Human uncertainty makes classification more robust

Joshua C. Peterson*, Ruairidh M. Battleday*, Thomas L. Griffiths, Olga Russakovsky
Princeton University, Department of Computer Science
{joshuacp,battleday,tomg,olgarus}@cs.princeton.edu

Abstract

The classification performance of deep neural networks has begun to asymptote at near-perfect levels. However, their ability to generalize outside the training set and their robustness to adversarial attacks have not. In this paper, we make progress on this problem by training with full label distributions that reflect human perceptual uncertainty. We first present a new benchmark dataset which we call CIFAR10H, containing a full distribution of human labels for each image of the CIFAR10 test set. We then show that, while contemporary classifiers fail to exhibit human-like uncertainty on their own, explicit training on our dataset closes this gap, supports improved generalization to increasingly out-of-training-distribution test datasets, and confers robustness to adversarial attacks.

1. Introduction

On natural-image classification benchmarks, state-of-the-art convolutional neural network (CNN) models have been said to equal or even surpass human performance, as measured in terms of "top-1 accuracy"—the correspondence between the most probable label indicated by the model and the "ground truth" label for a test set of held-out images. As accuracy gains have begun to asymptote at near-perfect levels [11], there has been increasing focus on out-of-training-set performance—in particular, the ability to generalize to related stimuli [39], and robustness to adversarial examples [29]. On these tasks, by contrast, CNNs tend to perform rather poorly, whereas humans continue to perform well.

To redress this problem, and provide a better standard for training classifiers, we suggest an alternative objective: not just trying to capture the most likely label, but trying to capture the full distribution over labels. Errors in classification can be just as informative as the correct answers—a network that confuses a dog with a cat, for example, might be judged to generalize better than one that confuses it with a truck (see [1]). Indeed, consider the examples shown in Figure 1, in which the CNN can be underconfident, overconfident, or systematically incorrect, and yet receive a perfect accuracy score. Capturing this similarity structure is a key part of effective generalization [19], and an important consideration when building classification models for real-world applications, for example, object avoidance in driverless cars.

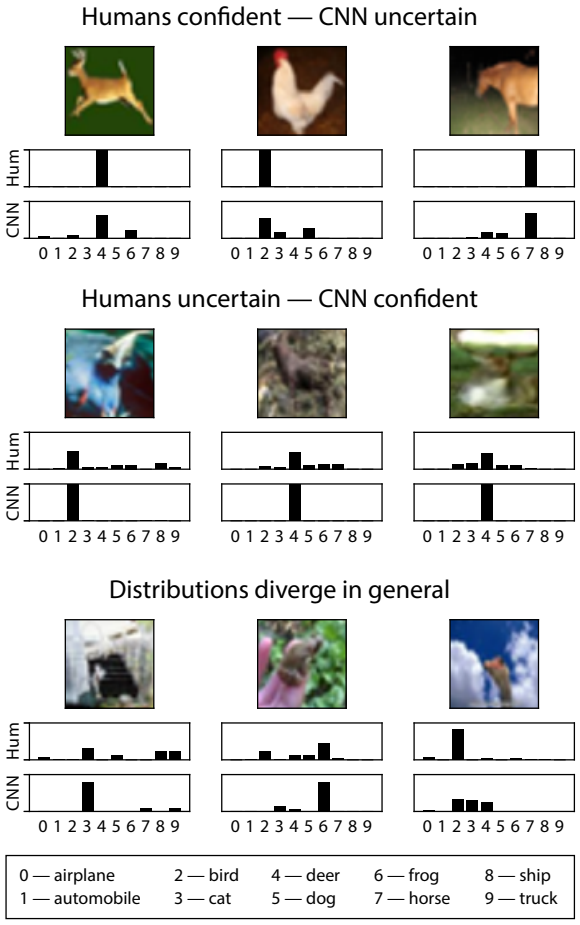

Figure 1: CIFAR10 images for which humans and our best traditionally-trained CNN (Shake-Shake [11]) agree in their top guess, but systematically differ over other choices.

* contributed equally

Predicting more complete distributions of labels requires first measuring those distributions. Given that we cannot directly extract ground truth perceptual similarity from the world, human categorization behavior is a natural candidate for such a comparison. Indeed, there is often a lack of human consensus on the category of an object, and human errors often convey important information about the structure of the visual world [31]. Beyond complementing training paradigms, collecting these full label distributions from humans to better model human biases and predict their errors is interesting in itself—this time, for example, to help a driverless car infer the actions of nearby human drivers. Finally, although there has been much work scaling the number of images in datasets [18], and investigating label noise [40, 12, 48], little effort has been put into identifying the benefits from increasing the richness of (informative) label distributions for image classification tasks.

To these ends, we make the following contributions:

- We present a novel soft-label dataset which we call `CIFAR10H`, comprising full label distributions for the entire 10,000-image `CIFAR10` test set, utilizing over 500k crowdsourced human categorization judgments.
- We show that when state-of-the-art CNN classifiers are trained using these soft labels, they generalize better to out-of-sample datasets than hard-label controls.
- We present a performance benchmark assessing model fit to human labels, and show that models trained using alternative label distributions do not approximate human uncertainty as well.
- We show that when CNNs are trained to perform well on this benchmark they are significantly more resistant to adversarial attacks.

Taken together, our results support more fine-grained evaluations of model generalization behavior and demonstrate the potential utility of one method for integrating human perceptual similarity into paradigms for training classifiers.

