On the Informativeness of Supervision Signals (Supplementary Material)

Ilia Sucholutsky[1] **Ruairidh M. Battleday**[1] **Katherine M. Collins**[2] **Raja Marjieh**[3] **Joshua C. Peterson**[1]
Pulkit Singh[1] **Umang Bhatt**[2,4] **Nori Jacoby**[5] **Adrian Weller**[2,4] **Thomas L. Griffiths**[2,1]

[1]Dept. of Computer Science, Princeton University,
[2]Dept. of Engineering, University of Cambridge,
[3]Dept. of Psychology, Princeton University,
[4]Alan Turing Institute,
[5]Max Planck Institute for Empirical Aesthetics

A PRACTICAL GUIDANCE ON HUMAN SOFT LABEL ELICITATION

Our framework provides users with a way to quantify the relative amount of information contained in each type of label so that they can optimize which labels to collect for their dataset. Specifically, the findings from our theory, simulations, and experimental results are that the relative informativeness of labels depends on three factors associated with the dataset (the number of labeled examples, the number of classes, and the latent dimensionality) and two factors associated with labels (error rate and sparsity). Out of these, most factors are not under the user's control but the primary factor that users can control (in supervised learning settings) is label sparsity. Our guidance to users is thus the following:

- Use our framework to estimate the relative informativeness of the label types you are considering collecting. The key parameter to optimize is label sparsity so compute the informativeness of soft labels with different levels of sparsity.
- Pick out promising label types and run a small pilot study collecting each label type for a small set of objects. Compute error rates and per-label costs for each label type.
- Update the relative informativeness estimates based on error rates and calculate the cost-benefit tradeoff for each label type. Pick the type with the most favorable tradeoff.

For users who want a simpler procedure, we offer the following rule-of-thumb guidance: *Generally, softer labels are preferable in smaller data regimes (e.g. one-shot and less-than-one-shot learning) while harder labels are preferable in big data regimes (i.e. many-shot learning).*

B LABEL OPTIMIZATION SIMULATIONS

See Figure 1.

C ELICITED-HUMAN VS MODEL-PREDICTED ENTROPY

We investigate the hypothesis that the labels which confer the best downstream performance may strike a natural resonance with the models they are used as supervision signals for. In Figure 2, we compare the entropy of the training labels against the entropy of the trained models' predicted distributions. In other words, we compare the probability distributions produced by each model, to the probability distributions that the model was trained on. We find a remarkable alignment between the entropy of models' predictions trained on the `CIFAR-10S` varieties. Future work could investigate the links between the inductive biases of models and the labels best suited for training specific architectures.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

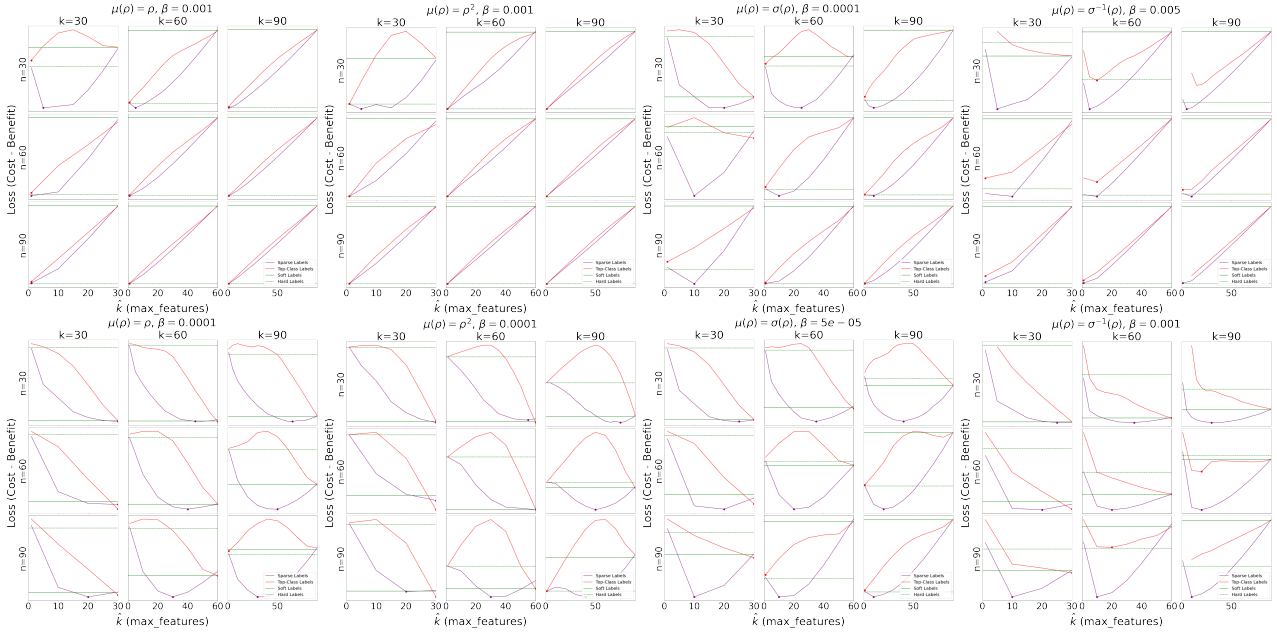

Figure 1: Loss curves for soft labels (solid green), sparse labels (purple), top-class labels (red), and hard labels (dashed green) based on subjective utility function ($u(\rho)$), cost weighting parameter (β), sparsity ($\hat{k}$), number of points (n), and number of classes (k). Global minima for sparse labels and top-class labels are marked with a point.

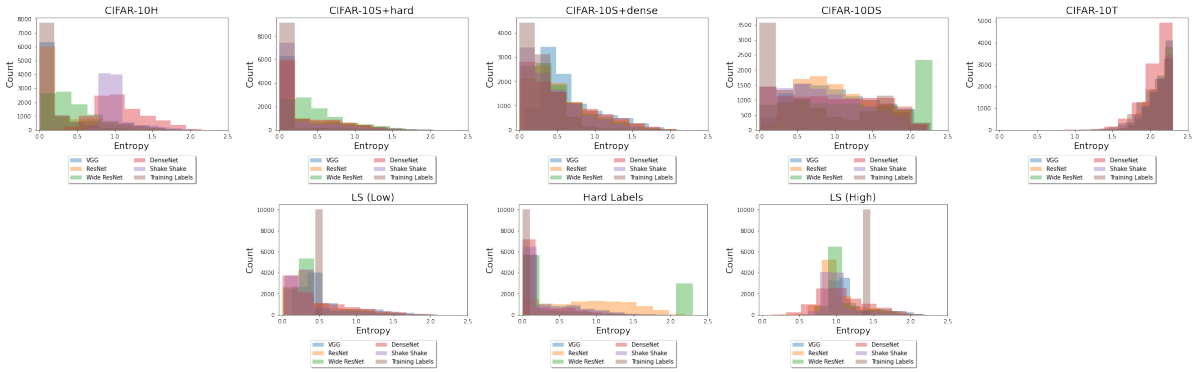

Figure 2: Comparing the entropy of the models' predictions against the entropy of the labels used to train them. The training label type is listed as the title for the respective histogram.

D ADDITIONAL DETAILS ON HUMAN SOFT LABELS

D.1 COLLECTING `CIFAR-10DS` AND SIMILARITY JUDGMENTS

Soft labels for `CIFAR-10DS`, as well as similarity judgments, were collected on Amazon Mechanical Turk (AMT). The recruitment and experimental pipelines were automated using the PsyNet framework for online experiment design Harrison et al. [2020]. Prior to participation in the studies, participants provided informed consent in accordance with an Institutional Review Board (IRB), and were paid at a rate of $12 per hour. In addition, participants were required to have successfully completed at least 2000 tasks on AMT.

To collect `CIFAR-10DS`, participants observed individual images and were given a set of sliders (10, one for each category) ranging from 0 to 1 and were asked to move the sliders in accordance with how well they thought each category matched a given image, with 0 being "not at all matching" and 1 being "completely matching". We aimed for about 10 multi-ratings per image and each participant completed 50 such multi-ratings.

