40,000 images, we created a 400-page labeling guide describing the fine-grained class distinctions. In addition, we enlisted the help of a dog competition judge with 20 years of experience to validate particularly hard instances among the 118 dog breeds in ImageNet. Interestingly, we find that `top-1` accuracy is almost perfectly linearly correlated with multi-label accuracy for the models in our testbed.

Human vs. machine comparison. Building on our multi-label annotations, we find that the highest accuracy achieved by one of our human labelers is comparable to the best model in 2020. Our five labelers are 3% to 9% better than the performance levels from early 2015, when claims of super-human accuracy on ImageNet first appeared (He et al., 2015). Moreover, there are important differences in the errors of trained humans and models. Humans make more mistakes among the fine-grained distinctions between animal species (especially dog breeds) and achieve higher accuracy on the 590 classes of inanimate objects. In contrast, the accuracies of models are more uniform. This shows that there is still room for improving existing classification models: the best human achieves more than 99% accuracy on the object classes for both ImageNet and ImageNetV2, while the best network achieves only 95% and 89% on the object classes in ImageNet and ImageNetV2, respectively (see Tables 1 and 2).

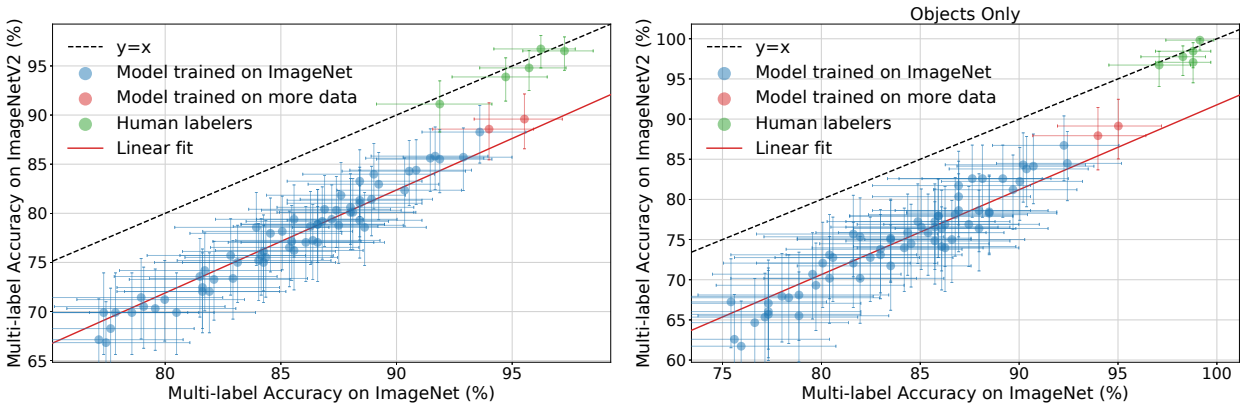

Figure 1. Multi-label accuracies for both CNN models and human participants on the ImageNet validation set versus their accuracies on the ImageNetV2 test set. The left plot shows accuracies on all 1,000 classes, the right plot show accuracies on the 590 object classes. The confidence intervals are 95% Clopper-Pearson confidence intervals.

Table 1. Human and model multi-label accuracy on the ImageNet validation dataset and ImageNetV2. The gap is measured as multi-label accuracy on ImageNet validation minus multi-label accuracy on ImageNetV2. The confidence intervals are 95% Clopper-Pearson intervals. The AdvProp model (Xie et al., 2019a) is an EfficentNet-B8 trained with AdvProp and the FixRes model (Mahajan et al., 2018; Touvron et al., 2019) is a ResNext-32x48d trained on one billion images from Instagram.

ImageNet multi-label accuracy (%)			
Participant	Original	V2	**Gap**
`ResNet50`	84.2 [81.8, 86.4]	75.7 [73.2, 78.7]	**8.4**
`AdvProp`	93.6 [91.9, 95.0]	88.3 [86.5, 90.6]	**5.3**
`FixRes`	95.5 [94.0, 96.7]	89.6 [87.9, 91.8]	**5.9**
Human A	91.9 [90.0, 93.5]	91.1 [89.6, 93.2]	**0.8**
Human B	94.7 [93.1, 96.0]	93.9 [92.6, 95.6]	**0.8**
Human C	96.2 [94.9, 97.3]	96.7 [95.9, 98.1]	**-0.5**
Human D	95.7 [94.3, 96.9]	94.8 [93.7, 96.4]	**0.9**
Human E	97.3 [96.0, 98.2]	96.5 [95.6, 97.9]	**0.7**

Robustness to distribution shift. The common practice in machine learning benchmarks is to draw training and test sets from the same distribution. This setup may favor trained models over human labelers, who are less focused on one specific distribution. To investigate this hypothesis, we labeled images from the original ImageNet validation set and the ImageNetV2 replication experiment. Recht et al. (2019) closely followed the original ImageNet creation process to build a new test set but found that even the best models achieve 11% lower top-1 accuracy than on the original test set. We show that all models in our testbed still see a 5% to 8% accuracy drop for our multi-label annotations. In contrast, all five human labelers have the same accuracy on ImageNet and ImageNetV2 up to 1% (see Figure 1). This demonstrates that robustness to small distribution shifts is a real problem for neural networks. Neural networks have made little to no progress in this dimension over the past decade and even models trained on 1,000 times more data than the standard ImageNet training set (Mahajan et al., 2018; Touvron et al., 2019) do not close this gap.