2. Related Work

Hierarchical Classification. Work on using class confusion or hierarchy to improve classification accuracy or robustness dates back to early works of *e.g.,* Griffin and Perona [14], Marszalek and Schmid [34], or Zweig and Weinshall [53]. Class label hierarchies have been used to enable *e.g.,* sharing of representations [47, 9, 22], effective combination of models [23], or improved accuracy of classification through hierarchical prediction [32, 8]. Benchmarks have occasionally proposed using hierarchical metrics for evaluation (*e.g.,* the hierarchical error rate of ILSVRC 2010 and 2011 [41]). Overall though the dominant paradigm has focused on evaluating the top-K accuracy rather analyzing the errors of the system, and the hierarchical structure has been used mostly for training. We argue it is time to rethink this. First, modern large-scale open-world complex datasets no longer guarantee non-overlapping object classes [26], making hierarchical class confusion particularly meaningful. Second, existing methods are becoming remarkably good at top-K accuracy, so an increasing focus on their robustness with regard to adversarial examples [44, 13, 2] or distributional shift [45, 39] is warranted. In this work we present to our knowledge the first large-scale evaluation of generalization to human uncertainty in image classification.

Knowledge Distillation. The label hierarchies used to aid recognition can be manually constructed [6, 3], derived from linguistic knowledge bases [10, 9], or learned automatically [14, 19]. Our work is closest to the former (manual construction), although instead of explicitly constructing a class hierarchy we rely on human confusion between the classes to infer the relationship between the classes for a given image. While being derived from *human* confusion, our work bears some resemblance to the knowledge distillation approach of [19]. In knowledge distillation, these labels are provided by the smoothed softmax probabilities from a pre-trained classification model. When soft labels are combined with ground truths, a form of model transfer and compression is achieved, because the softmax probabilities carry crucial information. The rationale for this process is similar to our own: networks (and humans) gain great robustness from distilling important information about similarity structure into the distributions we infer over images and their categories. However, the use of a network to provide them (*i.e.,* the standard application of knowledge distillation) is itself problematic without a gold standard to compare to: there is no guarantee that the similarity structure a model has learned is correct.

Soft Labels. One of the core contributions of our work is around using the soft labels provided through human confusion as a replacement for one-hot label encodings. Several methods have been proposed as alternatives to one-hot encodings, *e.g.,* using heuristics to smooth the top-1 label during large-scale 1000+ way classification [43] or incorporating test-time human uncertainty into a collaborative computer vision system [4]. *mixup* [51] is another recently developed method for automatically generating soft labels based on convex combinations of pairs of examples and their hard labels, and has been shown to improve generalization and adversarial robustness while reducing memorization. However, since the linearity constraint is constant across all pairs of classes, and the labels are one-hot, it is difficult to see how the softness in such labels is a full measure of perceptual likeness.

Human studies. Lastly, there are a number of studies that also use human experts to provide distributional information over training labels in related classification fields, such as medical diagnosis systems [35, 36]. While the theoretical cases these studies present support our own, they do not provide a large-scale testbed for evaluation of other classification models. Notably, the human uncertainty labels frequently don't need to be explicitly collected but will become automatically available in the process of data collection. Much of crowdsourcing work focuses on *reconciling* human labels and mitigating their disagreement (c.f., Kovashka et al. [25] for a survey). Our approach proposes utilizing these human disagreements to improve the accuracy and robustness of a model, complementing existing work aimed at leveraging "errors" in human labeling [27].

3. From Labels to Label Distributions

The standard practice for image classification tasks is to train using "ground truth" labels provided in common benchmark datasets, for example, `ILSVRC12` [41], and `CIFAR10` [28], where the "true" category for each image is decided through human consensus (the *modal* choice) or by the database creators. Although a useful simplification in many cases, we suggest that this approximation introduces a bias into the learning framework that has important distributional implications. To see this, first consider the standard loss minimization objective during training given below:

$$\min_{\theta} \sum_{i=1}^{n} \mathcal{L}(f_\theta, x_i, y_i), \tag{1}$$

in which the loss $\mathcal{L}$ for a model with parameters θ is minimized with respect to observed data samples $\{x_i, y_i\}_{i=1}^n$. Our goal in training a model in this way is to generalize well to unseen data: to minimize the expected loss over unobserved labels given observed images $\{x_j\}_{j=1}^m$ drawn from the same underlying data distribution in the future:

$$\frac{1}{m} \sum_{j=1}^{m} \sum_{c} \mathcal{L}(f_\theta, x_j, y_j = c)\, p(y_j = c|x_j). \tag{2}$$

When we consider the second term in this product, we can see that using modal labels during dataset construction would only be an optimal estimator if for any stimulus x, the underlying conditional data distribution $p(y|x)$ is zero for every category c apart from the one assigned by human consensus. By contrast, when we consider the network and human confusions seen in Figure 1, we can see there do exist cases in which this assumption violates human allocation of probabilities.

How, then, can we reach a more natural approximation of $p(y|x)$? For some problems, it is easy to just sample from some real set of data $p(x, y)$, but for image classification, we must rely on humans as a gold standard for providing a good estimate of $p(y|x)$. If we expect the human image label distribution $p_{\text{hum}}(y|x)$ to better reflect the natural distribution over categories given an image, we can use it as an improved estimator for $p(y|x)$.

