HUMANS AS A DIRECT MEASURE OF TRUE LABEL UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

Waiting on results and the rest of the manuscript...

1 INTRODUCTION

The impressive performance of convolutional neural networks (CNNs) in solving natural-image classification has been one of the primary drivers of the deep learning revolution, providing much of its initial stimulus (Krizhevsky et al., 2012), development through model competition on benchmark datasets (Krizhevsky, 2009), and interest for technological application (Lakhani & Sundaram, 2017). On these benchmarks, state-of-the-art models have been said to begin to equal or even surpass human performance, often measured in terms of the accuracy of a model's top category choice for a test set of held-out images.

As performance gains have begun to asymptote at near-perfect levels, there has been increasing focus on their out-of-training-set performance—in particular, their ability to generalize to related stimuli (Recht et al., 2018), and their robustness to adversarial examples (Kurakin et al., 2016). On these tasks, by contrast, CNNs tend to perform rather poorly, whereas humans continue to perform well. A key recent observation has been that under current training paradigms, CNNs could memorize training examples rather than learning to generalize from them—a phenomenon that would lead to the good in-sample and poor out-of-sample performance described above (Zhang et al., 2017).

One strategy to address these shortcomings has been to augment datasets by sampling extra training inputs using domain-specific alterations—such as slight image rotations and horizontal reflections—or distributional assumptions over the perceptual space surrounding empirical examples (Schott et al., 2018; Simonyan & Zisserman, 2014). However, there also is another rich source of information that can also be leveraged for this purpose: the uncertainty over labels inherent in the underlying data distribution. Instead of assigning the modal—or "ground truth"—category to an image, we can sample image-labels pairs from a distribution over category labels for a particular image, $p(c|x)$. This enables us to harness relevant structure in the label data, for example, cross-category confusions, and learn perceptual boundaries that can confer the desiderata discussed above. Our motivation for this extension is that there is often a lack of absolute human consensus on the category of an object in the world, and the nature of categorization decisions away from the ground truth often convey important information about the structure of the visual world in general.

In this paper, we show that sets of human image classifications can be used to infer $p(c|x)$, and is an ecological and efficient strategy to capture distributional assumptions during training that confer better generalization and robustness. We do so by presenting a novel image dataset we call `CIFAR10H`, which comprises over 500k human classifications over the test set of the `CIFAR10` natural image dataset (Krizhevsky, 2009). This allows us to assess the generalization of pretrained CNN classification models on labels sampled from distributions over images, and to show that when such models are tuned to better predict them, their generalization abilities improve.

2 EMPIRICAL RISK MINIMIZATION BEYOND THE MODE

In statistical learning, our goal is to learn the best model of an underlying data distribution over a set of random variables, for example, features X and labels Y. In general, we are given a family of models f, and are tasked to find a member of that family, indexed by parameters θ, that minimizes the expected loss:

$$min_\theta \int \mathcal{L}(f_\theta, x, y)\, p(x, y)\, dx dy, \tag{1}$$

where $\mathcal{L}$ is a loss function that penalizes deviations of model predictions from the data distribution, and $p(x, y)$ is the data distribution itself. Since in general, we do not know $p(x, y)$, we approximate it by the empirical loss over a set of n samples $(x_1, y_1), \cdots, (x_n, y_n)$:

$$\int \mathcal{L}(f_\theta, x, y)\, p(x, y)\, dx\, dy \approx \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f_\theta, x_i, y_i). \tag{2}$$

In supervised learning, we use a training subset of samples to learn parameters θ that minimize the loss between model outputs $f_\theta(x)$ for inputs x, and the "true" outputs y associated with them. For natural image classification, our family of functions f are CNNs parameterized by θ, our inputs x are vectors of pixel intensities, and our targets y are category labels corresponding to each image. The probabilistic interpretation of these targets is that of samples drawn from the true $p(y|x)$. For a given image x_i generated from its true category c_t in a set of k categories, $p(y_i = c_t|x_i)$ could take on any probability.

The standard practice for computing this loss has been to use "ground truth" labels (in the form of "one-hot" vectors) provided in common benchmark datasets, for example, `ILSVRC12` and `CIFAR10`, to train and evaluate models. This corresponds to the assumption that training labels are always sampled from the true category of their respective images; in other words, that $p(y_i = c_t|x_i) = 1$. Here, the "true" image categories are decided through human consensus (taking the mode of the distribution over images), or by the database creators. However, this approximation introduces a bias into the learning paradigm that has important distributional implications. Instead of a particular image—-and stimulus vector—-being associated with a probability mass function over labels, all probability mass is reallocated to the modal category. This forces the network to learn that all instances of a category are equally and totally likely, and destroys information about how likely it is to come from others. For example, if an image is known to have been drawn from category c_t, we assume $p(x_i|y_i = c_t)$ is infinitely greater than $p(x_i|y_i = c_f)$ for all $c_f \neq c_t$. This removes all supervisory information about x_i's similarity to other classes, sampling only from the mode, and potentially placing an artificial bound on the expected risk.