Based on our investigation, we make the following recommendations for future machine evaluations on ImageNet:

1. Measure multi-label accuracy. While `top-1` accuracy is still a good predictor of multi-label accuracy for models, this is not guaranteed for the future. Moreover, multi-label accuracy is a more meaningful metric for the ImageNet classification task.

2. Report performance on dogs, other animals, and inanimate objects separately. Label noise and ambiguities are a smaller concern on the 590 object classes where human labelers can achieve 99%+ accuracy.

3. Evaluate performance to distribution shift. Our experiments show that the distribution shift from ImageNet to ImageNetV2 does not pose a challenge to humans but leads to substantial accuracy drops for all models. The robustness of humans proves that classifiers with the same accuracy on ImageNet and ImageNetV2 exist. Finding machine models that exhibit similar robustness is an important direction for future research.

Finally, we caution that our human accuracies should not be seen as the best possible performance on this task. We conjecture that longer training can still lead to higher accuracy for humans, and we expect that automated methods will also continue to improve their accuracy on ImageNet. Moreover, classification problems such as ImageNet attempt to render

immeasurable ambiguous, cultural, and subjective aspects of their tasks measurable in a way that does not capture important dimensions of human experience. This makes an overly broad interpretation of such benchmarks problematic and omits important dimensions of human abilities. Nevertheless, we still hope that our multi-label annotations and the associated evaluation methodology will be useful to measure further progress of machine learning methods.

2. Experiment setup

We conducted our experiment in four phases: (i) initial multi-label annotation, (ii) human labeler training, (iii) human labeler evaluation, and (iv) final annotation review. Figure 2 provides a detailed timeline of the experiment. In total, five human labelers participated in the experiment, denoted A through E. All five participants are anonymized authors of this manuscript. While evaluating more humans would have provided additional information, the scale of the experiment made it difficult to incentivize others to invest the time and effort required to familiarize themselves with the 1,000 ImageNet classes and to label thousands of images.

In detail, the four phases of the experiment were:

1. Initial multi-label annotation. Labelers A, B, and C provided *multi-label* annotations for a subset of size 20,000 from the ImageNet validation set and 20,683 images from all three ImageNetV2 test sets collected by Recht et al. (2019). At this point, labelers A, B, and C already had extensive experience with the ImageNet dataset. We further discuss the annotation process in Section 3.

2. Human labeler training. Using a subset of the remaining 30,000 unannotated images in the ImageNet validation set, labelers A, B, C, D, and E underwent extensive training to understand the intricacies of fine-grained class distinctions in the ImageNet class hierarchy. The exact training process is detailed in Section 4.

3. Human labeler evaluation. For the human labeler evaluation, we generated a class-balanced random sample containing 1,000 images from the 20,000 annotated images of the ImageNet validation set and 1,000 images from ImageNetV2. We combined the two sets and randomly shuffled the resulting 2,000 images. Then, the five participants labeled these images over the course of 28 days.

4. Final annotation review. Lastly, all labelers reviewed the additional annotations generated in the human labeler evaluation phase. We discuss the main results from our evaluation in Section 5.

3. Multi-label annotations

In this section, we describe the details of the multi-label annotation process for the ImageNet validation dataset and ImageNetV2. We first explain why multi-label annotations are necessary for proper accuracy evaluation on ImageNet by outlining the pitfalls of the two most widely used accuracy metrics, `top-1` and `top-5`.

`Top-1` accuracy. `Top-1` accuracy is the standard accuracy measure used in the classification literature. It measures the proportion of examples for which the predicted label matches the single target label. However, the assumption that each image has a single ground truth label from a fixed set of classes is often incorrect. ImageNet images, such as Figure 3a, often contain multiple objects belonging to different classes (e.g. `desk`, `laptop`, `keyboard`, `space bar`, `screen`, and `mouse` frequently all appear in the same image). Moreover, even for images for which a class is prominent the ImageNet label might refer to another class present in the image. For example, in Figure 3b the class `gown` is central and appears in the foreground, but the ImageNet label is `picket fence`. As a result, one is not guaranteed to achieve high `top-1` accuracy by identifying the main objects in images. In other words, `top-1` accuracy can be overly stringent by penalizing predictions that appear in the image but do not correspond to the target label.

