Combining Human Predictions with Model Probabilities via Confusion Matrices and Calibration

Anonymous Author(s)
Affiliation
Address
email

Abstract

An increasingly common use case for machine learning models is augmenting the abilities of human decision makers. For classification tasks where neither the human or model are perfectly accurate, a key step in obtaining high performance is combining their individual predictions in a manner that leverages their relative strengths. In this work, we develop a set of algorithms that combine the probabilistic output of a model with the class-level output of a human. We show theoretically that the accuracy of our combination model is driven not only by the individual human and model accuracies, but also by the model's confidence. Empirical results on image classification with CIFAR-10 and a subset of ImageNet demonstrate that such human-model combinations consistently have higher accuracies than the model or human alone, and that the parameters of the combination method can be estimated effectively with as few as ten labeled datapoints.

1 Introduction

One of the main goals of machine learning is to develop algorithms that can operate robustly in an autonomous fashion without human supervision. However, there are many applications where hybrid human-machine approaches are likely to be a preferred mode of operation, for a variety of different reasons, such as improving trust between humans and machines, and allowing for a human or a model to take over in situations where the one or the other lacks expertise [1, 2, 3, 4, 5, 6, 7].

The performance benefits of combining multiple predictors, rather than relying on a single predictor, have been clearly demonstrated in past work in a variety of fields. For example, in machine learning there is a rich vein of research over the past few decades on combining models using a variety of different estimation and algorithmic approaches [8, 9, 10, 11]. This work emphasizes the point that combinations of models that have diversity in how they make predictions can systematically outperform a single model. In parallel, in the behavioral science literature, there has been extensive prior work studying combinations of human opinions where, again, diverse combinations tend to outperform any single individual [12, 13].

This naturally leads to questions about hybrid combinations of human and machine predictions, rather than just combining one type or the other. For example, one motivation for hybrid combinations is empirical evidence that human and machine classifiers do not make the same types of errors for problems such as image classification [14, 15, 16], i.e., they are diverse in their predictions. These ideas have begun to have impact in real-world applications, where hybrid human-machine teams have been found to be effective in areas such as crowdsourcing [17], citizen science [18], speech transcription [19], face identification [20], and clinical radiology [21, 22, 23].

In this paper we focus on a specific, simple, and important instantiation of the general problem of hybrid combinations of human and machine predictions. In our problem we consider a K-way classification problem such as image classification, with a single human making hard classification

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

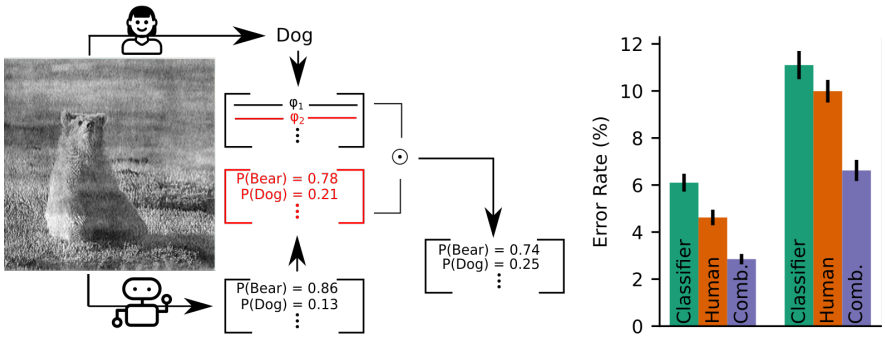

Figure 1: Left: Combining a human's label and a classifier's probabilities for an ImageNet-16H image (true label: bear). Right: Human-machine combinations (purple) achieve lower error rates on average than the human or classifier alone (ResNet-164 on CIFAR-10H, VGG-19 on ImageNet-16H).

decisions (no confidence estimates) and a single classification model providing class-conditional probability vectors. While humans can provide confidence estimates with their label predictions, calibrated self-assessment of confidence can be difficult [24, 25, 26] as well as being time-consuming.

A key question in this context is whether the non-probabilistic information from the human predictor can be effectively combined with the probabilistic information from a machine learning model. We answer this question in the affirmative and show both theoretically and empirically how relatively simple probabilistic combination techniques can robustly outperform each of a human and machine on their own. In particular we show that a human can augment their predictions with those of a classification model, improving classification accuracy and producing calibrated predictions, even when the model is less accurate than the human. Similarly, from the model's perspective, accuracy can often be significantly improved by augmenting its' class probabilities with a human's labels, while improving calibration performance, even when the human is less accurate than the model.