As for similarity judgments, participants were presented with pairs of unlabeled images and were required to rate their

similarity on a 7-point Likert scale ranging from 0 (“completely dissimilar”) to 6 (“completely similar”). Here we aimed for 5 judgments per pair of images and each participant completely an average of 80 such judgments.

D.2 IN-FILLING CIFAR-10S LABELS

The CIFAR-10S labels collected in Collins et al. [2022] included only 1,000 of the full 10,000 CIFAR-10 test set. Note, however, that these 1,000 examples were already enriched to be those that are naturally more confusing – so it can be considered a sensible sampling of what -10S labels *may* look like more generally. However, for adequate comparison against the other label types, we needed to choose a labeling method to label the remaining 9,000. We elected two variants: 1) using hard labels, or 2) simulating CIFAR-10S labels via sparsified version of CIFAR-10DS. The former represents a real-world cost efficient scenario; we could imagine a researcher only having the budget to annotate a subset of a dataset with soft labels. The second case is designed to mimic what the labels may have been like had we elicited CIFAR-10S over the full set. Taking only the scalar values for the top two highest sliders from CIFAR-10DS offered a nice entropy- and conceptual-match (entropy of 0.69 for CIFAR-10S to 0.75 for the adjusted -10DS labels). Future work could explore automated measures to extend label conversions (e.g., learning a mapping from CIFAR-10DS to simulated CIFAR-10S labels). We note that the CIFAR-10S labels used in this work are the T2 Clamp varieties, with a redistribution factor of 10% following Collins et al..

CIFAR-10T The CIFAR-10T labels are a novel set of labels we crowdsourced, comprising over 350,000 typicality ratings for each image under the ground truth category (about 35 judgements per image). 1759 unique participants were recruited on Amazon Mechanical Turk, and presented with a sequence of 200 randomly sampled CIFAR-10 test set images, upsampled to 160x160 pixels (see Peterson et al. [2019], Battleday et al. [2020]. Participants were given the category of each image, and asked to rate how typical it was of the category on a sliding scale of “Not at all typical" to “Extremely typical". We interpret an image’s typicality as the probability of the ground truth class, and spread the remaining probability mass over the 9 remaining labels—a smoothed version of a *sparse* soft label with K=1).

D.3 ADDITIONAL SIMILARITY JUDGMENT STUDIES

We extend the GNMDS analyses in the main text by examining the similarity structure of image representations extracted from the penultimate layer of each network. For each image and network, we derive an abstract vector representation by storing the unit activations of the last layer during classification. Then, for each image we compute the pairwise cosine similarity between the representations derived from our classifiers. We correlate these to the ground truth similarity ratings, and present the results in Figure 3. The images used for these analyses are discussed below, and displayed in Figures 6 and 7.

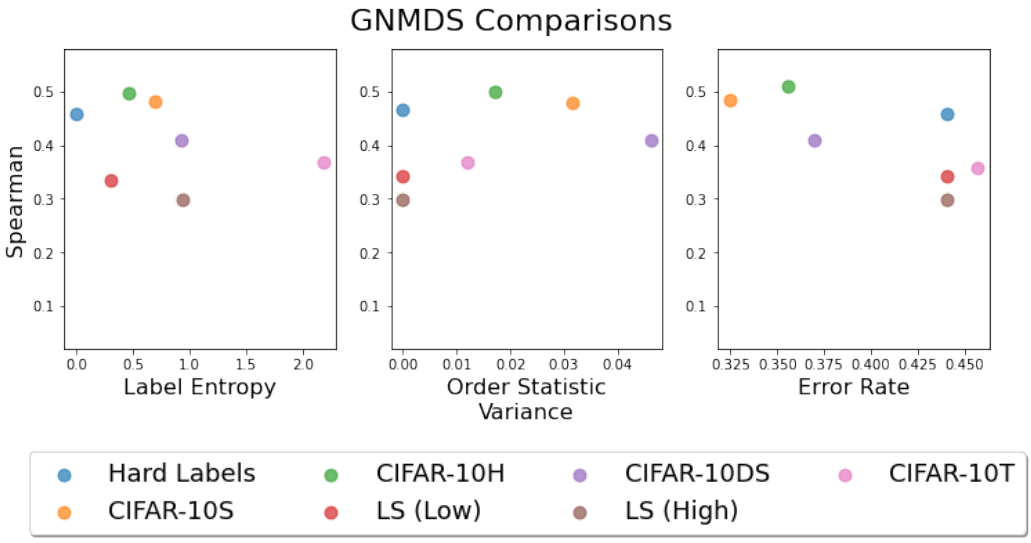

Figure 3: Correlation between ground-truth similarity judgments and the cosine similarity of image representations for different model architectures.

E ADDITIONAL COMPUTATIONAL EXPERIMENT DETAILS AND OBSERVATIONS

E.1 MODELS

We use ten fold cross-validation to partition the images of the CIFAR-10 test subset into train and validation sets for each set of soft labels. We train a number of models using stochastic gradient descent over a range of learning rates and seeds, and use the best performing seed for all subsequent analyses (Table 1). We use ten fold cross-validation to partition the images of the CIFAR-10 test subset into train and validation sets for each set of soft labels.

Table 1: Image Classifiers.

Model	Key Features	Parameters	Citation
VGG	very deep connections	14,728,266	Simonyan and Zisserman [2014]
ResNet	residual connections		He et al. [2016]
WRN	wide residual connections	36,479,194	Zagoruyko and Komodakis [2016]
DenseNet	dense connections	769,162	Huang et al. [2017]
Shake shake	shake shake regularization	11,709,514	Gastaldi [2017]

E.2 DATASETS

In order of increasing distributional shift, CIFAR-10 50K is the *training* subset of CIFAR10 (50,000 images; Krizhevsky et al. [2009], CIFAR10.1v6,v4 are two near-sample datasets constructed from the same TinyImages classes Torralba et al. [2008] as CIFAR-10 (2,000 images each; Recht et al. [2018]), our subset of CINIC10 contains rescaled images from ImageNet using the CIFAR-10 classes (210,000 images; Darlow et al. [2018]), and ImageNet-Far contains a label-coarsened version of rescaled ImageNet images such that the CIFAR classes now contain a more diverse range of examples (for example, now "deer" contains "ibex" and "gazelle"; 63,895 images; Peterson et al. [2019], Darlow et al. [2018]).

E.3 SOFTNESS, TASK ACCURACY, AND INFORMATION CONTENT

In the main text, we depicted the relationship between label softness and task performance using crossentropy (CE) as our principal metric. We focus on CE as it better captures the fidelity of the models' predictive distributions. This is particularly important when we evaluate the model on held-out soft labels; CE offers more information about model performance than just top-1 accuracy. However, we include top-1 accuracy in Figure 4 for completeness.

We also depict performance in Figure 5 as function of the information content of the labels. Here, we use the Spearman rank correlation coefficient between the CIFAR-10 GNMDS and elicited similarity judgments as a proxy for information content. Note, here, CIFAR-10S+hard and CIFAR-10S+dense have the same score, as the similarity judgments are collected over the 1000 shared original CIFAR-10S examples.

E.4 EFFECTIVE DIMENSIONALITY OF SOFT LABELS

Our theory and simulations address the number of features available for representation learning but did not discuss the nature of these features—i.e., whether they are essential or superfluous. Estimating the effective dimensionality of a dataset is tied to the nature of the computational task required. For classification, there is a range of methods for estimating this (e.g., [Vapnik and Chervonenkis, 1971, Jha et al., 2023]).