`Top-5` accuracy. To partially remedy issues with `top-1`, the organizers of the ImageNet challenge (Russakovsky et al., 2015) measured `top-5` accuracy, which considers a classification correct if *any* of the five predictions matches the target label. However, allowing five guesses on *all* images on fine-grained classification tasks such as ImageNet can make certain class distinctions trivial. For example, there are five turtles in the ImageNet class hierarchy (`mud turtle`, `box turtle`, `loggerhead turtle`, `leatherback turtle`, and `terrapin`), which can be difficult to distinguish, but given an image of a turtle, a classifier can guess all five turtle classes to ensure that it predicts the correct label.

Multi-label accuracy. For multi-label accuracy, every image has a set of target labels and a prediction is marked correct if it corresponds to *any* of the target labels for that image. Due to the limitations of `top-1` and `top-5` accuracy, as well as ambiguity in the target class for many images, multi-label annotations are necessary for rigorous accuracy evaluation on ImageNet.

3.1. Types of multi-label annotations

Next, we discuss three categories of multi-label annotations that arose in our study, exemplified in Figure 3.

Multiple objects or organisms. For images that contain multiple objects or organisms corresponding to classes in the ImageNet hierarchy, we added an additional target label

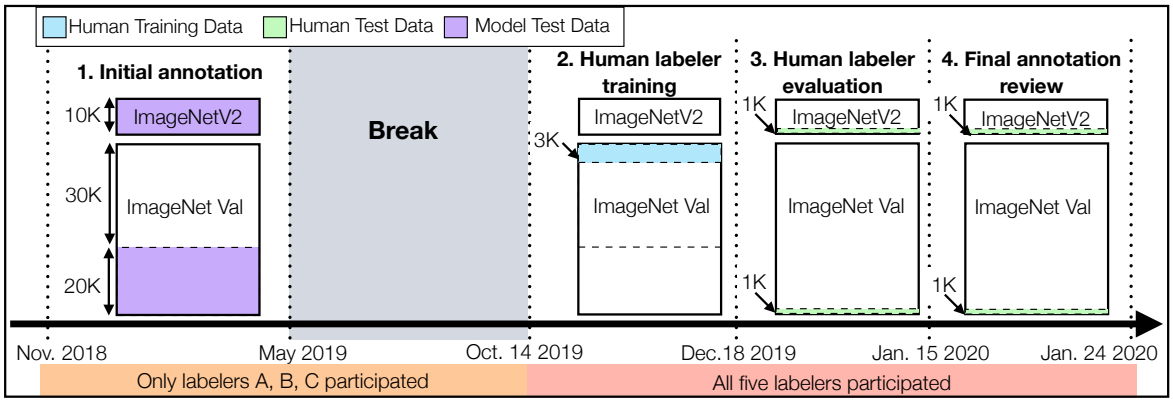

Figure 2. Timeline for the four phases of our experiment. 1) *Initial multi-label annotation:* First, starting in November 2018, human labelers A, B, and C annotated a set of images from ImageNetV2 and the ImageNet validation set with multi-label annotations. 2) *Human labeler training:* Then, after a long break, on October 14, 2019, all five participants began training on a 3,000 image subset of the original ImageNet validation set. (Humans were *not* trained on ImageNetV2.) 3) *Human labeler evaluation:* Next, starting on December 18, 2019, humans labeled 2,000 images from a random, class-balanced sample including 1,000 images from the ImageNet validation dataset and 1,000 images from ImageNetV2. The evaluation dataset did not include any of the images used in training. 4) *Final annotation review:* Finally, all five labelers reviewed the annotations collected for the 2,000 image evaluation dataset.

for each entity in the scene. For example, Figure 3a shows an image with target label `desk` that also contains multiple different objects corresponding to ImageNet classes. When there are multiple correct objects or organisms, the target class does not always correspond to the most central or largest entity in the scene. For example, in Figure 3b, the target class `picket fence` appears in the background of the image, but classes `groom`, `bow tie`, `suit`, `gown`, and `hoopskirt` all appear in the foreground.

Synonym or subset relationships. If two classes are synonyms of each other, or a class is a subset of another class, we considered both classes to be correct target labels. For example, the ImageNet class `tusker` is defined as any animal with visible tusks. Since `warthog`, `African elephant` and `Indian elephant` all have prominent tusks, these classes are all technically subsets of `tusker`. Figure 3c shows an `African elephant` that additionally has `tusker` as a correct label.

Unclear images. In certain cases, we could not ascertain whether a label was correct due to ambiguities in the image or in the class hierarchy. Figure 3d shows a scene which could arguably be either a `lakeshore` or a `seashore`.

3.2. Collecting multi-label annotations

Next, we detail the process we used to collect multi-label annotations. We first collected the `top-1` predictions of 72 pre-trained ImageNet models published from 2012 to 2018. Then, over a period of three months, participants A, B and C reviewed all predictions made by the models on 40,683 images from ImageNet and ImageNetV2. Participants first researched class distinctions extensively – the details of this research are covered in 4. The three participants then categorized *every* unique prediction made by the 72 models on the 40,683 images (a total of 182,597 unique predictions) into `correct` or `incorrect`, thereby allowing each image to have multiple `correct` labels.