Figure 1 shows an example using our proposed methodology for an image from the ImageNet dataset. The human incorrectly predicts the label `dog`. The classification model (a VGG-19 deep network) predicts the correct label `bear` with a probability of 0.86 (uncalibrated) and 0.78 (calibrated). The combined prediction is for `bear` with a confidence of 0.74, with `dog` having a higher confidence given the human's prediction. The histograms on the right show the overall average reduction in error rate on test images (ResNet-164 on CIFAR-10H [27], VGG-19 on ImageNet-16H [28]): even though the classifiers are on average less accurate (6.10%, 11.1% error) than the human (4.62%, 9.99% error), the resulting combinations have lower error rates than either (2.85%, 6.62% error).

The primary contributions of our work are as follows:

- We propose and investigate a general framework for combining predictions using instance-level confidence from a model and class-level information from a human. The methods we propose are straightforward to implement in practice and label-efficient.[1]
- We empirically validate our approach on the CIFAR-10H and ImageNet-16H image classification datasets and show that human-machine combinations in this context are systematically more accurate and better calibrated than either alone.
- We develop a theoretical understanding in this framework of the key tradeoffs related to calibration and accuracy for both the individual human and model, introducing the notion of model and human confidence ratios. We illustrate how these factors affect the combination, showing for example how two models with the same accuracy but with different calibration properties can have different performance when combined with human predictions.

The paper begins in Section 2 by introducing notation and background concepts. Section 3 discusses related work and in Section 4 we propose a number of different estimation methods. Section 5 describes our experimental results with two image classification datasets, with individual human labelers, individual models, and combinations. Complementing the experimental results in Section 6

[1]Code for our estimation methods and experiments is available at `https://github.com/neurips21-anon/conf_matrix_and_calibration`.

we develop theoretical results that characterize combination performance. Section 7 concludes the paper including a discussion of potential societal impact and limitations.

2 Combining Human Labels and Model Probabilities

Notation. We consider a K-ary classification problem, where the goal is to predict a label $y \in \mathcal{Y} = \{1, \ldots, K\}$ from features $x \in \mathcal{X}$. The random variable (x, y) follows an unknown distribution with support $\mathcal{X} \times \mathcal{Y}$. We assume access to an individual human labeler represented by the function $h : \mathcal{X} \to \mathcal{Y}$, where $h(x) \in \mathcal{Y}$ is the label predicted by the human. In addition, we have access to a trained machine classifier $m : \mathcal{X} \to \mathbb{R}^K$, where $m(x)$ is the normalized probability vector output by the classifier. Both the human and classifier are assumed to be noisy labelers relative to the ground truth y. The true labels could be determined, for example, by expert labelers or additional information not contained in x.

Combining Predictions. Given an input $x \in \mathcal{X}$, our goal is to predict a true label conditioned on the predictions $h(x)$ and $m(x)$. The key challenge in combining human and classifier predictions is simultaneously leveraging both the class-level outputs from the human and the predictive distributions output by the classifier. Although we focus on the particular case where h is a human and m is a classifier, our setup could be applied more generally to combinations of a non-probabilistic and a probabilistic labeler.

There are a variety of functional forms that could be used to combine the predictions. We pursue a probabilistic approach, where the conditional distribution over labels that we seek can be factored via Bayes' rule as

$$p(y|h(x), m(x)) \propto p(h(x)|y, m(x))p(y|m(x)). \tag{1}$$

It is natural in this context to pursue a conditional independence (CI) approach, where the human labels $h(x)$ and the probabilistic predictions $m(x)$ are assumed to be conditionally independent given y. Under this assumption we can write

$$p(y|h(x), m(x)) \propto p(h(x)|y)p(y|m(x)) \tag{2}$$

where the right-hand side terms have a natural interpretation in terms of calibrated probabilities at the class level $\big(p(h(x)|y)\big)$ and at the instance level $\big(p(y|m(x))\big)$.

We parameterize the first term of Equation (2) by the the confusion matrix for the labeler h, which we denote by φ with entries $\varphi_{ij} = p(h(x) = i|y = j)$. On the other hand, the probabilistic output of the classifier $m(x)$ may differ from $p(y|m(x))$. For example, modern neural networks tend to be overconfident in their predictions [29]. To remedy this, post-hoc calibration maps $m(x)$ to well-calibrated probabilities via a learned calibration map with parameters θ. In this work, we use $m^\theta(x)$ to denote the output of the classifier after applying such a calibration map. The second term in Equation (2) is then parameterized by the calibrated classifier probabilities $m^\theta(x)$. Altogether, our method expresses the predicted probability for class j as:

$$\boxed{p(y = j|h(x) = i, m(x)) = \frac{\varphi_{ij} m_j^\theta(x)}{\sum_{i=1}^K \varphi_{ij} m_j^\theta(x)}} \tag{3}$$

The CI assumption above is common (both implicitly and explicitly) in prior work on combining predictions, such as additive classifier ensembles [10, 11] and (log-)linear opinion pools [30, 31]. As our primary motivation is to develop a relatively simple and robust methodology for combining human and model predictions, the additional functional or parametric assumptions (and parameters) required to specify a joint model for $p(h(x), m(x)|y)$ are beyond our scope. In addition, although the CI assumption is unlikely to hold exactly, prior work [32] notes that a CI model can be an optimal discriminant even when the CI assumption is violated. As further motivation, for the two datasets we use in this paper, CIFAR-10H and ImageNet-16H, the conditional dependence of $h(x)$ and $m(x)$ appears to be relatively weak (see Appendix B for details).