In the case where $f_\theta(x)$ is a distribution $p_\theta(y|x)$ and $\mathcal{L}(f, x, y)$ is the negative log-likelihood, the expected loss reduces to the cross-entropy between the human distribution and that predicted by the classifier:

$$-\frac{1}{m} \sum_{j=1}^{m} \sum_{c} p_{\text{hum}}(y_j = c|x_j) \log p_\theta(y_j = c|x_j). \tag{3}$$

This implies that the optimal strategy for gathering training pairs $\{x_i, y_i\}_{i=1}^n$ is to sample them from $p_{\text{hum}}(y|x)$. Our dataset provides this distribution directly, so that models may be trained on human labels or evaluated against them, or better approximations of $p(y|x)$ for natural images be found. In turn, better approximation of this underlying data distribution should be expected to give better generalization and robustness.

4. Dataset Construction

While larger-scale popular datasets such as ImageNet [41], Places [52], or COCO [33] might seem like the best starting point, `CIFAR10` in particular has several unique and attractive properties. First, the dataset is still of enough interest to the community that state-of-the-art image classifiers are being developed on it [11, 21]. Second, the dataset is small enough to allow us to collect *substantial* human data for the entire test set of images. Third, the low resolution of the images is useful for producing variation in human responses. Human error rates for high resolution images with non-overlapping object categories are sufficiently low that it is hard to get a meaningful signal from a relatively small number of responses. Finally, `CIFAR10` contains a number of examples that are close to the category boundaries, in contrast with other datasets that are more carefully curated such that each image is selected to be a good example of the category. Our final `CIFAR10H` behavioral dataset consists of $511{,}400$ human categorization decisions over the $10{,}000$-image testing subset of `CIFAR10` (approx. 50 judgments per image).

4.1. Image Stimuli

We collected human judgments for all $10{,}000$ 32×32 color images in the *testing* subset of `CIFAR10`. This contains $1{,}000$ images for each of the following 10 categories: *airplane*, *automobile*, *bird*, *cat*, *deer*, *dog*, *frog*, *horse*, *ship*, and *truck*. This allows us to evalulate models pretrained on the `CIFAR10` training set using the same testing images, but in terms of a different distribution over labels, detailed in the next section.

4.2. Human Judgments

We collected 511,400 human classifications over our stimulus set via Amazon Mechanical Turk [5]—to our knowledge, the largest of its kind reported in a single study to date. In the task, participants were asked to categorize each image by clicking one of the 10 labels surrounding it as quickly and accurately as possible (but with no time limit). Label positions were shuffled between candidates. After an initial training phase, each participant (2,571 total) categorized 200 images, 20 from each category. Every 20 trials, an obvious image was presented as an attention check, and participants who scored below 75% on these were removed from the final analysis (14 total). We collected 51 judgments per image on average (range: $47 - 63$). Average completion time was 15 minutes, and workers were paid \$1.50 total. Examples of distributions over categorization judgments for a selection of images is shown in Figure 1.

5. Generalization Under Distributional Shift

Our general strategy is to train a range of classifiers using our soft labels and assess their performance on held-out validation sets and a number of generalization datasets with increasing distributional shift. We expect the human information about image label uncertainty to be most useful when test datasets are increasingly out-of-distribution.

5.1. Setup

Models. We trained eight CNN architectures (VGG [42], ResNet [16], Wide ResNet [50], ResNet preact [17], ResNext [49], DenseNet [20], PyramidNet [15], and Shake-Shake [11]) to minimize the crossentropy loss between softmax outputs and the full human-label distributions for images in `CIFAR10H`. The models were trained using `PyTorch` [38], adapting the repository found in the footnote.[1] For each architecture, we train 10 models using 10-fold cross validation (using 9,000 images for training each time) and at test time average the results across the 10 runs. We use k-fold instead of a single validation set in order to obtain more stable results. We used the default hyperparameters in the repository for all models, following [39] for the sake of reproducibility, except for the learning rate. We trained each model for a maximum of 150 epochs using the Adam [24] optimizer, and performed a grid-search over base learning rates 0.2, 0.1, 0.01, and 0.001 (we found 0.1 to be optimal in all cases).

[1] `github.com/hysts/pytorch_image_classification`; model identifiers `vgg_15_BN_64`, `resnet_basic_110`, `wrn_28_10`, `resnet_preact_bottleneck_164`, `resnext_29_8x64d`, `densenet_BC_100_12`, `pyramidnet_basic_110_270`, `shake_shake_26_2x64d_SSI_cutout16` (output folder names).

Test Datasets. A key prediction from section 3 is that the uncertainty in our labels will be increasingly informative when generalizing to increasingly out-of-training-sample distributions. We test this prediction empirically by examining generalization ability to the following datasets:

`CIFAR10`: This is the standard within-dataset evaluation. Since our `CIFAR10H` soft labels are for the `CIFAR10` test set, here we use the 50,000-images of the standard `CIFAR10` training set to instead evaluate the models.

`CIFAR10.1v6,v4`: These are two 2,000-image near-sample datasets created by [39] to assess overfitting to `CIFAR10` "test" data often used for validation. The images are taken from TinyImages [46] and match the sub-class distributions in `CIFAR10`. `v6` has 200 images per class while `v4` is the original class-unbalanced version (90% overlap).

`CINIC10`: This is an out-of-sample generalization test. The `CINIC10` dataset collected by [7] contains both `CIFAR10` images and rescaled ImageNet images from equivalent classes [7]. For example, images from the *airplane, aeroplane, plane (airliner)* and *airplane, aeroplane, plane (bomber)* ImageNet classes were allocated to the *airplane* `CIFAR10` top-level class. Here we use only the 210,000 images taken from ImageNet.