In total, we found that 18.2% of the ImageNet validation images have more than one correct label. Among images with multiple correct labels, the mean number of correct labels per image is 2.3.

The multi-label accuracy metric. Multi-label accuracy is computed by counting a prediction as correct if and only if it was marked correct by the expert reviewers during the annotation stage. We note that we performed a second annotation stage after the human labelers completed the experiment, as explained in Section 4.3.

In Figure 4, we plot each model's `top-5` and `top-1` accuracy versus its multi-label accuracy. *Every* model prediction was reviewed individually for correctness. Higher `top-1` and `top-5` accuracy correspond to higher multi-label accuracy with relatively few changes in model rankings across the different metrics. However, for all models, `top-1` accuracy underestimates multi-label accuracy (models see a median improvement of 8.9% when comparing multi-label to `top-1`) while `top-5` overestimates multi-label accuracy (models see a median drop of 7.4% when comparing multi-label accuracy to `top-5`). While multi-label accuracy is highly correlated with `top-1`and `top-5`accuracy, we assert that neither `top-1`nor `top-5`measure a semantically meaningful notion of accuracy.

(a) desk (b) picket fence (c) African elephant (d) lakeshore

Figure 3. Examples from the ImageNet validation of scenarios where multi-label annotations are necessary. *Multiple objects or organisms:* In Figure 3a, the ImageNet label is desk but screen, monitor, coffee mug and many more objects in the scene could count as correct labels. Figure 3b shows a scene where the target label picket fence is counterintuitive because it appears in the background of the image while classes groom, bowtie, suit, gown, and possibly hoopskirt are more prominently displayed in the foreground. b) *Synonym or subset relationships:* This image has ImageNet label African elephant, but can be labeled tusker as well, because every African elephant with tusks is a tusker. c) *Unclear images:* This image is labeled lakeshore, but could also be labeled seashore as there is not enough information in the scene to distinguish the water body between a lake or sea.

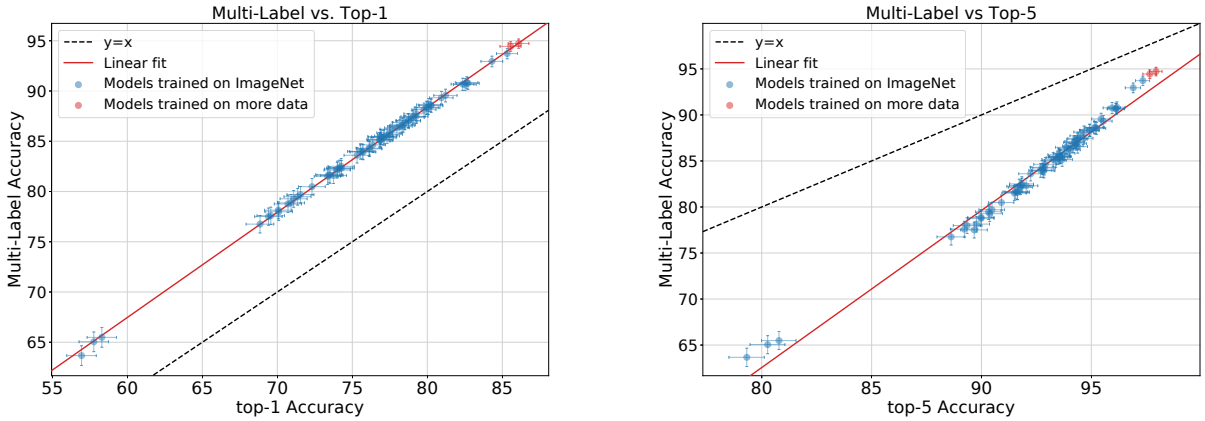

Figure 4. The relationship between top-1, top-5, and multi-label accuracy on ImageNet test for all 72 models in our test bed. The left figure plots multi-label vs. top-1 accuracy accuracy. Multi-label accuracy makes the task easier than top-1 accuracy, with a median improvement of 8.9% between top-1 and multi-label scores. The right figure plots multi-label vs. top-5 accuracy accuracy. Multi-label accuracy is more stringent than top-5 accuracy, with a median drop of 7.4% between top-5 and multi-label scores.

4. Human accuracy measurement process

We now describe the human evaluation portion of our experiment. Annotators A, B, & C participated in the initial annotation review and thus saw all 40,683 evaluation images and labels from ImageNet and ImageNetV2. To remedy the possibility that annotators A, B & C unintentionally memorized the evaluation labels, two precautions were taken. First, annotators A, B, & C did not look at the data for six months. Second, we introduced annotators D & E, neither of whom had seen the test images prior to evaluation.