3 Related Work

We summarize below relevant aspects of related literature. While there is a significant amount of prior work in machine learning and related fields on combining predictors, this work has in general not

addressed the specific problem of combining hard label predictions from a human with probabilistic label predictions from a model.

Ensembles and Opinion Pools. There is a rich literature in machine learning on studying predictions based on ensembles of classification models. For non-probabilistic classifiers, the most common aggregation methods are variants of (weighted) majority voting [9, 11]. However, in our case of only two predictors, a weighted majority vote ensemble can never improve accuracy over its components. Beyond majority voting, naive Bayes aggregation [33, 10] fits a class-level confusion matrix to each predictor. Kim and Ghahramani [34] develop a fully Bayesian extension of this, which relaxes the independence assumption by explicitly modeling correlations between predictors. However, because these confusion-matrix aggregation methods are at the class level they are unable to take full advantage of the instance-level uncertainties produced by the probabilistic labeler.

In the context of aggregating predictions from multiple humans, there has been a considerable amount of prior work in the behavioral sciences and forecasting literature. Approaches include additive linear and log-linear opinion pools for subjective distributions [30, 31], techniques for weighting linear combinations of real-valued human predictions [13, 35], and voting methods for combining label predictions from more than two human predictors [36]. A key difference between these methods and our work is that they do not address the problem of how to combine probabilistic and non-probabilistic predictions in a human-machine context.

Leveraging Human and Model Predictions. Combining human predictions with model predictions to solve classification problems has been a topic of recent interest in a number of different areas. For example, in [37] simple averaging is used to combine the labels of multiple human annotators with the output of a classifier for astronomical image classification, achieving better performance than with either the humans or the classifier. In crowd-sourcing, classification models have been used to automatically filter examples to improve human annotation efficiency [17, 38, 2, 5]. A similar line of research focuses on algorithmic deferral techniques where a model defers to human predictions based on the model's confidence [39, 40], as well as work on adapting prediction models to the human decision maker [41, 42, 43, 44]. The results in [41] in particular describe experiments with the same CIFAR-10H dataset that we use in this paper. However, in addition to being different to our work in terms of its focus on deferral (rather than combining) we also note that in [41] that improvements in performance are demonstrated using relatively large numbers of human labels. In contrast, as we demonstrate in Section 5 on the CIFAR-10H and ImageNet-16H datasets, our methods require only a small number of human labels to yield combined predictions that are more accurate than either human or model alone. In general, existing work on filtering and deferral strategies complements the combining methods that we develop in this paper. All of these approaches are useful in a broad range of human-AI applications, but in different contexts.

4 Estimation Methods and Algorithms

Combining human and machine predictions via Equation (3) requires learning two sets of parameters: confusion matrix parameters for the human and calibration parameters for the classifier. The choice of procedure used to infer these parameters impacts the label efficiency and quality of the resulting combination. In this section, we detail several inference procedures and empirically evaluate them in the context of human-machine combinations.

To estimate our combination model, we assume access to a combination dataset $\mathcal{D}_C = \{(h(x_\ell), m(x_\ell), y_\ell)\}_{\ell=1}^n$ with human labels, machine probabilities, and ground truth labels. We assume the classifier is pre-trained with true labels on a separate training set $\mathcal{D}_T$.

Confusion Matrix Estimation. Recall that φ denotes a a confusion matrix for the human of shape $K \times K$, where φ_{ij} is $p(h(x) = i|y = j)$. The most straightforward estimate for this quantity is the maximum likelihood estimate, where φ_{ij} is estimated by the number of datapoints in $\mathcal{D}_C$ where the human labeler predicts $h(x) = i$ when the ground truth is $y = j$, normalized by the number of points in $\mathcal{D}_C$ where $y = j$.

However, as the size of the confusion matrix is quadratic in K, this estimate will have high variance for small amounts of labeled data and collecting enough human labels to overcome this variance could be prohibitively expensive. We can instead take a Bayesian approach and incorporate informative

Table 1: Summary of combination methods studied in this work. Except for logistic regression, parameter counts correspond to calibration using MAP temperature scaling (one parameter), and confusion matrices are fit with MAP inference. The human output is always a label.