`ImageNet-Far`: Finally, as stronger exemplar of distributional shift, we built ImageNet-Far. As above, we used rescaled ImageNet images, but chose classes that might not be under direct inheritance from the `CIFAR10`-synonymous classes. For example, for the `CIFAR10` label *deer*, we included the ImageNet categories *ibex*, *gazelle*, and for the `CIFAR10` label *horse* we included the ImageNet category *zebra*, which was not included in `CINIC10`.

Generalization Measures. We evaluate each model on each test set in terms of both *accuracy* and *crossentropy*. Accuracy remains a centrally important measure of classification performance for the task of out-of-sample generalization. As accuracy ignores the probability assigned to a guess, we also employ the crossentropy metric to evaluate model behavior: how confident it is in its top prediction, and whether its distribution over alternative categories is sensible. Note that this interpretation arises naturally when computing crossentropy with a one-hot vector, as only the probability mass allocated to the ground-truth choice contributes to the score. Crossentropy becomes even more informative when computed with respect to human soft labels that distribute the mass unlike one-hot vectors. In this case, the second guess of the network, which provides a sense of the most confusable classes for an image, will likely be a large secondary contributor to the loss. To provide a more readily interpretable heuristic measure of this, we introduce a new accuracy measure called *second-best accuracy* (SBA). While top-1 accuracy may largely asymptote, we expect that gains in SBA may still have a way to go.

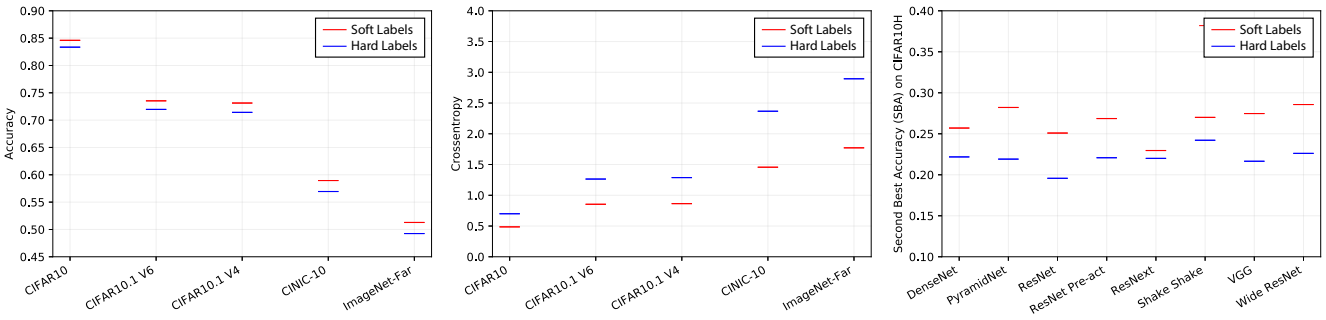

Figure 2: Generalization results. Left: accuracy against ground-truth labels, for increasingly out-of-training-sample distributions, averaged across CNNs. Accuracy was higher using human labels for every individual CNN and dataset. Center: crossentropy against ground-truth labels, averaged across CNNs. Loss was lower using human labels for every individual CNN and dataset. Right: Second best accuracy (SBA) for all models using `CIFAR10H` held out set, averaged across folds.

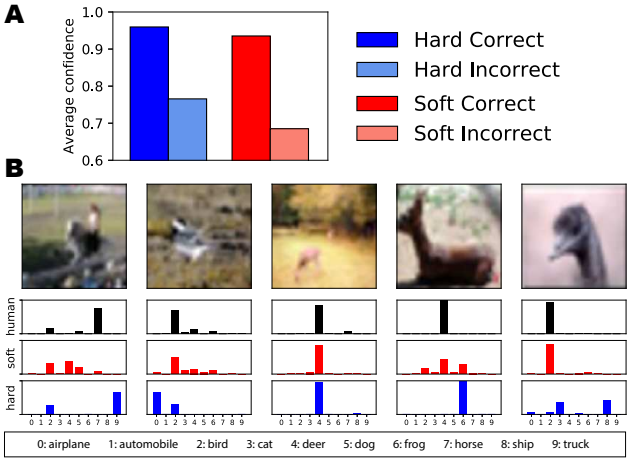

Figure 3: (A) Mean confidence for correctly/incorrectly classified examples after hard/soft label training. Soft-label models are far less confident when incorrect than hard-label controls, and only slightly less confident when correct. (B) Soft label training yields predictions that distribute probability mass more like people, with the same top choice.

5.2. Human Labels Improve Generalization

We train each CNN described above on both one-hot labels (default, control) and on `CIFAR10H` soft human labels (ours), and evaluate on each of the proposed test sets with increasingly out-of-sample distributions.

Our first finding is that when we train CNNs on `CIFAR10H` soft labels, their accuracy improves on all generalization datasets compared to our control (Figure 2, left). This pattern was replicated across individual cross-validation folds for every individual model (not shown). A key feature of this boost in generalization is that it increases as test datasets become increasingly out-of-training-distribution (horizontal axis, left to right). For example, while using human soft labels gives us only a 1% improvement (from 83.5% to 84.5%) when evaluated on `CIFAR10`, the same models when evaluated on ImageNet-Far achieved an accuracy gain of 2% on average (from 49.4% to 51.4%).

This pattern is even more evident when we consider the crossentropy metric (Figure 2, center). For example, while using human soft labels gives us a 29% reduction in crossentropy (from 0.7 to 0.5) when evaluated on `CIFAR10`, the same models when evaluated on ImageNet-Far achieve a reduction of 38% on average (from 2.9 to 1.8). These results imply that models trained on our soft labels show better confidence in their correct choices, and allocate more probability to the ground-truth during errors.