Under what circumstances is this a reasonable assumption or approximation? In many problems, $p(x_i|y_i = c_t)$ is clearly not equal for all examples from c_t (for example, some images of dogs are more likely than others). Without this information, our classifier will not know that the penalty for mistaking a dog for a cat should be less than mistaking a dog for a car. More importantly, in domains where category exemplar feature overlap is common (for example, dogs and cats share many joint features sets), we are forcing our classifier to solve the wrong problem. In order to satisfy the constraint of wholly non-confusable category exemplars, our networks may resort to memorization (which notably satisfies this condition given any dataset with non-identical exemplars across categories).

In this paper, we show that information about the uncertainty over image categories is indeed useful, and can be used to improve inferences as to the true underlying data distribution over images *and* labels. If we expect human image labels to reflect the natural distribution over categories given an image, we can use these samples to infer a posterior over true category values, $p(y|\tilde{y})$, that improves our approximation of the expected loss as follows:

$$\int \mathcal{L}(f_\theta, x, y)\, p(x, y)\, dx\, dy \approx \sum_{i=1}^{n} \int \mathcal{L}(f_\theta, x_i, y_i)\, p(y|\tilde{y})\, dy_i. \tag{3}$$

Our main contribution is to show that human categorizations of natural images follow a distribution over potential labels that can be used for this purpose, and that when we do so network performance generalizes better to unseen natural images. This approach naturally complements those seeking to improve generalization by using distributional assumptions around $p(x|y)$ to augment training sets (Zhang et al., 2017).

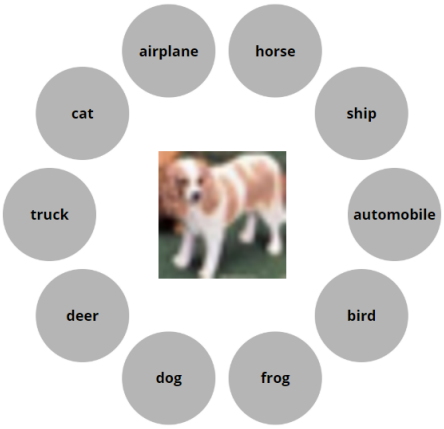

Figure 1: JP: Experimental task figure. I'm thinking we can put the focus cross on the left, and a histogram example of the repeated dog image on the right. Can polish later...

3 Dataset Construction

The human decisions and accompanying CNN representations we investigate are based on the `CIFAR10` dataset, which comprises $60{,}000$ 32×32 color images from 10 categories of natural objects (Krizhevsky, 2009). Human judgments were collected for all $10{,}000$ images in the *testing* subset, which contains $1{,}000$ images for each of the following ten categories: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck.

3.1 Human Experiments

Our `CIFAR10H` behavioral dataset consists of $511{,}400$ human categorization decisions made over this stimulus set collected via Amazon Mechanical Turk (Buhrmester et al., 2011)—to our knowledge, the largest reported in a single study to date. Participants saw an image, presented centrally, and were asked to categorize it by pressing one of the ten labels surrounding the image as quickly and accurately as possible. Label positions were shuffled between candidates. There was an initial training phase, during which candidates had to score at least 75% accuracy, split into 3 blocks of 20 images taken from the `CIFAR10` training set (6 per category, total). If a candidate failed any block they were asked to redo it until passing the threshold accuracy. For the main experiment, each participant ($2{,}571$ total) categorized 200 images, 20 from each category. Every 20 trials there was an attention check image—a carefully selected unambiguous member of a particular category. Participants who scored below 75% on these checks were removed from the final analysis (14 participants failed checks).

The mean number of judgments per image was 51 (range: $47 - 63$). The mean accuracy per subject was 95% (range: $71\% - 100\%$). The mean accuracy per image was 95% (range: $0\% - 100\%$). Average completion time was 15 minutes, and workers were paid $\$1.50$ total. See Figure 2 for a schematic of the judgments.

References

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.

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

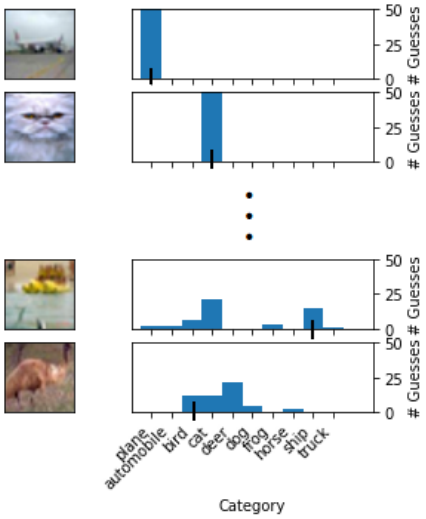

Figure 2: RB: Experimental task figure. We should indicate ground truth a little better. Can polish later...

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

Paras Lakhani and Baskaran Sundaram. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284(2): 574–582, 2017.

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

Lukas Schott, Jonas Rauber, Wieland Brendel, and Matthias Bethge. Robust perception through analysis by synthesis. *CoRR*, abs/1805.09190, 2018. URL `http://arxiv.org/abs/1805.09190`.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.