Method Name	Acronym	Parameters	Model Output	Ground Truth?	Label Efficient?
Logistic Regression	LR	$k^2 + 2k$	Probabilities (P)	✓	X
Calibrated Machine Probs & Single-Parameter Confusion	SP	2	Probabilities (P)	✓	✓
Machine Labels & Human Labels	L+L	$2k^2$	Labels (L)	✓	X
Calibrated Machine Probs. & Human Labels	P+L	$k^2 + 1$	Probabilities (P)	✓	✓
Calibrated Machine Probs. & Human Labels (EM)	P+L-EM	$k^2 + 1$	Probabilities (P)	X	✓

prior information. Given the true label $y = j$, the human label $h|y = j \sim \text{Discrete}(\varphi_{*j})$ is assumed to be drawn from a discrete distribution with parameters corresponding to the jth row in the confusion matrix. We place a conjugate Dirichlet prior $\varphi_{*j} \sim \text{Dirichlet}(\alpha_j)$ over each column with parameters $\alpha_j \in \mathbb{R}^k$. The prior parameters α_j are chosen such that

$$(\alpha_j)_i = \begin{cases} \beta & i \neq j \\ \gamma & i = j \end{cases}$$

That is, the prior matrix is $\gamma \in \mathbb{R}_{>0}$ along the diagonal and $\beta \in \mathbb{R}_{>0}$ on the off-diagonal. We choose β and γ such that the resulting Dirichlet distribution has mode equal to the train-set accuracy of the classifier, which can be obtained without additional human labels. This choice of prior reflects our belief that the confusion matrix will have a diagonally dominant structure. Posterior estimates of the confusion matrix can then be obtained straightforwardly by conjugacy.

Calibration Parameter Estimation. Scaling-based calibration maps are typically fit by optimizing the log-likelihood [29, 45, 46]. In this section, we detail a Bayesian version of temperature scaling [29], allowing us to incorporate informative prior information. A temperature $T \in (0, 1)$ indicates underconfidence, and $T \in (1, \infty)$ indicates overconfidence. To account for this difference in scale, we place a Gaussian prior on the log-temperature $\log T = \tau \sim \mathcal{N}(\mu, \sigma^2)$. As this is a non-conjugate prior, the maximum a posteriori (MAP) temperature is estimated via gradient-based optimization. In our experiments, we choose $\sigma = 0.5$ for the CIFAR-10 models and $\sigma = 0.75$ for the ImageNet models, and we use $\mu = 0.5$ throughout. These parameters were chosen to reflect our belief that deep models tend to be overconfident and to concentrate the prior on reasonable temperature values. In Appendix G, we derive a fully Bayesian approach where we marginalize over the posterior distribution over temperature (e.g. using Monte Carlo methods [47]). However, we find empirically that the simpler MAP approach is more effective and, as a result, focus on MAP estimation in this paper.

Learning without Ground Truth. Requiring both human and ground truth labels in $\mathcal{D}_C$ can be a potentially limiting assumption in domains where ground truth labels are unavailable or expensive to obtain. To avoid this, we propose an unsupervised approach that is able to learn both calibration parameters and confusion matrix parameters from a combination dataset of the form $\mathcal{D}_C = \{(h(x_\ell), m(x_\ell))\}$, consisting only of human labels and machine probabilities. We treat the ground truth labels as latent and fit the required parameters using Expectation-Maximization (EM) [48] (details in Appendix I). This approach can be seen as a novel extension of the Dawid-Skene model [49], where calibration parameters are fit for the model rather than a confusion matrix. We perform both maximum likelihood and MAP estimation with this method, using the same priors as above.

For clarity of exposition, we focus on three types of combinations in our results:

- **Machine Labels & Human Labels (L+L)**, a baseline where the instance-level probabilities from the model are discarded and instead a confusion matrix is fit for the model, as well as a confusion matrix for the human. The confusion matrices are estimated via supervised MAP inference. This can be viewed as a naive Bayes' combination [10, Chapter 4] for non-probabilistic predictors.
- **Calibrated Machine Probabilities & Human Labels (P+L)** combined via Equation (3) using *supervised* MAP estimates for both the calibration parameters and confusion matrix parameters.
- **Calibrated Machine Probabilities & Human Labels (P+L-EM)** combined via Equation (3) using *unsupervised* MAP estimates fit with our EM algorithm, i.e. using human labels but no ground truth.

In terms of complexity, these methods sit between a simple model with only one parameter for the human confusion matrix **(SP)** (where the diagonal entry corresponds to the human's marginal accuracy), and a full multinomial logistic regression model (**LR**).[2] We find that **LR** can obtain slightly lower error rates in some cases, but requires significantly more labeled data than the other methods to do so. At the other extreme, **SP** is highly data efficient, but underfits compared to our preferred methods. We provide additional discussion of these methods to Appendix H. The various estimation methods discussed in this section are summarized in Table 1.