Finally, CNNs trained on our soft labels consistently show significant boosts in SBA compared to controls, performing on average 5% better (Figure 2, right). This shows improvement in generalization in a broader sense: the distribution of the most likely two categories has important consequences for the graceful degradation in generalization we hope a good model provides, as well as for the nature of guesses made by a classification model when it is wrong.

Figure 3 provides an additional picture of model behavior on our validation folds beyond overall generalization performance. Encouragingly, we find that soft-label-trained models are significantly less confident when incorrect than hard-label-trained controls, but only marginally less confident when correct (Figure 3a), and more generally provide a better fit to patterns of human uncertainty (Figure 3b).

6. Alternative Soft Label Methods

Above, we show out-of-sample classification benefits arise from training on our human labels. One natural question that arises is whether this improvement is the result of simply training with soft labels (*i.e.,* allowing the model to distribute the probability mass over more than one class), or due to the fact that this distribution explicitly mimics human uncertainty. Here we show the answer is the latter.

ResNet [16]	c10H	c10	v4	v6
Trained CIFAR10	0.82	0.25	0.84	0.82
FT CIFAR10	0.57	0.19	0.60	0.58
FT CIFAR10 with *mixup* [51]	0.36	**0.18**	**0.48**	0.46
FT CIFAR10H category soft targets	0.42	0.21	0.53	0.51
FT CIFAR10H soft targets (ours)	**0.35**	0.19	0.50	0.49
FT CIFAR10H sampled hard targets (ours)	**0.35**	0.19	**0.48**	**0.46**
ResNet preact [17]	c10H	c10	v4	v6
Trained CIFAR10	0.75	0.20	0.69	0.66
FT CIFAR10	0.65	0.19	0.61	0.59
FT CIFAR10 with *mixup* [51]	0.40	**0.18**	0.45	0.43
FT CIFAR10H category soft targets	0.44	0.23	0.47	0.46
FT CIFAR10H soft targets (ours)	0.35	0.21	0.49	0.48
FT CIFAR10H sampled hard targets (ours)	**0.34**	0.19	**0.42**	**0.41**
VGG [42]	c10H	c10	v4	v6
Trained CIFAR10	0.71	0.26	0.79	0.76
FT CIFAR10	0.54	**0.20**	0.62	0.59
FT CIFAR10 with *mixup* [51]	0.47	**0.20**	0.56	0.53
FT CIFAR10H category soft targets	0.42	0.22	0.51	0.49
FT CIFAR10H soft targets (ours)	**0.34**	0.21	**0.49**	0.48
FT CIFAR10H sampled hard targets (ours)	0.35	0.21	**0.49**	**0.47**
DenseNet [20]	c10H	c10	v4	v6
Trained CIFAR10	0.61	0.15	0.54	0.54
FT CIFAR10	0.59	0.14	0.51	0.50
FT CIFAR10 with *mixup* [51]	0.36	**0.13**	0.43	0.42
FT CIFAR10H category soft targets	0.39	0.18	0.42	0.42
FT CIFAR10H soft targets (ours)	0.32	0.17	**0.40**	0.40
FT CIFAR10H sampled hard targets (ours)	**0.31**	0.16	**0.40**	**0.39**
PyramidNet [15]	c10H	c10	v4	v6
Trained CIFAR10	0.54	0.12	0.42	0.42
FT CIFAR10	0.51	**0.11**	0.38	0.38
FT CIFAR10 with *mixup* [51]	0.49	**0.11**	0.40	0.40
FT CIFAR10H category soft targets	0.36	0.14	**0.32**	**0.32**
FT CIFAR10H soft targets (ours)	**0.28**	0.13	0.35	0.34
FT CIFAR10H sampled hard targets (ours)	**0.28**	0.12	**0.32**	**0.32**
ResNext [49]	c10H	c10	v4	v6
Trained CIFAR10	0.47	**0.10**	0.37	0.36
FT CIFAR10	0.46	**0.10**	0.35	0.34
FT CIFAR10 with *mixup* [51]	0.47	**0.10**	0.37	0.36
FT CIFAR10H category soft targets	0.37	0.17	0.37	0.36
FT CIFAR10H soft targets (ours)	0.29	0.13	**0.34**	**0.33**
FT CIFAR10H sampled hard targets (ours)	**0.28**	0.13	**0.34**	**0.33**
Wide ResNet [50]	c10H	c10	v4	v6
Trained CIFAR10	0.46	0.14	0.40	0.39
FT CIFAR10	0.42	**0.12**	0.37	0.36
FT CIFAR10 with *mixup* [51]	0.40	**0.12**	0.37	0.36
FT CIFAR10H category soft targets	0.36	0.15	0.33	0.33
FT CIFAR10H soft targets (ours)	**0.27**	0.13	0.32	0.31
FT CIFAR10H sampled hard targets (ours)	0.28	0.13	**0.31**	**0.30**
Shake-Shake [11]	c10H	c10	v4	v6
Trained CIFAR10	0.60	0.09	0.34	0.33
FT CIFAR10	0.51	**0.07**	0.28	0.27
FT CIFAR10 with *mixup* [51]	0.63	0.08	0.34	0.33
FT CIFAR10H category soft targets	0.33	0.12	0.28	0.28
FT CIFAR10H soft targets (ours)	**0.26**	0.10	**0.27**	**0.26**
FT CIFAR10H sampled hard targets (ours)	0.27	0.10	**0.27**	0.27

Table 1: Crossentropy for each holdout set (columns from left to right: holdout human soft labels (c10H), holdout ground truth labels (c10), the entire CIFAR10.1v4 dataset, and the entire CIFAR10.1v6 dataset. Crossentropy for our human labels decreases substantially after fine-tuning (FT), especially when using human targets. Fine-tuning on human targets also produces the best generalization in terms crossentropy on CIFAR10.1.

