Human uncertainty improves object classification

Joshua C. Peterson* (peterson.c.joshua@gmail.com)

Ruairidh M. Battleday* (battleday@princeton.edu)

Thomas L. Griffiths (tomg@princeton.edu)

Department of Computer Science
Princeton University Princeton, NJ 08540 USA

Abstract

Despite the continued improvement in deep network classifiers, humans remain the enduring gold standard for strong generalization and robustness. In this work, we show that incorporating more human-like perceptual uncertainty into classification models can help narrow this gap. In particular, we show that training state-of-the-art convolutional neural networks with human-derived distributions over labels, as opposed to ground-truth labels, improves their generalization to out-of-sample datasets and robustness to adversarial attacks. These findings suggest that more accurately capturing uncertainty over image labels is critical to forming a robust visual model of the world. To facilitate further advancements of this kind, we propose our human-derived "soft" label distributions for the `CIFAR10` test set, which we call `CIFAR10H`, as a new benchmark.

Keywords: object classification; convolutional neural networks; generalization; statistical learning.

Introduction and Methods

A striking feature of human categorization abilities is graceful degradation: generalization to increasingly out-of-training-sample images or perceptual distortions remains satisfyingly robust. On the other hand, state-of-the-art natural image classifiers, such as convolutional neural networks (CNNs), do not exhibit the same pattern: they generalize poorly (Recht, Roelofs, Schmidt, & Shankar, 2018), succumb to perceptually insignificant adversarial attacks (Kurakin, Goodfellow, & Bengio, 2016), and exhibit different patterns of uncertainty than humans (see Figure 1).

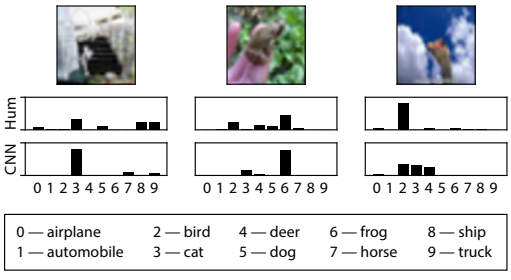

Figure 1: Perceptual uncertainty from humans and CNNs on natural images.

In this work, we show that training CNNs with human-derived label distributions (guess proportions for each class) helps to close this gap. In particular, we train CNNs using a novel dataset, which we call `CIFAR10H`. This comprises full label distributions for all $10{,}000$ images from the `CIFAR10` test set. We then compare the performance of these networks to typically-trained controls on a number of increasingly out-of-sample generalization datasets. We also compare their robustness to adversarial attacks via Projected Gradient Descent (PGD) (Kurakin et al., 2016).

Results and Discussion

We find that replacing hard labels (*i.e.,* "ground truth" labels) with human soft labels (*i.e.,* full label distributions) results in better performance, in terms of loss and accuracy, on all of our increasingly distant generalization datasets (see Figure 2). Notably, this benefit increases as test sets move increasingly out-of-distribution. We also find that training on full label distributions improves robustness to adversarial attacks.

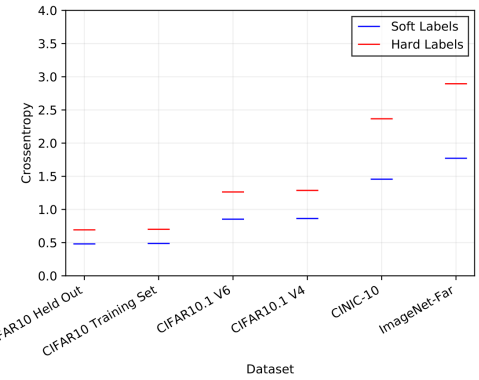

Figure 2: Average CNN performance (crossentropy loss) on increasingly distant generalization datasets (lower is better).

References

Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2018). Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*.