5 Experiments

Datasets and Models. We evaluate various combination strategies on two pre-existing image classification datasets that include human annotations: CIFAR-10H [27] and ImageNet-16H [28]. CIFAR-10H contains 10-way human classifications for 10,000 images from the standard CIFAR10 test set [50]. ImageNet-16H contains 16-way human classifications for noisy images from the ImageNet test set [51], distorted by phase noise at each spatial frequency based on four levels of phase noise (80, 95, 110, and 125). The human classifications for CIFAR-10H and ImageNet-16H come from the Amazon Mechanical Turk platform.

For both CIFAR-10H and ImageNet-16H, we select a single human label for each image by randomly sampling from the available human annotations. We experiment with four CNN models on CIFAR-10H (ResNet-110, Resnet-164 [52], PreResNet-164 [53], DenseNet [54]), and eight models (four VGG-19 [55], four GoogLeNet [56]) of varying accuracy on ImageNet-16H. Our models are chosen to span a range of performance, from below human accuracy to exceeding human accuracy. See Appendix E for details regarding our model architectures and training procedures.

Both datasets are partitioned into three disjoint subsets: (i) a model training set $\mathcal{D}_T$, (ii) a combination training set $\mathcal{D}_C$, and (iii) an evaluation set $\mathcal{D}_E$. The model training set is the same as the suggested training split for the original CIFAR-10 and ImageNet datasets, and is used to fit the classification models. The combination training set is used to estimate any calibration parameters and confusion matrices, and the held-out evaluation set is used solely for testing. The combination training set and evaluation set are subsets of the original evaluation sets, where 70% of the data is used for fitting the combinations and 30% is used for evaluation. The true labels (ground truth) for both CIFAR-10 and ImageNet correspond to the originally-provided labels for each of these datasets. In our experiments, (i) is fixed and we average over randomly selected splits for (ii) and (iii). The sampled human label for each image is fixed across all trials.

Calibration Methods. In our experiments, model calibration is done by MAP temperature scaling (TS) [29]. In Appendix C, we experiment with two additional calibration methods: ensemble temperature scaling [45] and I-Max binning [57]. We find that the combination performance is robust to the choice of calibration map, and hence restrict our focus to TS in this section. In addition, we find in general that combinations using uncalibrated model probabilities produce less accurate combinations than combinations using calibrated model probabilities (see Appendix C).

Learning Curves. Given that it is highly desirable to learn human-machine combinations from small amounts of data, we empirically study the data efficiency of the various inference methods previously described. In Figure 2, we plot the combination error rate on the evaluation data as a function of dataset size for the four CIFAR-10 models (first row) and the VGG-19 model on ImageNet (second row). For supervised methods, the dataset size corresponds to the number of $(h(x), m(x), y)$ triples used for learning, whereas for unsupervised methods the dataset size corresponds to the number of $(h(x), m(x))$ pairs without ground truth y.

As Figure 2 demonstrates, the P+L method is able to learn a human-model combination that outperforms both the human and model alone, with very few datapoints. While the P+L-EM method requires more human labels than the P+L method, it is able to learn such a combination without any ground-truth labels. The baseline L+L method fails to learn an effective combination on the CIFAR-10 dataset, and only does so on the VGG-19 ImageNet dataset with a large number of ground

[2] In fact, the CI combination with temperature scaling can be seen as a special case of multinomial logistic regression taking $m(x)$ and $h(x)$ as inputs (see Appendix F).