6.1. Setup

Training. We set out to demonstrate that training with human labels provides benefits even over competitive baselines. We use the same CNN architectures and setup as in Section 5.1 with one notable exception: we pre-train the networks before incorporating the soft labels (this allows us to achieve the best possible fit to humans). To do so, we train using the standard CIFAR10 training protocol using 50,000 images and the optimal hyperparameters in the repository, either largely replicating or surpassing the original accuracies proposed in the papers for each architecture. We then fine-tune each pretrained model using either hard-label controls or our human soft labels on the CIFAR10 test set. This fine-tuning phase mirrors the training phrase from Section 5.1: we used 10-folds, trained for 150 epochs, and searched over learning rates 0.1, 0.01, and 0.001.

Evaluation. We evaluate the results on the holdout folds of CIFAR10H with both human soft labels and ground truth hard labels, as well as on the ground truth hard labels of both the CIFAR10.1v4 and CIFAR10.1v6 datasets. We also shift our attention to evaluating crossentropy rather than accuracy. With CIFAR10 pretraining, the accuracy of all models is high, but this gives no indication of the level of confidence or the "reasonableness" of errors. Crossentropy, on the other hand, does exactly that: measures the level of confidence when evaluated on hard labels and the "reasonableness" of errors when evaluated on human soft labels.

6.2. Methods

To test for simpler and potentially equally effective alternatives to approximating the uncertainty in human judgments, we include a number of competitive baselines below.

Ground Truth Control. The first baseline we consider is a "control" fine-tuning condition where we use identical image data splits, but fine-tune using the ground-truth hard labels. This is expected to improve upon the pretrained model as it utilizes the additional $9{,}000$ images previously unseen.

Class-level Penalty. One much simpler alternative to image-level human soft labels is class-level soft labels. That is, instead of specifying how much each image resembles each category, we could simply specify which classes are more confusable on average using a class-level penalty. However, while we know, for example, that dogs and cats are likely more confusable on average than dogs and cars, it's not clear what the optimal class-level penalties should be. Since exhaustively searching for competitive inter-class penalties is inefficient, we propose to generate gold-standard penalties by summing and re-normalizing our human probabilities within each class (*i.e.,* resulting in exactly 10 unique soft-label vectors). This also allows us to determine if image-level information in our human soft labels

is actually being utilized as opposed to class-level statistics across image exemplars. In this baseline, fine-tuning simply uses these greatly compressed soft vectors as targets.

Knowledge Distillation. As discussed in Section 2, softmax probabilities of a trained neural network can be used as soft labels because they contain information inferred by the network about the similarity between categories and among images. The pretrained networks from this section provide such probabilities and so provide a corresponding baseline. However, we can infer from the results in Section 5.2 that hard-label-trained CNNs infer class probabilities that do not approximate those of humans, because incorporating explicit supervision to humans provides different results in terms of generalization. So, to provide a stronger baseline in this respect, we include an ensemble of the predictions from all eight models (*i.e.,* providing soft predictions due to uncertainty from variation across models).

***mixup*.** *mixup* is a technique for soft label generation that improves the generalization of natural image classification models trained on `CIFAR10` among others [51]—see Section 2. As such, it provides an interesting and competitive baseline with which to compare training with human soft labels. Concretely, *mixup* generates soft labels by taking convex combinations of pairs of examples, encouraging linear behavior between them. These combinations constitute virtual training examples ($\bar{x}$, $\bar{y}$) that are sampled from a vicinial distribution, and take on the form

$$\bar{x} = \lambda x_i + (1 - \lambda) x_j$$
$$\bar{y} = \lambda y_i + (1 - \lambda) y_j,$$

where (x_i, x_j) are examples from the dataset, and (y_i, y_j) are their labels. The strength of the interpolation $\lambda \in [0, 1]$ is sampled according to Beta(α, α), where α is a hyperparameter. For our *mixup* baseline, we apply this procedure to the ground truth labels corresponding to each of the same 10 splits used above. For each architecture, we searched for the best value of α from 0.1 to 1.0 in increments of 0.1.

Soft Labels Versus Sampling. Finally, we run one additional experiment beyond the soft label baselines above. Results from Section 5 suggest that human soft labels are useful, but how should we best incorporate them into training? In Section 3, we justified using human probabilities as targets to minimize the expected loss. However, another valid option is to sample from $p_{\text{hum}}(y|x)$, *i.e.,* sample one-hot labels from categorical distribution parameterized by the human probabilities conditioned on each image. If we sample a new label each time the image is presented to the network for a new gradient update, the label uncertainty will still be incorporated, but there will be additional variation in the gradients that could act as further regularization. To test for any such advantages of label sampling, we fine-tuned a second corresponding set of models using this method, sampling a new label for each image on each epoch.

6.3. Human Soft Labels Beat Alternatives

Results are summarized for each architecture and method in Table 1. The first column is our primary measure of fit to humans; the last two assess further generalization.