4.1. Human labeler training

After a six month period of inactivity, in October 2019, all five participants began a training regimen for the labeling task. Previously, participants A, B, C undertook a similar training for the initial multi-label annotation review. All training was carried out using a the 30,000 ImageNet validation images that would not be used for the final evaluation. The primary goal of training was to familiarize humans with the ImageNet class hierarchy.

The initial human accuracy study by Russakovsky et al. (2015) details three main failure modes of humans: fine-grained distinctions, class unawareness, and insufficient training images. We address all three failure modes with our training regimen:

Fine-grained distinctions. There are many difficult class distinctions in ImageNet, but humans tend to struggle with fine-grained distinctions within the 410 animal classes and 118 dog classes. Even the scientific community disagrees about the exact taxonomy of specific species. For instance, while tiger beetles are often classified as a subfamily

of `ground beetle`, this classification isn't universally accepted among entomologists (bug, 2019; Wikipedia contributors, 2019). Similar issues arise in other animal families, such as the mustelines, monkeys, and wolves.

To help humans perform well on fine-grained class distinctions, we created training tasks containing only images from certain animal families. The training tasks gave labelers immediate feedback on whether they had made the correct prediction or not. These targeted training tasks were created after labelers identified classes for which they wanted additional training. Labelers trained on class-specific tasks for dogs, insects, monkeys, terriers, electric rays and sting rays, and marmots and beavers. After training, labelers reviewed each other's annotations as a group and discussed the class distinctions. Labelers also wrote a labeling guide containing useful information for distinguishing similar classes, discussed in more detail in Section 4.2.

Information from the American Kennel Club (akc) was frequently used to understand and disambiguate difficult dog breeds. We also reached out to a member of the local chapter of the club for aid with dog identification. Since some dogs may be mixed-breeds, it may be impossible to disambiguate between similar dog breeds from pictures alone. Fortunately, the ImageNet dog labels are of high quality as they are derived from the Flickr image description, which are often authored by the owner of the dog.

Class unawareness. For the 590 object categories in ImageNet, *recall* is the primary difficulty for untrained humans. To address this, we built a labeling user interface that allowed annotators to either search for a specific ImageNet class or explore a graphical representation of the ImageNet classes based on the WordNet (Miller, 1995) hierarchy.

Insufficient training images. The two annotators in (Russakovsky et al., 2015) trained on 500 and 100 images respectively, and then had access to 13 training images per class while labeling. In our experiment, human labelers had access to 100 training images per class while labeling.

4.2. Labeling guide

During training, the participants constructed a *labeling guide* that distilled class specific analysis learned during training into key discriminative traits that could be referenced by the labelers during the final labeling evaluation. The labeling guide contained detailed entries for 431 classes.

4.3. Final evaluation and annotation review.

On December 18th 2019, $1{,}000$ images were sampled from ImageNet Validation and $1{,}000$ images were sampled from ImageNetV2 and shuffled together. The datasets were sampled in a class balanced manner.

Between December 19th 2019 and January 16th 2020 all 5 participants labeled 2,000 images in order to produce the main results of this work. The only resources the labelers had access to during evaluation were 100 randomly sampled images from the ImageNet training set for each class, and the labeling guide. The participants spent a median of 26 seconds per image, with a median labeling time of 36 hours for the entire labeling task.

After the labeling task was completed, an additional multi-label annotation session was necessary. Since each image only contained reviewed labels for classes predicted by *models*, to ensure a fair multi-label accuracy, the human predictions for the 2,000 images had to be manually reviewed. To minimize bias, participants were not allowed to view their predicted labels after the task, and random model predictions were seeded into the annotation review such that every image had both model and human predictions to be reviewed. Compared to labels from the initial annotation review from November 2018, after the final annotation review, labels were unchanged for 1320 images, added for 531 images, and modified for 239 images. The modifications were due to a much greater knowledge of fine-grained class distinctions by the participants after the training phase.

5. Main Results

In this section we discuss two key facets of our experimental findings: a comparison of human and machine accuracies on ImageNet, and a comparison of human and machine robustness to the distribution shift between the ImageNet validation set and the ImageNet-V2 test set. We also consider these comparisons on three restricted sets of images.

The main results of our work are illustrated in Figure 1. We can see that all the human labelers fall close to the dotted line, indicating they their accuracies on the two datasets are within 1%. Moreover, we can see that the accuracies of three of the human labelers are better than the performance of the best model on both the original ImageNet validation set and on the ImageNet-V2 test set. Importantly, we note that labelers D and E, who did not participate in the initial annotation period, performed better than the best model.

Figure 1 shows that the ImageNet validation set confidence intervals of the best 4 humans labelers and of the best model overlap. However, McNemar's paired test rejects the null hypothesis that the `FixResNeXt` model (the best model) and Human E (the best human labeler) have the same accuracy on the ImageNet validation set distribution with a p-value of 0.037. In Figure 1 we observe that the confidence intervals of Humans C, D, and E on the ImageNetV2 test set do not overlap with the confidence interval of the best model. McNemar's test between Human B and the `FixResNeXt` model on ImageNetV2 yields a p-value of 2×10^{-4}.