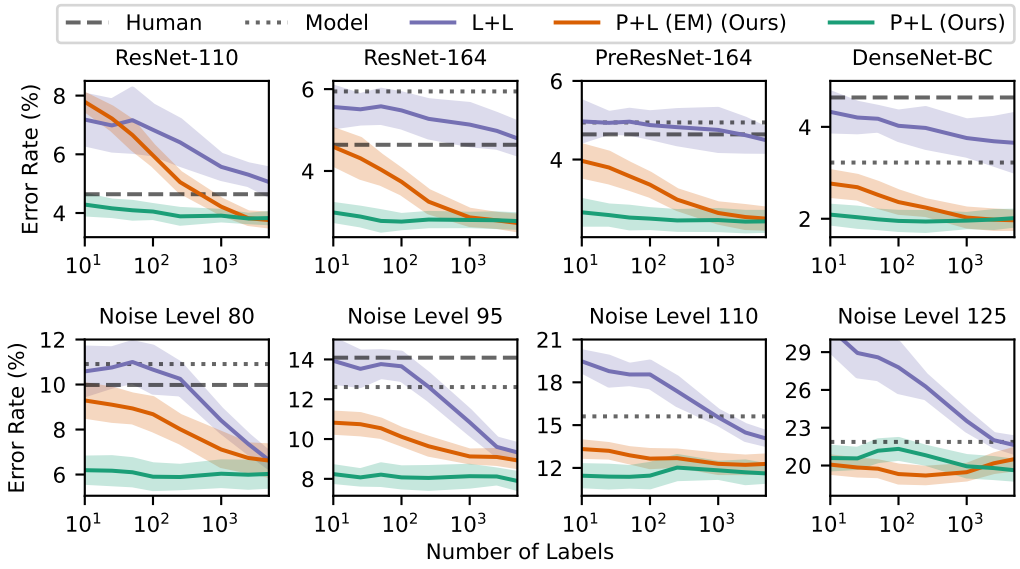

Figure 2: Learning curves for various models on CIFAR-10H (top) and VGG-19 on ImageNet-16H (bottom). For the supervised methods (L+L, P+L), the x-axis corresponds to the number of human labels, machine predictions, and ground truth labels. For the unsupervised method (P+L (EM)), the x-axis corresponds to the number of human labels and machine predictions (but no ground truth).

truth labeled datapoints. This demonstrates that the instance level probabilities from the model are a key component in efficiently learning human-model combinations with high accuracy.

In Appendix H, we provide similar plots for GoogLeNet and for maximum likelihood based inference procedures. In general, we find that maximum likelihood estimation requires more data than MAP estimation, and hence we focus our presentation of results on MAP.

Calibration Properties of Combinations. In addition to the error rate, we study the calibration properties of P+L combinations. In Table 2, we report the ECE [29], classwise-ECE (cwECE) [46, 57], and negative log-likelihood (NLL) for our various CIFAR-10 models (Model) and the resulting human-machine combinations (Comb.) on the held-out evaluation set. The ECE and cwECE are evaluated using 15 bins containing an equal number of data points. In addition to P+L combinations fit with 10 or 5000 labeled datapoints, we evaluate a combination consisting of the uncalibrated classifier probabilities with the MAP human confusion matrix estimated with 5000 labeled datapoints (No Calibration).

Combining classifier probabilities with human labels generally results in a combination that is better calibrated than the model alone. Moreover, MAP TS can be fit using a very small number of labeled datapoints. Our results show that the calibration properties (of both the classifier alone and the resulting human-machine combination) significantly improve with only ten labeled examples. However, increasing the number of labeled examples to 5000 does not result in further calibration gains. We provide similar results for our ImageNet-16H models in Appendix D.

6 Theoretical Analysis

Confidence Ratios. The key quantity in our analysis is the *confidence ratio* of a predictor, which is a random variable representing a predictor's confidence for the correct class relative to the predictor's confidence for other classes. This quantity can be thought of as the predictor's instance-level odds

Metric	Model Name	No Calibration		10 Datapoints		5000 Datapoints	
		Model	Comb.	Model	Comb.	Model	Comb.
ECE (10^{-2})	ResNet-110	5.23 ± 0.35	2.08 ± 0.25	3.03 ± 0.58	1.30 ± 0.23	2.99 ± 0.36	1.76 ± 0.18
	ResNet-164	2.98 ± 0.34	1.63 ± 0.23	1.95 ± 0.33	1.25 ± 0.18	1.89 ± 0.32	1.39 ± 0.18
	PreResNet-164	3.03 ± 0.29	1.87 ± 0.22	2.31 ± 0.33	1.40 ± 0.26	2.27 ± 0.31	1.43 ± 0.21
	DenseNet-BC	2.18 ± 0.27	1.53 ± 0.20	1.76 ± 0.28	1.34 ± 0.14	1.73 ± 0.28	1.27 ± 0.13
cwECE (10^{-2})	ResNet-110	0.81 ± 0.07	0.23 ± 0.05	0.58 ± 0.07	0.24 ± 0.05	0.58 ± 0.06	0.19 ± 0.06
	ResNet-164	0.39 ± 0.06	0.15 ± 0.03	0.31 ± 0.05	0.15 ± 0.04	0.31 ± 0.05	0.13 ± 0.03
	PreResNet-164	0.29 ± 0.04	0.13 ± 0.03	0.28 ± 0.04	0.13 ± 0.03	0.28 ± 0.04	0.13 ± 0.03
	DenseNet-BC	0.23 ± 0.03	0.11 ± 0.02	0.24 ± 0.02	0.12 ± 0.02	0.24 ± 0.02	0.11 ± 0.02
NLL	ResNet-110	0.40 ± 0.02	0.16 ± 0.01	0.35 ± 0.02	0.15 ± 0.01	0.35 ± 0.02	0.14 ± 0.01
	ResNet-164	0.24 ± 0.02	0.11 ± 0.01	0.20 ± 0.01	0.10 ± 0.01	0.20 ± 0.01	0.10 ± 0.01
	PreResNet-164	0.23 ± 0.02	0.13 ± 0.02	0.19 ± 0.02	0.11 ± 0.01	0.19 ± 0.02	0.10 ± 0.01
	DenseNet-BC	0.17 ± 0.01	0.10 ± 0.01	0.14 ± 0.01	0.09 ± 0.01	0.14 ± 0.01	0.08 ± 0.01