E.5 VARYING LABEL SOFTNESS

We further investigate how the amount of softness we elicit from humans when constructing our supervision signals impacts downstream performance: a selective ablation for increasing levels of sparsity. In our real-world soft label experiments, we in-fill missing CIFAR-10S labels by simulating if CIFAR-10DS labels had only provided uncertainty over $\hat{k} = 2$ labels. As we have softness over all $k = 10$, we can simulate varying $\hat{k}$. We train additional models in-filling with $\hat{k} = 3$ and 4,

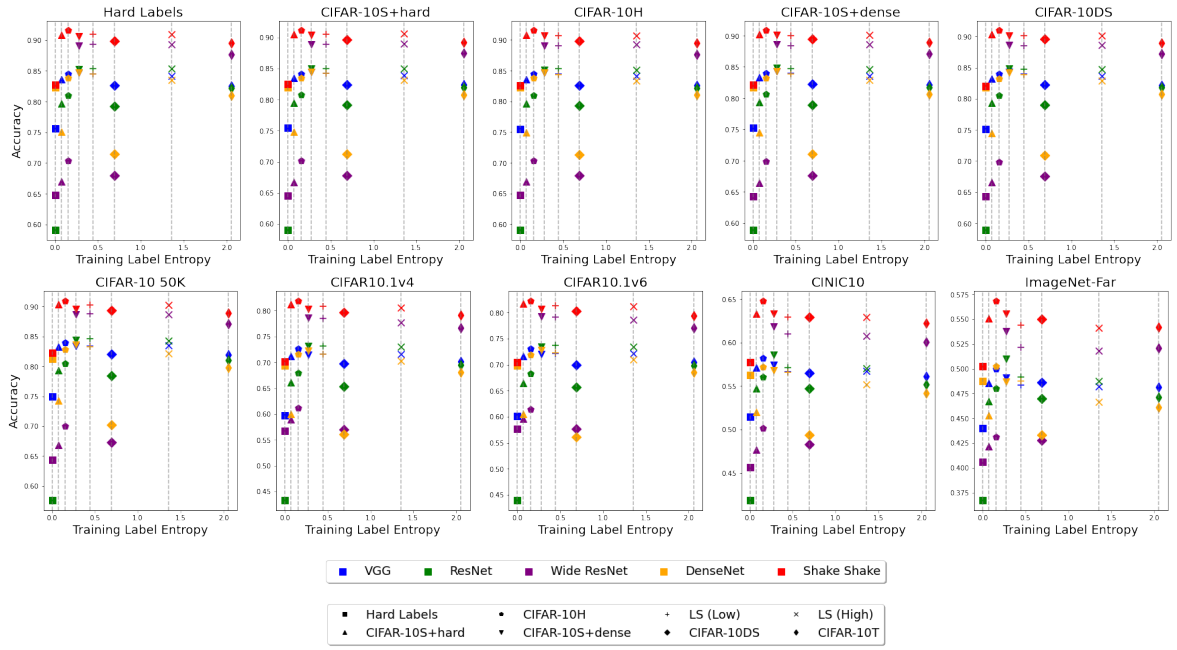

Figure 4: Cross-label (top) and generalization results (bottom), scored by top-1 accuracy against the respective labels.

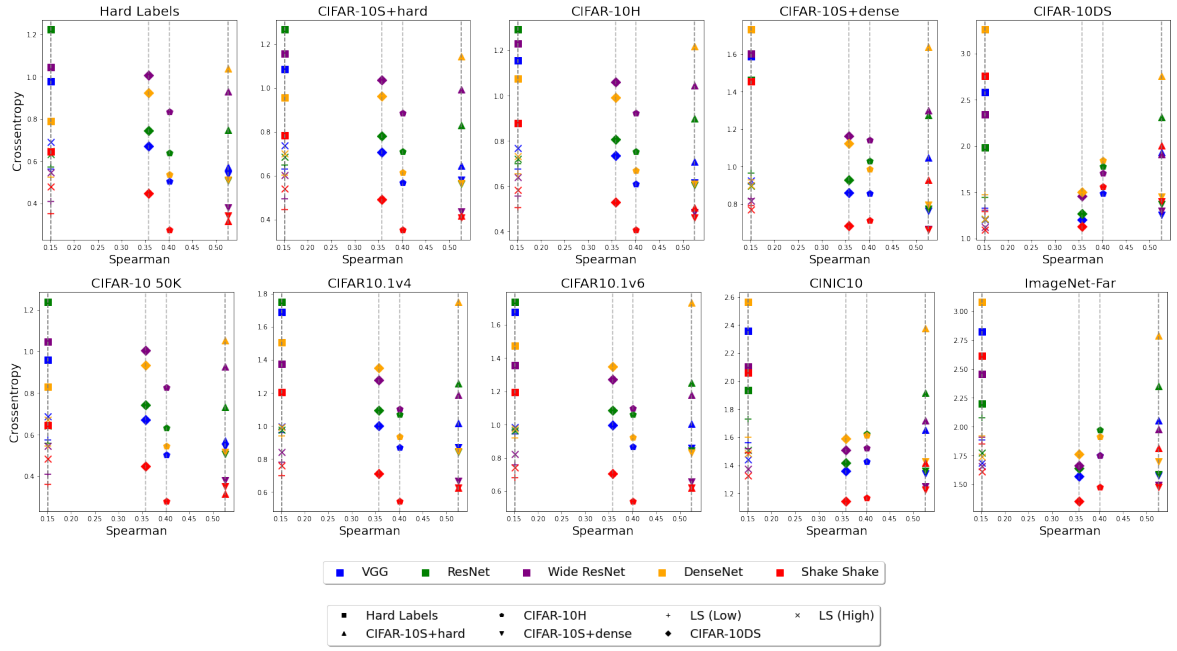

Figure 5: **Top:** Model performance on different label types at test time, as a function of information content of the labels. Information content is approximated by the Spearman rank correlation with similarity judgments. **Bottom:** Generalization performance under increasing distributional shift, as a function of training label information content.

respectively. We find that the extra softness does not add substantial value over $\hat{k} = 2$ when evaluating on near-domain generalization (i.e. CIFAR10-50k, .1v4, and .1v6), but does appear to have a positive effect when evaluating on further out-of-domain generalization (e.g., CINIC10 and ImageNet-Far).

E.6 REPRESENTATIVE IMAGES AND MODEL PREDICTIONS

In Figures 6 and 7 we present the images used as the basis of the similarity experiments (see above for details on label and similarity judgment collection). These images were chosen *using our trained classification models* to include images that

Table 2: Crossentropy (lower is better) as a function of varying the classes we permit human uncertainty specification over.

Classes per Label	CIFAR10-50k	CIFAR10.1v4	CIFAR10.1v6	CINIC10	ImageNet-Far
k = 2	**0.46**	**0.77**	**0.76**	1.32	1.57
k = 3	0.48	**0.77**	**0.76**	1.30	1.51
k = 4	0.49	**0.77**	0.77	**1.26**	**1.47**
k = 10	0.76	1.09	1.08	1.40	1.60

were likely to have high label entropy (Figure 6), and images where model predictions diverged (Figure 7). The models making the predictions had not been trained on these images (i.e., the predictions were based on the held-out cross-validation folds).