Note that for pretrained models (first row of each subtable) crossentropy to ground truth labels is always lower than human soft labels, verifying what we expected: human soft labels provide additional information that is not inferred via training with ground truth. This is a first test that the information (informative probabilities) usually inferred by these networks using hard labels (*i.e.,* knowledge distillation) does not agree with humans. We further tested an ensemble of all eight networks in the top rows (*i.e.,* with no fine-tuning on human soft labels), and while this model is more like humans than any individual hard-label-trained model (crossentropy is 0.41), it is still not a substitute for human supervision. The benefit from our labels also appears to manifest during generalization, as in the last two columns (*i.e.,* `v4` and `v6` holdout sets) they show higher crossentropy than alternative approaches. Next, looking at the same top rows, note that there is little correspondence between recency of the architecture and fit to humans. In fact, Shake-Shake is the state-of-the-art of the eight yet is not one of the top three models in terms of fit to humans.

In the remaining rows of each sub-table, we can see an increase in fit to humans using our various fine-tuning schemes. This is expected in all cases given that all of these models are ultimately given more data than pretrained models. However, not all fine-tuning methods are equally effective. Importantly, fit to humans (second column) is best when either using our image-level soft labels or sampling hard labels using them (bottom two rows). Interestingly, category soft labels (4th rows) were also effective, but to a lesser degree. *mixup* was more effective than using ground truth labels alone, but less effective than any methods using human information. Lastly, we note that, while omitted for brevity, we found no loss in accuracy when using human labels in any of the conditions that utilized them.

7. Robustness to Adversarial Attacks

Because our soft labels contain information about the similarity structure of images that is relevant to the structure of perceptual boundaries, we might expect that representations learned in service of predicting them would be more robust to adversarial attacks, particularly in cases where similar categories make for good attack targets. Moreover, subsequent explorations of knowledge distillation [19, 37] have demonstrated that such practices can support adversarial robustness. If human judgments of perceptual similar-

	Accuracy		Crossentropy	
Architecture	C10	C10H	C10	C10H
VGG	7%	**8%**	7.9	**4.1**
DenseNet	17%	**19%**	6.9	**3.0**
PyramidNet	**22%**	19%	5.7	**2.8**
ResNet	15%	**23%**	6.1	**3.1**
ResNext	**25%**	24%	4.2	**2.7**
Wide ResNet	24%	**35%**	4.1	**2.2**
ResNet preact	17%	**29%**	6.3	**2.6**
Shake-Shake	**39%**	**39%**	4.0	**2.1**

Table 2: Accuracy and crossentropy after FGSM attacks on the CIFAR10-tuned (baseline) and CIFAR10H-tuned networks. Using human labels always results in lower (better) crossentropy, and in the majority of cases, higher accuracy.

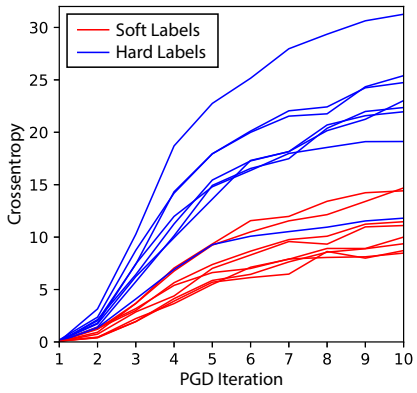

Figure 4: Crossentropy as a function of PGD iteration. Successive iterations increase crossentropy as expected, but more slowly after soft-label fine-tuning.

ity are superior to those inferred by CNNs—in the form of $p(y|x)$—we would expect distillation of human knowledge into a CNN would at the very least also increase robustness.

Setup. We use the same pretrained and fine-tuned (hard versus soft) models from Section 6. To measure robustness after each training scheme, we evaluate both accuracy and crossentropy (the latter again being a more sensitive measure of both confidence and entropy) against the hard class labels. As attack methods, we evaluate two additive noise attacks: the Fast Gradient Sign Method (FGSM) [29], and Projected Gradient Descent (PGD) [30], using the mister_ed toolkit[2] for PyTorch. For both methods, we explored ℓ_∞ bounds of 4 to 8 in increments of 1. Since we found no significant differences in the results, we report all attack results using a constant ℓ_∞ bound of 4 for brevity.

Human Soft Labels Confer Robustness. FGSM results are reported in Table 2, averaged over all 10,000 images in the CIFAR10 test set. In all cases, crossentropy (which attack methods seek to maximize) is much lower (roughly half) after attacking the human-tuned network compared to fine-tuning with original one-hot labels. For five out of eight architectures, accuracy also improves when using human soft targets. The two largest differences (Wide Resnet and ResNet preact) favor the human labels as well. Note that no explicit (defensive) training was required to obtain these improvements beyond previous training with human labels.

Without active defensive training, PGD is expected to drive accuracy to 0% given enough iterations. To explore the intrinsic defenses of our two label-training conditions to PGD attacks, we plot the increase in loss for each architecture and label-training scheme in Figure 4. While accuracy was driven to 0% for each network when trained on standard labels, and 1% for each network with human labels, loss for the former is driven up much more rapidly, whereas the latter asymptotes quickly. Put simply, a much higher degree of effort is required to successfully attack networks that behave more like humans.

[2] github.com/revbucket/mister_ed/

8. Discussion

In this work, we have demonstrated that incorporating information about human category uncertainty at the image-level can help protect against the perils of distributional shift and adversarial attacks. Notably, common classification benchmarks often do not naturally provide such protections on their own [45]. Further, besides explicitly incorporating this information, it gives a way of measuring whether our learning algorithms are inferring good similarity structure (beyond just top-1 performance). If we can begin to find good learning procedures that derive such information, we can obtain human-like robustness in our models without the need of explicit human supervision. However, developing such a robust models will take significant time and research—our dataset provides a first step (an initial gold standard with respect to a popular benchmark) in measuring this progress, even when not used for training.