Difficult images: One of the benefits of our experiments is

Table 2. Human and model multi-label accuracy on three subsets of the ImageNet and ImageNetV2 test sets. These results suggest that human labelers have an easier time identifying objects than dogs and organisms. Moreover, human labelers are highly accurate on images on which they spent little time to assign a label.

ImageNet multi-label accuracy (%)

	All Images		Without Dogs		Objects Only		Fast Images	
Participant	Original	V2	Original	V2	Original	V2	Original	V2
`resnet50`	84.2	75.7	84.9	76.8	82.5	72.8	86.8	79.6
`AdvProp`	93.6	88.3	94.1	89.3	92.3	86.7	94.9	91.3
`FixResNeXt`	95.5	89.6	96.0	90.1	95.0	89.1	96.2	92.3
Human A	91.9	91.1	94.2	93.4	97.0	96.7	97.6	97.5
Human B	94.7	93.9	96.9	96.0	98.3	97.8	98.5	98.5
Human C	96.2	96.7	98.4	98.6	99.1	99.8	99.1	99.7
Human D	95.7	94.8	97.3	96.6	98.8	98.4	99.3	98.3
Human E	97.2	96.5	98.7	97.3	98.8	97.0	99.5	98.6

the potential insight into the failure modes of image classification models. To have a point of comparison let us start with the human labelers. There were 10 images which were misclassified by *all* human labelers. These images consisted of one image of a monkey and nine images of dogs. On the other hand, there were 27 images misclassified by *all* 72 models considered by us. Interestingly, 19 out of these images correspond to object classes and 8 correspond to organism classes. We note that there are only two images that were misclassified by all models and human labelers, both of them containing dogs. Four of the 27 images which were difficult for the models are displayed in Figure 5. It is interesting that the failure cases of the models consist of many images of objects while the failure cases of human labelers are exclusively images of animals.

Figure 5. Four images which were misclassified by all 72models, two from ImageNet (top left and bottom right) and two from ImageNetv2. The target correct labels for these images are `cup`, `yawl`, `nail`, and `spotlight`

Accuracies without dogs: To understand the extent to which models are better than the human labelers at classifying dogs and animals, we compute their accuracies on two restricted sets of images. First, we computed accuracies by excluding the 118 dog classes. In this case, Table 2 shows an increase in the accuracy of the best model ((Touvron et al., 2019)) by 0.6% on ImageNet images and by 1.1% on ImageNetV2 images. However, the mean increase of the human labelers' accuracies is 1.9% on ImageNet and 1.8% on ImageNetV2. Before we interpret this result, we must establish whether the changes in accuracies shown in Table 2 are meaningful. There are 882 non-dog classes in ImageNet. We use the bootstrap to estimate changes in accuracies when the data is restricted to 882 classes. We compute accuracies over 1000 trials as follows: we sample without replacement 882 classes and compute the accuracies of the human labelers on the images whose main labels are in the sampled classes. All trials yield smaller changes in accuracy than those shown in Table 2. This simulation indicates that the increase in human performance on non-dog images is significant.

Therefore, the relative gap between human labelers and models increases on both ImageNet and ImageNetV2 when we remove the images containing dogs. This suggests that the dog images are more difficult for the human labelers participating in our experiment than for the models.

Accuracies on objects: To further understand the strengths and weaknesses of the models and human labelers, we compute their accuracies on the subset of data which have objects as their main labels, as opposed to organisms. There are 590 object classes. In Table 2 we can see the stark contrast in performance between human labelers and models on images of objects. The mean increase of the human labelers' accuracies is 3.3% on ImageNet and 3.4% on ImageNetV2, whereas the accuracy of the best model decreased by 0.5%

on both ImageNet and ImageNetV2. A bootstrap simulation similar to the one described for the "Without Dogs" comparison reveals that human accuracy increase is significant. This result suggests that images of objects are substantially easier for the human labelers than the models.

Accuracies on fast images: Whereas CNN models spend the same amount of time classifying different images, the human labelers spent anywhere from a couple of seconds to 40 minutes labeling one image. What does the amount of time spent by humans labeling an image say about that image? We compute accuracies of all models and human labelers on the subset of images for which the median time spent by the human labelers to label it was at most 60 seconds. Out of a total of 2000 images used in the evaluation, there are 756 such images from ImageNet (77% of images) and 714 such images from ImageNetV2 (73% of images). We observe a dramatic increase in the accuracies of the human labelers, suggesting that human labelers know when an image is difficult for them and spend more time labeling it. The accuracies of the models also increase on "Fast Images." This result is intuitive, suggesting that images that humans label quickly are more likely to be correctly classified by models. We present results for these images in Table 2.