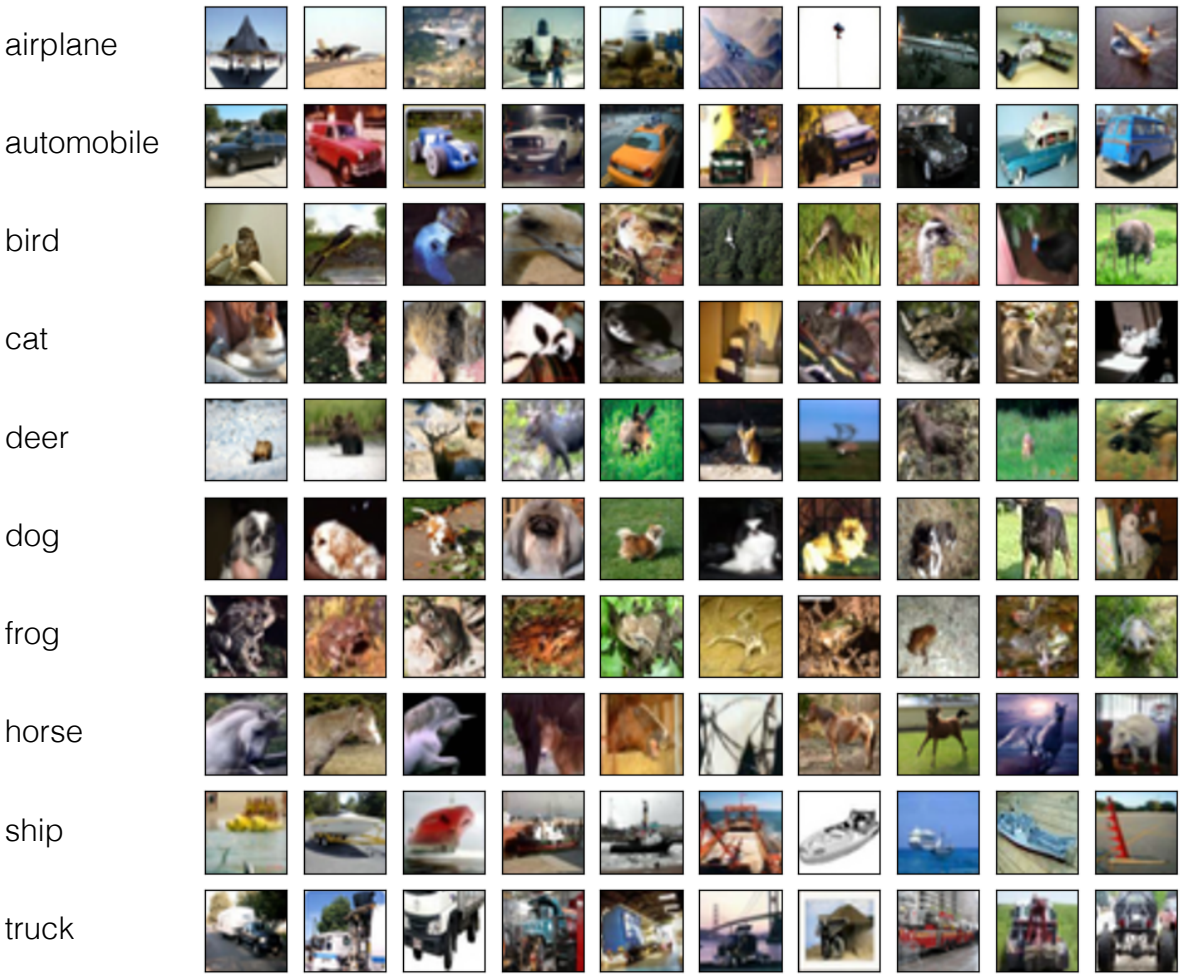

Figure 6: 100 images from the `CIFAR-10` testing subset. These were chosen to include images that were likely to have high label entropy.

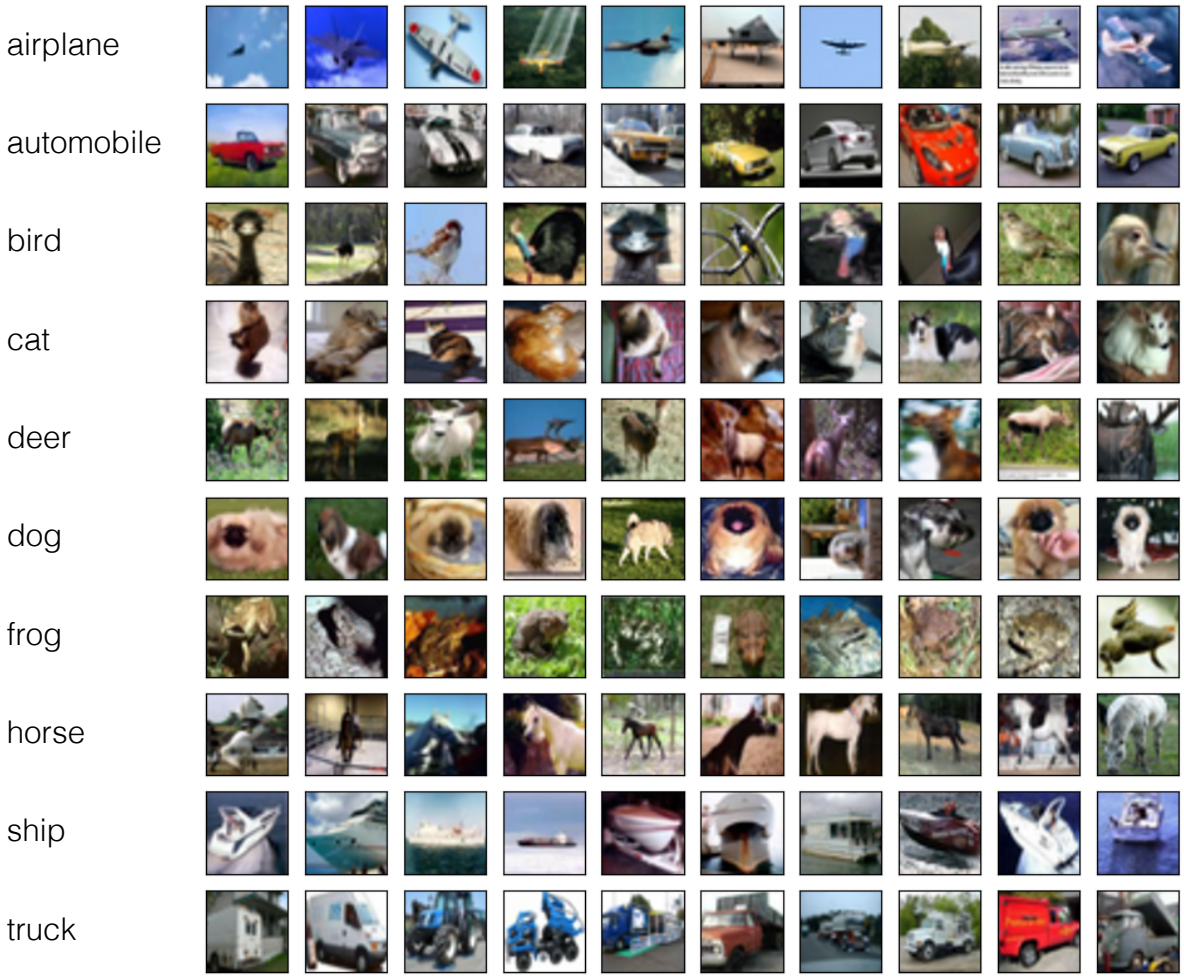

Figure 7: 100 images from the `CIFAR-10` testing subset. These were chosen to include images that were likely to cause model disagreement.

In Figures 8-17, we present four exemplars from each class, along with the soft labels and model predictions. We see that this analysis picks out genuinely ambiguous images, with borderline cases between two classes, many classes, noisy images, and categorically uncertain images. For each image, the top row are images in which models agree on high likely entropy (average prediction entropy). The bottom row is where models maximally disagree (average symmetric relative entropy between pairwise comparisons of models).

F ENLARGED MAIN RESULTS FIGURE

See Figure 18.