Although our data collection method does not immediately seem to scale to larger training sets, it's certainly possible to collect informative label distributions at a cost comparable to what we often spend on compute to find better top-1-fitting architectures. Interestingly, we found that the bulk of human uncertainty is concentrated in approximately 30% of the images in our dataset, meaning straightforward and much more efficient methods for mining only these more informative labels can be employed. In any case, we see the main contribution of such datasets as testing environments for algorithms intended for much larger datasets.

Acknowledgements. This work was supported by grant number 1718550 from the National Science Foundation.

References

[1] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59, 2018.

[2] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

[3] Jonathan Bragg, Mausam, and Daniel S. Weld. Crowdsourcing multi-label classification for taxonomy creation. In *Conference on Human Computation and Crowdsourcing (HCOMP)*, 2013.

[4] Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie. Visual recognition with humans in the loop. In *European Conference on Computer Vision (ECCV)*, 2010.

[5] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5, 2011.

[6] L. B. Chilton, G. Little, D. Edge, D. S. Weld, and J. A. Landay. Cascade: Crowdsourcing taxonomy creation. In *Conference on Human Factors in Computing Systems (CHI)*, 2013.

[7] Luke Nicholas Darlow, Elliot J. Crowley, Antreas Antoniou, and Amos J. Storkey. CINIC-10 is not imagenet or CIFAR-10. *arXiv preprint arXiv:1810.03505*, 2018.

[8] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *European Conference on Computer Vision (ECCV)*, 2014.

[9] R. Fergus, H. Bernal, Y. Weiss, and A. Torralba. Semantic label sharing for learning with many categories. In *European Conference on Computer Vision (ECCV)*, 2010.

[10] A. Frome, G.S. Corrado, J. Shlens, S. Bengio, J. Dean, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

[11] Xavier Gastaldi. Shake-shake regularization. *arXiv preprint arXiv:1705.07485*, 2017.

[12] Aritra Ghosh, Himanshu Kumar, and PS Sastry. Robust loss functions under label noise for deep neural networks. In *Conference on Artificial Intelligence (AAAI)*, 2017.

[13] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[14] G. Griffin and P. Perona. Learning and using taxonomies for fast visual categorization. In *Computer Vision and Pattern Recognition (CVPR)*, 2008.

[15] Dongyoon Han, Jiwhan Kim, and Junmo Kim. Deep pyramidal residual networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)*, 2016.

[18] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory F. Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

[19] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[20] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[21] Yanping Huang, Yonglong Cheng, Dehao Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, and Zhifeng Chen. Gpipe: Efficient training of giant neural networks using pipeline parallelism. *arXiv preprint arXiv:1811.06965*, 2018.

[22] S.J. Hwang, F. Sha, and K. Grauman. Sharing features between objects and their attributes. In *Computer Vision and Pattern Recognition (CVPR)*, 2011.

[23] Y. Jia, J.T. Abbott, J. Austerweil, T. Griffiths, and T. Darrell. Visual concept learning: Combining machine vision and bayesian generalization on concept hierarchies. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

[24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2014.

[25] Adriana Kovashka, Olga Russakovsky, Li Fei-Fei, and Kristen Grauman. Crowdsourcing in Computer Vision. *Foundation and Trends in Computer Graphics and Vision*, 10(3):177–243, 2016.

[26] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1), May 2017.

[27] Ranjay A. Krishna, Kenji Hata, Stephanie Chen, Joshua Kravitz, David A. Shamma, Li Fei-Fei, and Michael S. Bernstein. Embracing error to enable rapid crowdsourcing. In *Conference on Human Factors in Computing Systems (CHI)*, 2016.

[28] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[29] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations (ICLR)*, 2016.

[30] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations (ICLR)*, 2016.

[31] George Lakoff. *Women, fire, and dangerous things*. University of Chicago press, 2008.

[32] C.H. Lampert. Maximum margin multi-label structured prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2011.

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

[34] M. Marszalek and C. Schmid. Semantic hierarchies for visual object recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2007.

[35] An Thanh Nguyen, Byron C Wallace, and Matthew Lease. Combining crowd and expert labels using decision theoretic active learning. In *Conference on Human Computation and Crowdsourcing (HCOMP)*, 2015.

[36] Quang Nguyen, Hamed Valizadegan, and Milos Hauskrecht. Learning classification models with soft-label information. *Journal of the American Medical Informatics Association*, 21(3):501–508, 2014.

[37] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE, 2016.

[38] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NeurIPS Autodiff Workshop*, 2017.

[39] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.

[40] David Rolnick, Andreas Veit, Serge J. Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.

[41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2014.

[43] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[44] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2013.

[45] Antonio Torralba, Alexei A Efros, et al. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR)*, 2011.

[46] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970, 2008.

[47] Antonio Torralba, Kevin P Murphy, and William T Freeman. Sharing visual features for multiclass and multiview object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):854–869, 2007.

[48] Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2017.

[49] Saining Xie, Ross B Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[50] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference (BMVC)*, 2016.

[51] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2017.

[52] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[53] A. Zweig and D. Weinshall. Exploiting object hierarchy: combining models from different category levels. In *International Conference on Computer Vision (ICCV)*, 2007.