Table 2: Calibration metrics for various CIFAR-10H models with P+L combinations. Even a small amount of labeled data (10 labels) reduces the calibration error of both the classifier and combination.

for making a correct prediction. More specifically, the confidence ratios r_m and r_h for machine and human labelers are defined as

$$r_m(x) = \frac{m_y^\theta(x)}{1 - m_y^\theta(x)} \qquad r_h(x) = \frac{\varphi_{h(x)y}}{1 - \varphi_{h(x)y}} \tag{4}$$

We note that unlike the machine classifier, the human does not directly output such confidences – rather, this quantity is estimated empirically through the human's confusion matrix.

If the model has a confidence ratio of $r_m(x) > 1$ (indicating that the model has confidence greater than 0.5 for the correct class), then the model is guaranteed to correctly label x. On the other hand, $r_m(x) > 1$ is not sufficient for the combination to correctly label x – instead, the model must be sufficiently confident in its prediction as well. The following theorem formalizes this notion by a lower bound on the accuracy of the combination in terms of the confidence ratios of the individual predictors. For binary classification tasks this lower bound is achieved, i.e. we have equality.

Theorem 1 (Combination Model Accuracy.)**.** *The accuracy of the P+L combination $c(x)$ is at least the probability that the confidence ratio for m exceeds the inverse confidence ratio for h.*

$$\mathbb{E}\left[\mathbb{1}\left(c(x) = y\right)\right] \geq p(r_m(x) \geq (r_h(x))^{-1}) \tag{5}$$

A detailed proof is provided in Appendix A. An analogous result holds for the combination of two probabilistic classifiers or two non-probabilistic classifiers. As our focus is on combining human predictions with model probabilities, we discuss these cases in Appendix A.

Theorem (1) is further illustrated in Figure 3 for a ResNet-164 classifier on CIFAR-10H (first row) and a VGG-19 classifier on ImageNet-16H (second row). For each row, we create a family of classification models where each model makes the same class-level predictions (and hence has the same error rate) but with different confidence ratio distributions. This is achieved by tempering the output probabilities of a base classifier via the map $m(X) \mapsto (m_1(X)^{1/T}, \ldots, m_k(X)^{1/T}) / \sum_{i=1}^K m_i(X)^{1/T}$ with temperature $T > 0$ [29, 45]. In the first column of each row, the solid curve plots the error rate of the combination of a fixed human with the various classification models. Despite each classification model having the same accuracy, the accuracy of the resulting human-model combination varies.

This behavior can be explained by Theorem (1), which tells us that the combination accuracy is driven not only by the human and machine accuracies but by their confidence ratios as well. At large temperatures (purple), the classifier becomes underconfident in its predictions, and the combination error rate approaches the human error rate. At small temperatures (green), the model becomes overconfident in its predictions, and the combination error rate approaches the model error rate. When the combination is fit by our P+L method (orange), the classifier is well-calibrated (reflected by its low ECE), and the resulting combination obtains a lower error rate than each of the under- and over-confident classifier combinations.

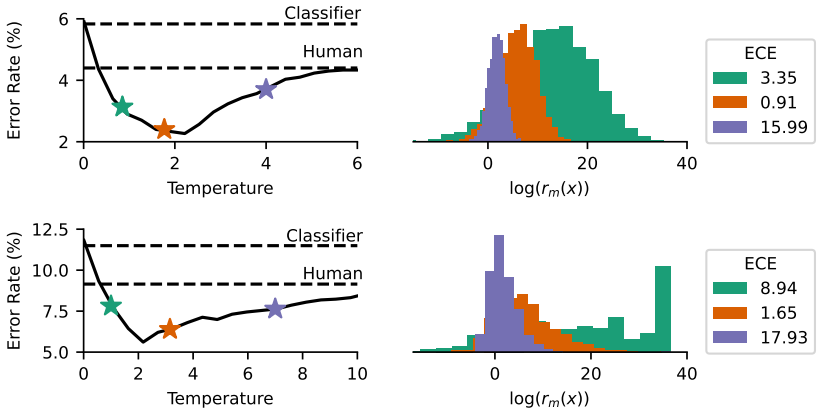

Figure 3: Left: Single human labeler combined with various classifiers of equal accuracy, resulting in combinations with varied accuracies. The orange point corresponds to the P+L combination. Right: First row: ResNet-164 on CIFAR-10H. Second row: VGG-19 on ImageNet-16H at noise level 80.