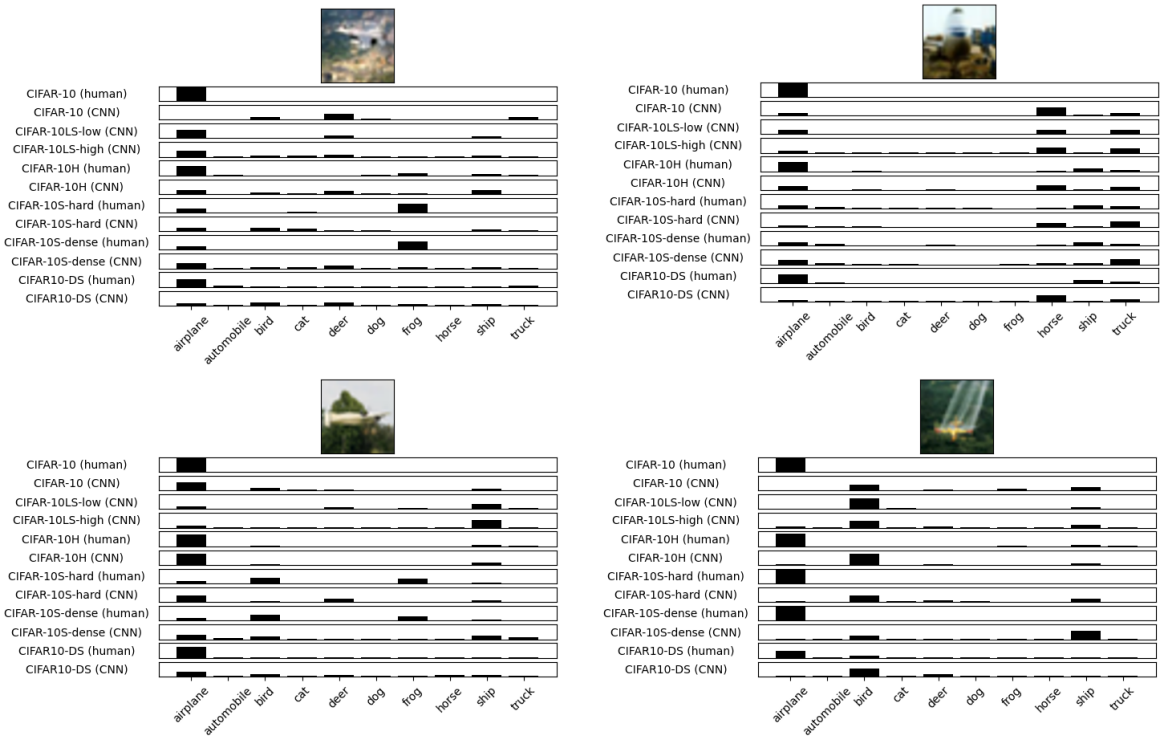

Figure 8: Representative ambiguous plane images. Top row: model agreement on high entropy image. Bottom row: maximal model disagreement.

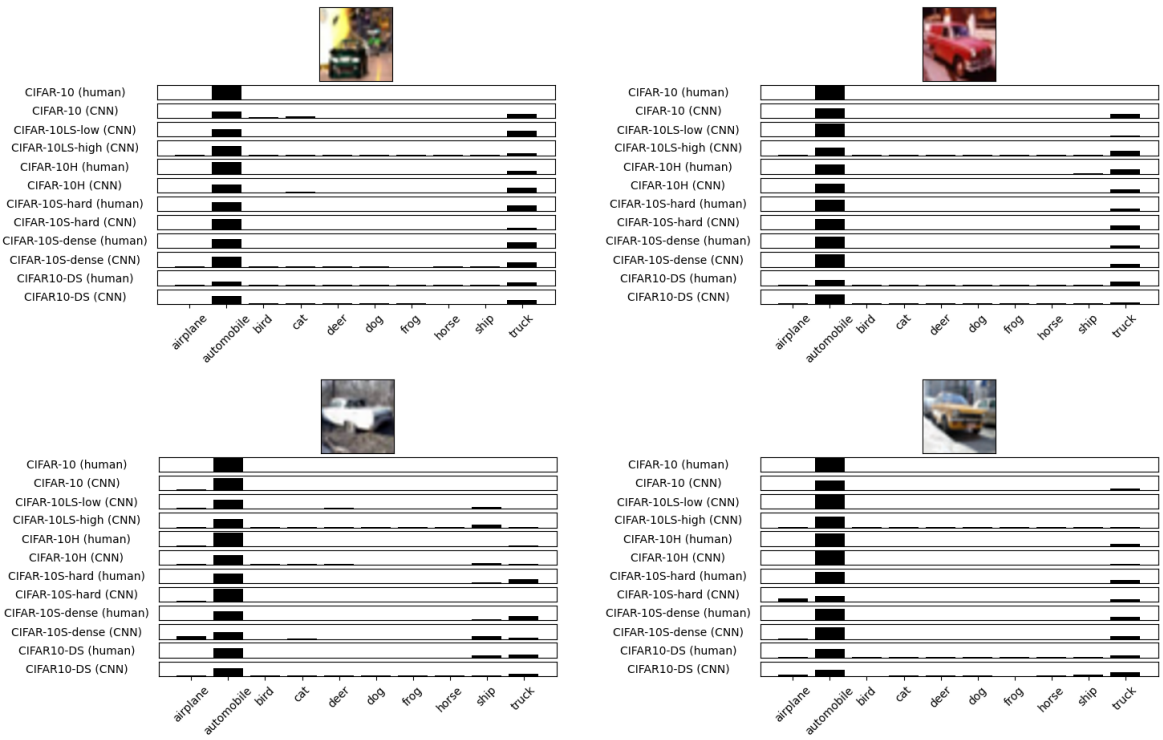

Figure 9: Representative ambiguous automobile images. Top row: model agreement on high entropy image. Bottom row: maximal model disagreement.

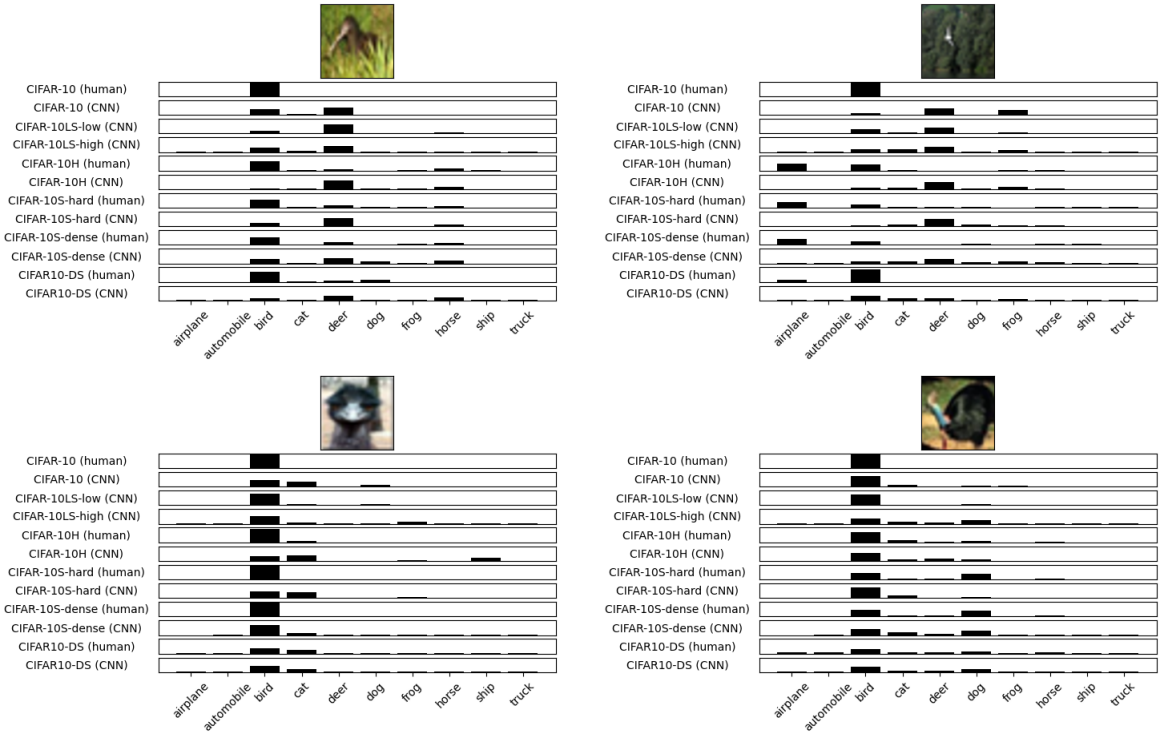

Figure 10: Representative ambiguous bird images. Top row: model agreement on high entropy image. Bottom row: maximal model disagreement.