6. Related Work

Human accuracy on ImageNet. In the context of the ImageNet challenge, Russakovsky et al. (2015) studied the accuracy of two trained humans on $1{,}500$ and 258 ImageNet images, respectively. The widely publicized human baseline on ImageNet is the top-5 accuracy of the labeler who labeled 1,500 images. As mentioned in the introduction, our study goes beyond their comparison in three ways: multi-label accuracy, more labelers, and a focus on robustness to small distribution shift. While some of our findings differ, other results from (Russakovsky et al., 2015) are consistent with ours. For instance, both experiments found that the time spent per image was approximately one minute, with a long tail due to difficult images.

Human performance in computer vision broadly. There have been several recent studies of humans in various areas of computer vision. For instance, Elsayed et al. (2018) construct adversarial examples that fool both models and *time-limited* humans. Geirhos et al. (2017) conducted psychophysical trials to investigate human robustness to synthetic distribution shfits, and Geirhos et al. (2018) studied characteristics used by humans to make object recognition decisions. In a similar vein, Zhu et al. (2016) contrast the effect of foreground and background objects on performance by humans and trained models.

Multi-label annotations. In this work, we contribute multi-label annotations for ImageNet and ImageNetV2. Previously, Stock & Cissé (2017) studied the multi-label nature of ImageNet and found that `top-1` accuracy can underestimate multi-label by as much as 13.2%. The results of this work largely agree with our study. We hope the public release of our multi-label annotations will allow an accurate evaluation of all future models.

ImageNet inconsistencies and label error. During our annotation review, we recorded all incorrectly classified images we found in ImageNet and ImageNetV2. With the help of experts from the Cornell Lab of Ornithology,Van Horn et al. (2015) estimate that at least 4% of birds are misclassified in ImageNet. Within the bird classes, (Van Horn et al., 2015) also observe inconsistencies in ImageNet's taxonomic structure which lead to weak class boundaries. We found that these taxonomic issues are present in the majority of the fine-grained organism classes.

Distribution shift. There is a growing body of work studying methods for addressing the challenge of distribution shift. For instance, the goal of distributionally robust optimization (DRO) is to find models that minimize the worst case expected error over a set of probability distributions (Abadeh et al., 2015; Ben-Tal et al., 2013; Delage & Ye, 2010; Duchi et al., 2019; Esfahani & Kuhn, 2018; Sagawa et al., 2019; Sinha et al., 2017). A similar yet different line of work has focused on finding models that have low error rates on adversarial examples (worst case small perturbations to data points in the test set) (Biggio & Roli, 2018; Biggio et al., 2013; Goodfellow et al., 2014; Madry et al., 2017). The work surrounding both DRO and adversarial examples has developed valuable ideas, but neither line of work has been shown to resolve the drop in accuracy between ImageNet and ImageNetV2.

7. Conclusion & Future Work

Achieving truly reliable machine learning will require a deeper understanding of the input changes a model should be robust to. Such understanding can likely guide research on more robust methods and is essential for developing meaningful tests of reliable performance. For tasks where human-like generalization is the ultimate goal, comparing model performance to human generalization can provide valuable information about the desired robustness properties. Our work is a step in this direction. Our results highlight that robustness to small, naturally occuring distribution shifts is a performance dimension not addressed by current benchmarks, but easily handled by humans. Besides the obvious direction of improving model robustness to such distribution shifts, there are further avenues for future work:

Robustness of non-expert labelers. A natural question is whether labelers with less training exhibit similar robustness to the distribution shift from ImageNet to ImageNetV2.

Since untrained labelers will likely be in a lower accuracy regime, this would further illustrate that human robustness is a more stable property than direct accuracy measurements.

Other generalization dimensions. What further dimensions of human generalization are current models clearly lacking? Other forms of natural distribution shift such as robustness to temporal changes could be one candidate (Gu et al., 2019; Shankar et al., 2019).

Acknowledgements

We would like to thank Nicholas Carlini, Moritz Hardt, Andrej Karpathy, Henry Martin, Christopher Re, Ashia Wilson, and Bin Yu for valuable conversations while working on this project. Finally, we would like to thank Sandra Pretari Hickson for sharing her rich expertise concerning dog breeds. She helped us clarify distinctions between similar dog breeds and identify dogs in particularly difficult images.

This research was generously supported in part by ONR awards N00014-17-1-2191, N00014-17-1- 2401, and N00014-18-1-2833, the DARPA Assured Autonomy (FA8750-18-C-0101) and Lagrange (W911NF-16-1-0552) programs, a Siemens Futuremakers Fellowship, an Amazon AWS AI Research Award.

References

American kennel club. URL https://www.akc.org/.

Subfamily cicindelinae - tiger beetles, Oct 2019. URL https://bugguide.net/node/view/375.

Abadeh, S. S., Esfahani, P. M. M., and Kuhn, D. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pp. 1576–1584, 2015.

Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Biggio, B. and Roli, F. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 2018. https://arxiv.org/abs/1712.03141.