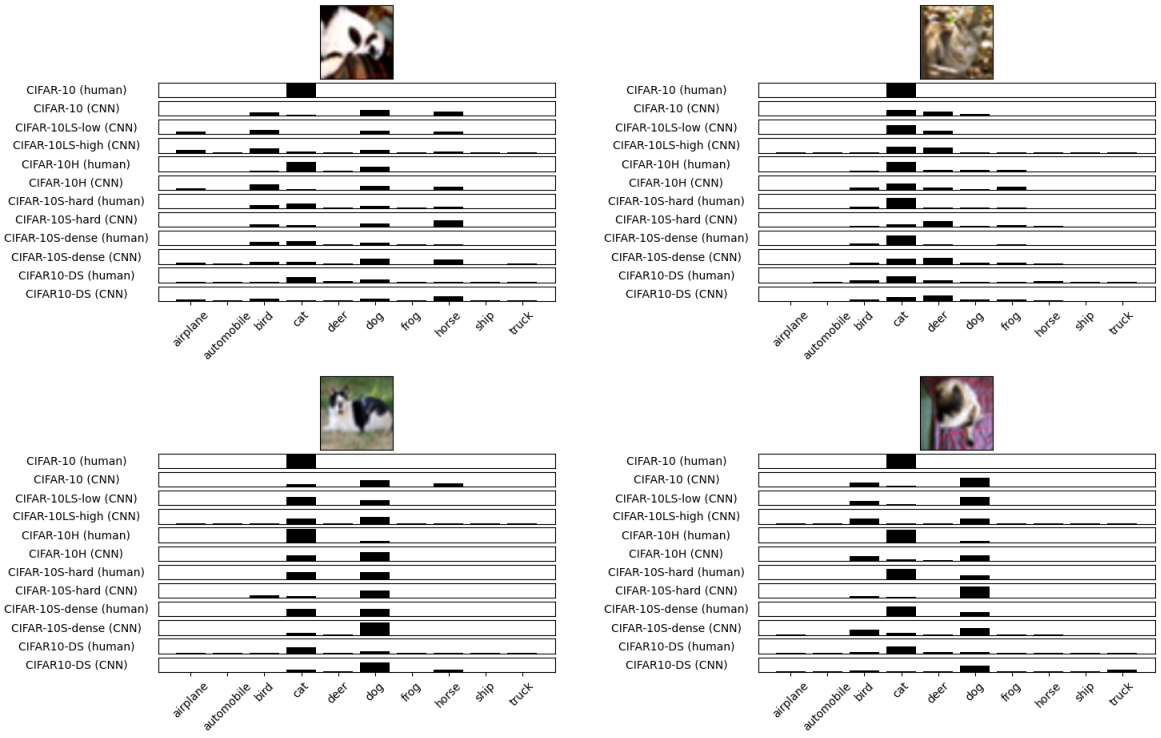

Figure 11: Representative ambiguous cat images. Top row: model agreement on high entropy image. Bottom row: maximal model disagreement.

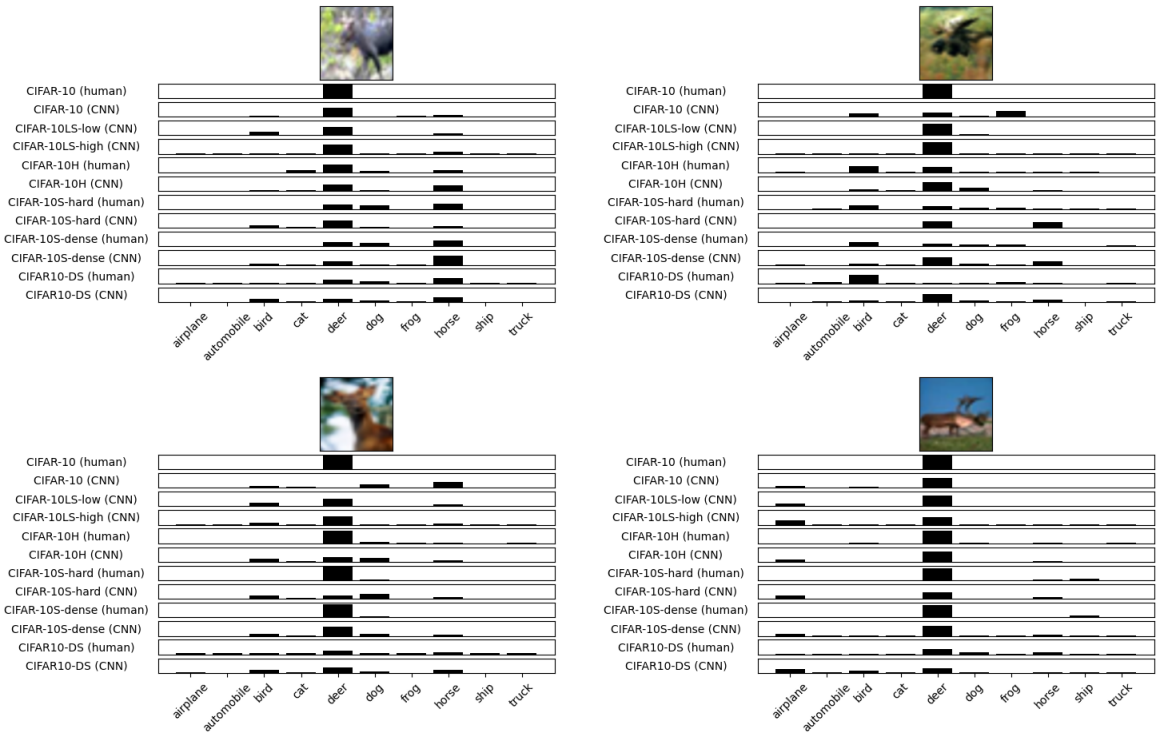

Figure 12: Representative ambiguous deer images. Top row: model agreement on high entropy image. Bottom row: maximal model disagreement.

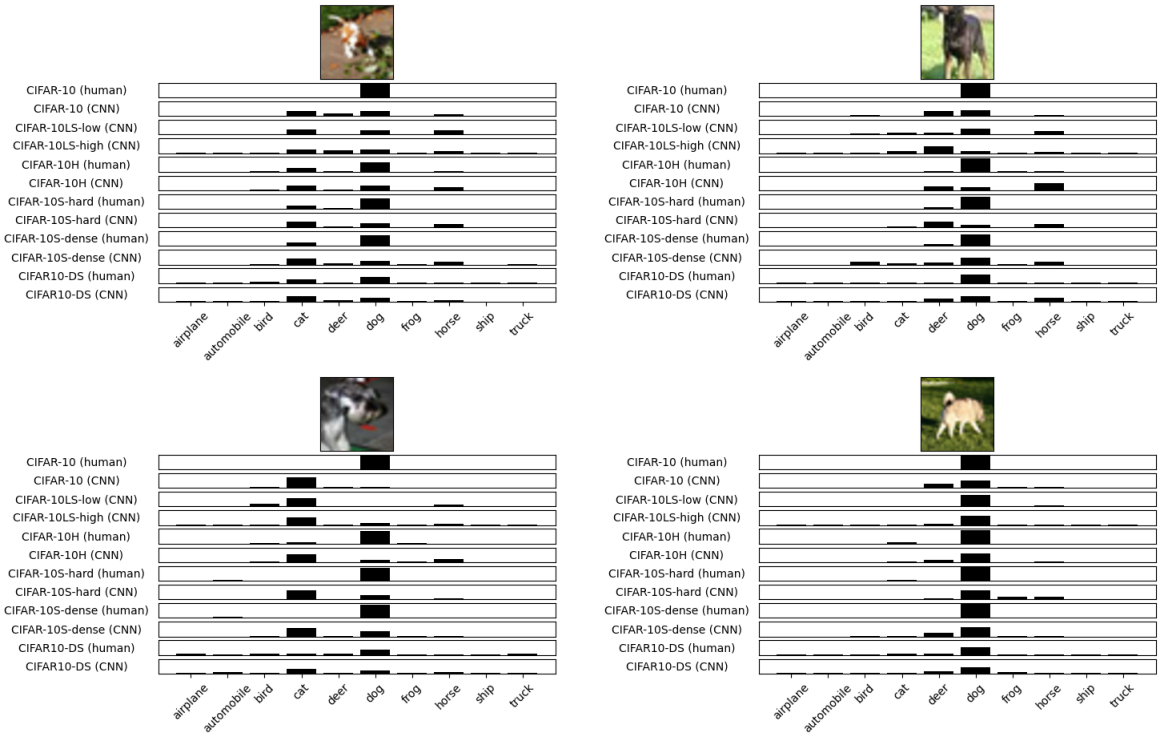

Figure 13: Representative ambiguous dog images. Top row: model agreement on high entropy image. Bottom row: maximal model disagreement.

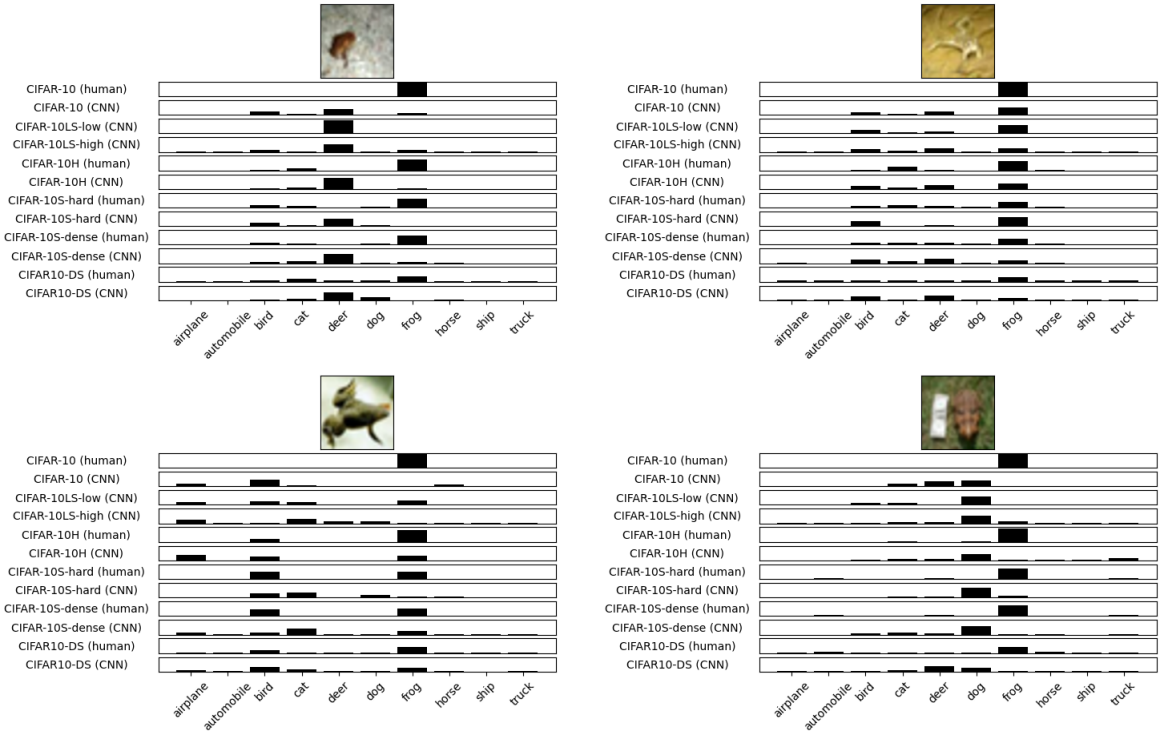

Figure 14: Representative ambiguous frog images. Top row: model agreement on high entropy image. Bottom row: maximal model disagreement.

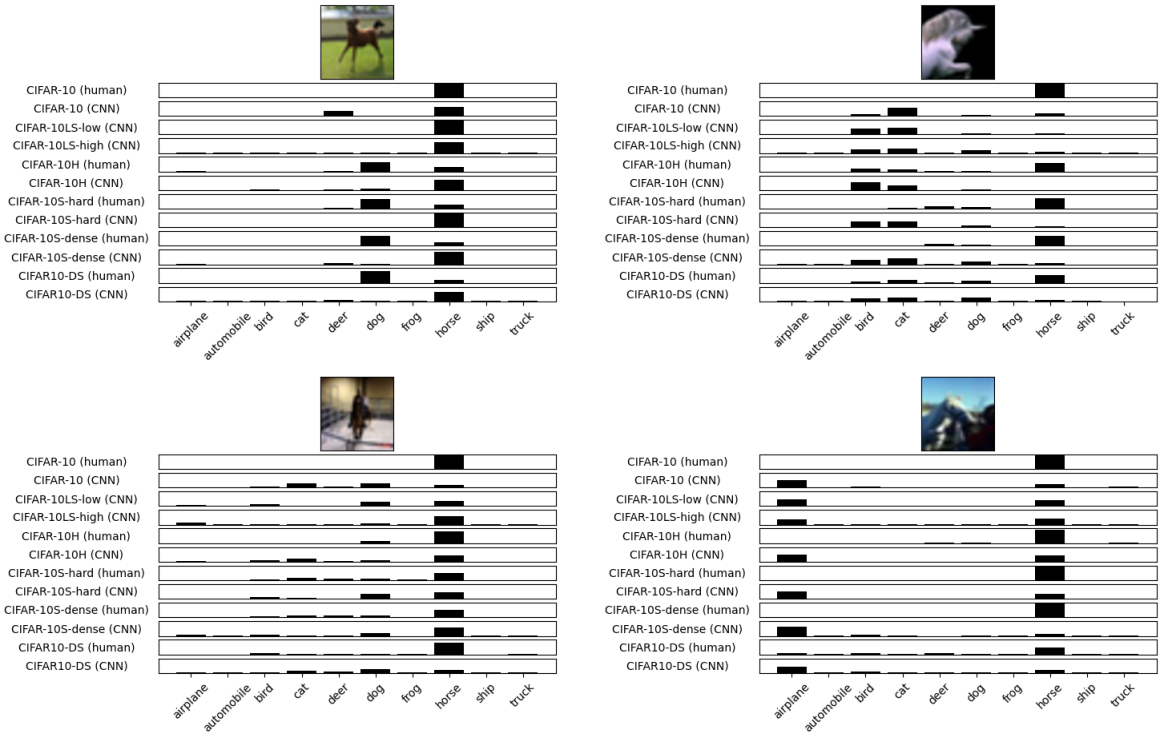

Figure 15: Representative ambiguous horse images. Top row: model agreement on high entropy image. Bottom row: maximal model disagreement.

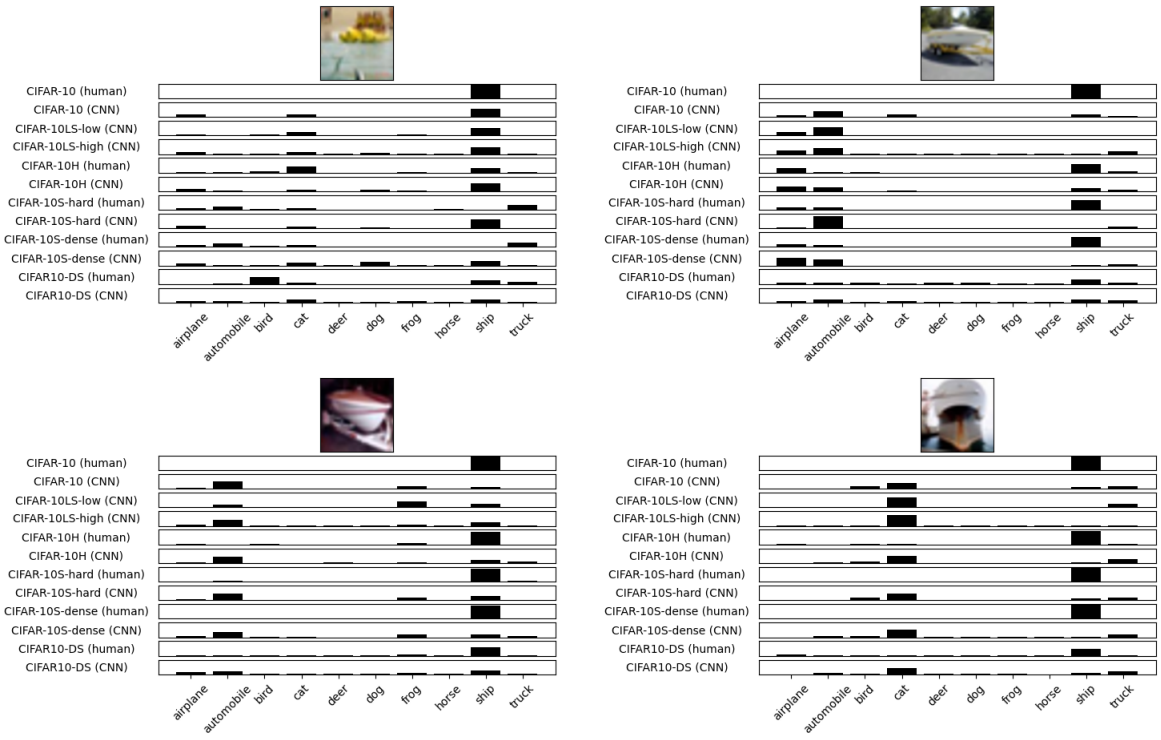

Figure 16: Representative ambiguous ship images. Top row: model agreement on high entropy image. Bottom row: maximal model disagreement.