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 *Machine Learning and Knowledge Discovery in Databases*, 2013. https://link.springer.com/chapter/10.1007/978-3-642-40994-3_25.

Chen, Y., Li, J., Xiao, H., Jin, X., Yan, S., and Feng, J. Dual path networks. In *Neural Information Processing Systems (NIPS)*, 2017. https://arxiv.org/abs/1707.01629.

Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. https://arxiv.org/abs/1610.02357.

Clinchant, S., Csurka, G., Perronnin, F., and Renders, J.-M. XRCE's participation to ImagEval. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.102.6670&rep=rep1&type=pdf, 2007.

Delage, E. and Ye, Y. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. http://www.image-net.org/papers/imagenet_cvpr09.pdf.

Duchi, J. C., Hashimoto, T., and Namkoong, H. Distributionally robust losses against mixture covariate shifts. *Under review*, 2019.

Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., and Sohl-Dickstein, J. Adversarial examples that fool both computer vision and time-limited humans. In *Advances in Neural Information Processing Systems*, pp. 3910–3920, 2018.

Esfahani, P. M. and Kuhn, D. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., and Wichmann, F. A. Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969*, 2017.

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. *CoRR*, abs/1811.12231, 2018. URL http://arxiv.org/abs/1811.12231.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Gu, K., Yang, B., Ngiam, J., Le, Q. V., and Shlens, J. Using videos to evaluate image model robustness. *CoRR*, abs/1904.10076, 2019. URL http://arxiv.org/abs/1904.10076.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on

imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016a. https://arxiv.org/abs/1512.03385.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)*, 2016b. https://arxiv.org/abs/1603.05027.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. https://arxiv.org/abs/1704.04861, 2017.

Hu, J., Shen, L., and Sun, G. Squeeze-and-excitation networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. https://arxiv.org/abs/1709.01507.

Huang, G., Liu, Z., Weinberger, K. Q., and van der Maaten, L. Densely connected convolutional networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. https://arxiv.org/abs/1608.06993.

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 <0.5MB model size. https://arxiv.org/abs/1602.07360, 2016.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, 2015. https://arxiv.org/abs/1502.03167.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012. https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.

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 *European Conference on Computer Vision (ECCV)*, 2018. https://arxiv.org/abs/1712.00559.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., and van der Maaten, L. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 181–196, 2018.

Miller, G. A. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Perronnin, F., Sánchez, J., and Mensink, T. Improving the Fisher kernel for large-scale image classification. In *European Conference on Computer Vision (ECCV)*, 2010. URL https://www.robots.ox.ac.uk/~vgg/rg/papers/peronnin_etal_ECCV10.pdf.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? 2019. URL http://arxiv.org/abs/1902.10811.

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

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Shankar, V., Dave, A., Roelofs, R., Ramanan, D., Recht, B., and Schmidt, L. A systematic framework for natural perturbations from videos. *CoRR*, abs/1906.02168, 2019. URL http://arxiv.org/abs/1906.02168.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. https://arxiv.org/abs/1409.1556, 2014.

Sinha, A., Namkoong, H., and Duchi, J. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

Stock, P. and Cissé, M. Convnets and imagenet beyond accuracy: Explanations, bias detection, adversarial examples and model criticism. *CoRR*, abs/1711.11443, 2017. URL http://arxiv.org/abs/1711.11443.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. https://arxiv.org/abs/1409.4842v1.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the Inception architecture for computer

vision. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. https://arxiv.org/abs/1512.00567.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. Inception-v4, Inception-Resnet and the impact of residual connections on learning. In *Conference On Artificial Intelligence (AAAI)*, 2017. https://arxiv.org/abs/1602.07261.

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

Touvron, H., Vedaldi, A., Douze, M., and Jégou, H. Fixing the train-test resolution discrepancy. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

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

Wikipedia contributors. Tiger beetle — Wikipedia, the free encyclopedia, 2019. URL https://en.wikipedia.org/w/index.php?title=Tiger_beetle&oldid=932794435. [Online; accessed 1-February-2020].

Xie, C., Tan, M., Gong, B., Wang, J., Yuille, A., and Le, Q. V. Adversarial examples improve image recognition, 2019a.

Xie, Q., Hovy, E., Luong, M.-T., and Le, Q. V. Self-training with noisy student improves imagenet classification. *arXiv preprint arXiv:1911.04252*, 2019b.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. https://arxiv.org/abs/1611.05431.

Yalniz, I. Z., Jégou, H., Chen, K., Paluri, M., and Mahajan, D. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019.

Zhang, X., Li, Z., Change Loy, C., and Lin, D. Polynet: A pursuit of structural diversity in very deep networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. https://arxiv.org/abs/1611.05725.

Zhu, Z., Xie, L., and Yuille, A. L. Object recognition with and without objects. *CoRR*, abs/1611.06596, 2016. URL http://arxiv.org/abs/1611.06596.

Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. Learning transferable architectures for scalable image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. https://arxiv.org/abs/1707.07012.