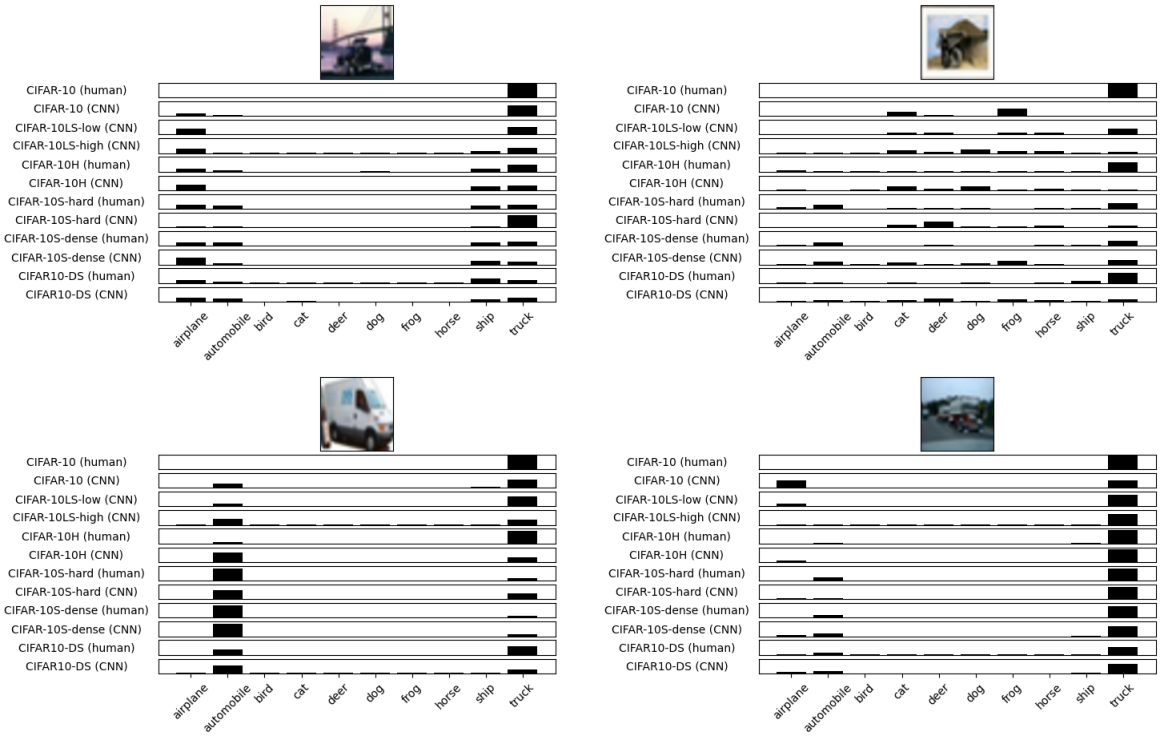

Figure 17: Representative ambiguous truck images. Top row: model agreement on high entropy image. Bottom row: maximal model disagreement.

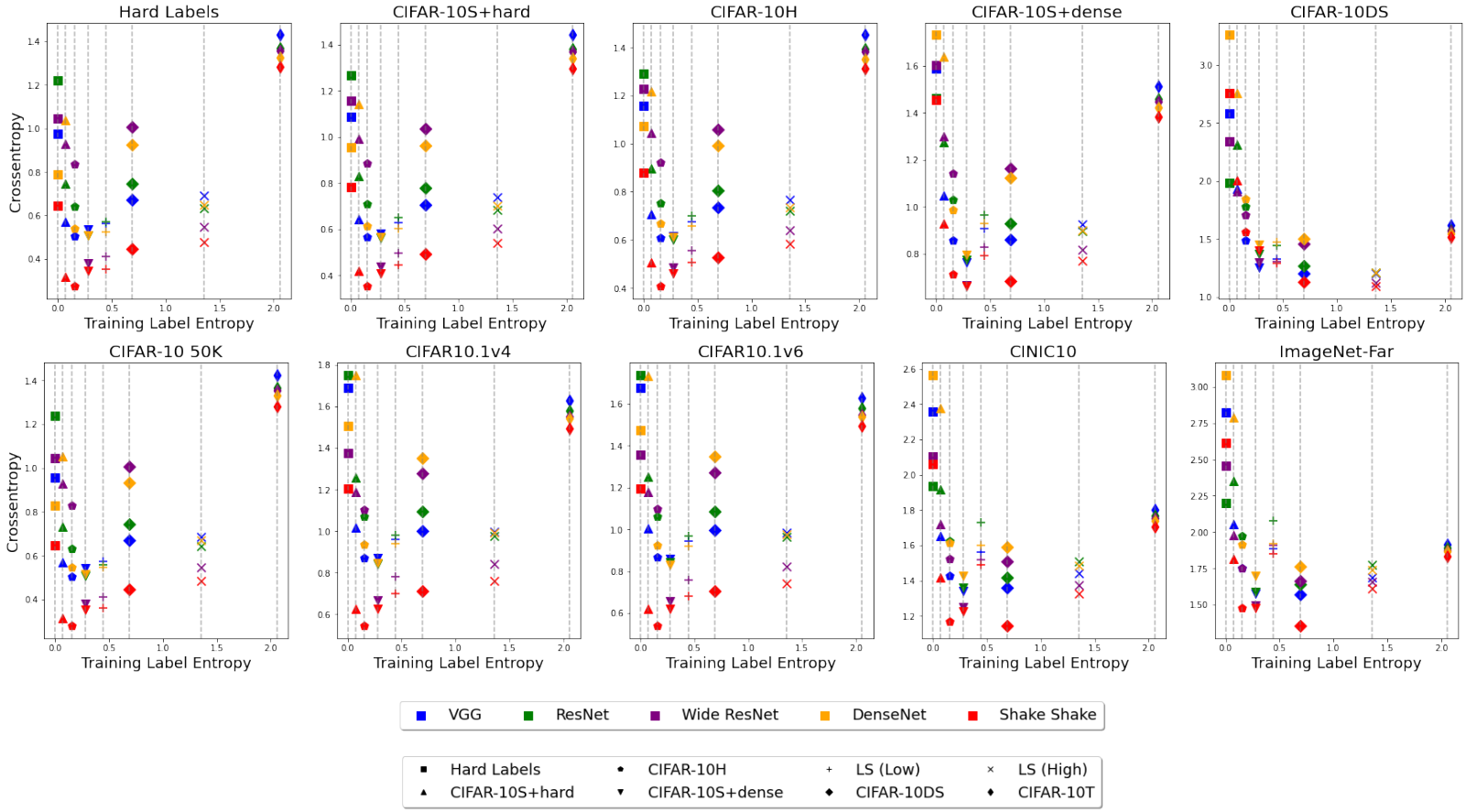

Figure 18: **Top:** Model performance on different label types at test time. **Bottom:** Generalization performance under increasing distributional shift. Each point represents the average score for a single model architecture (specified by color), trained on a particular label type (indicated via shape). Vertical lines represent points for a given label type.

References

Ruairidh M Battleday, Joshua C Peterson, and Thomas L Griffiths. Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature communications*, 11(1):5418, 2020.

Katherine M Collins, Umang Bhatt, and Adrian Weller. Eliciting and learning with soft labels from every annotator. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 40–52, 2022.

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.

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

Peter Harrison, Raja Marjieh, Federico Adolfi, Pol van Rijn, Manuel Anglada-Tort, Ofer Tchernichovski, Pauline Larrouy-Maestri, and Nori Jacoby. Gibbs sampling with people. *Advances in neural information processing systems*, 33: 10659–10671, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Aditi Jha, Joshua C Peterson, and Thomas L Griffiths. Extracting low-dimensional psychological representations from convolutional neural networks. *Cognitive Science*, 47(1):e13226, 2023.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9617–9626, 2019.

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

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

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.

V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971. doi: 10.1137/1116025. URL `https://doi.org/10.1137/1116025`.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.