Relationship between combination and calibration/confusion error. We additionally quantify the estimation error in our P+L method, incurred by empirically estimating the human confusion matrix and calibration parameters. The result below shows we can upper bound our estimation error by the estimation error for the confusion matrix and the ℓ_1 marginal calibration error (MCE) [58].

Theorem 2 (Estimation Error Upper Bound.)**.** *Let $\eta(x, y) = |p(h(x)|y)p(y|m(x)) - m_y^\theta(x)\widehat{\varphi}_{h(x)y}|$ be the estimation error (up to normalizing constants), where $\widehat{\varphi}_{ij}$ represents an estimate of $p(h(x) = i|y = j)$. In expectation over random (x, y) pairs,*

$$\mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[\eta(X, Y)\right] \leq ||\varphi - \widehat{\varphi}||_1 + MCE(m^\theta) \tag{6}$$

(Proof in Appendix A). With a sufficient amount of labeled data, the confusion matrix error $||\varphi - \widehat{\varphi}||_1$ can be made arbitrarily small (i.e. the MAP confusion matrix estimate is unbiased). This is not necessarily the case for MCE(m^θ). However, if an asymptotically unbiased calibration method is used, this result guarantees that our posterior estimation error will converge to zero.

7 Limitations, Societal Impact, and Conclusions

Limitations. One limitation of our work is that our experiments only involve two datasets and both involve image classification. Thus, there is no guarantee that similar results (in terms of combined improvements) are achievable for other tasks, such as question-answering from text data or more general problem-solving tasks. Another potential limitation is our reliance on conditional independence in our approach. A reverse view of this is that both our theoretical and experimental results demonstrate that there is ample room for improving human and machine performance by combining their predictions, even without taking dependence into account.

Potential Societal Impacts. Combining human and machine predictions to improve overall classification accuracy has the potential for positive societal benefit, particularly for example in high-stakes applications such as medical image diagnosis and autonomous driving. However, there are also potential negative societal impacts. For example, if there is a lack of transparency in terms of how the system operates (e.g., how predictions are being combined to arrive at a final result), augmenting an individual's predictions with a machine's could have negative psychological consequences for the individual, such as decreasing trust, reducing individual autonomy, and eventual disengagement.

Conclusions. We investigated methods for combining predictions using instance-level confidence from a model and class-level information from a human. Across a variety of image classification experiments our proposed combination framework leads to systematic increases in accuracy over both the model and human alone while requiring few human labels. Supporting theory illustrates how combined human-model performance is affected by calibration properties of the model.

References

[1] Ece Kamar. Directions in hybrid intelligence: Complementing AI systems with human intelligence. In *Proceedings of IJCAI*, pages 4070–4073, 2016.

[2] Jennifer Wortman Vaughan. Making better use of the crowd: How crowdsourcing can advance machine learning research. *J. Mach. Learn. Res.*, 18(1):7026–7071, 2017.

[3] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1): 237–293, 2018.

[4] Mark O Riedl. Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, 1(1):33–36, 2019.

[5] Laura Trouille, Chris J Lintott, and Lucy F Fortson. Citizen science frontiers: Efficiency, engagement, and serendipitous discovery with human–machine systems. *Proceedings of the National Academy of Sciences*, 116(6):1902–1909, 2019.

[6] Matthew Johnson and Alonso Vera. No AI is an island: the case for teaming intelligence. *AI Magazine*, 40(1):16–28, 2019.

[7] Zahra Zahedi and Subbarao Kambhampati. Human-AI symbiosis: A survey of current approaches. *arXiv preprint arXiv:2103.09990*, 2021.

[8] Josef Kittler, Mohamad Hatef, Robert PW Duin, and Jiri Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[9] Thomas G Dietterich. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15. Springer, 2000.

[10] Ludmila I Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley & Sons, 2014.

[11] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.

[12] Lu Hong and Scott E Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46): 16385–16389, 2004.

[13] PJ Lamberson and Scott E Page. Optimal forecasting groups. *Management Science*, 58(4): 805–810, 2012.

[14] Robert Geirhos, Kristof Meding, and Felix A Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, 33, 2020.

[15] Amir Rosenfeld, Markus D Solbach, and John K Tsotsos. Totally looks like – how humans compare, compared to machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1961–1964, 2018.

[16] Thomas Serre. Deep learning: the good, the bad, and the ugly. *Annual Review of Vision Science*, 5:399–426, 2019.

[17] Ece Kamar, Severin Hacker, and Eric Horvitz. Combining human and machine intelligence in large-scale crowdsourcing. In *AAMAS*, volume 12, pages 467–474, 2012.

[18] Melanie R Beck, Claudia Scarlata, Lucy F Fortson, Chris J Lintott, BD Simmons, Melanie A Galloway, Kyle W Willett, Hugh Dickinson, Karen L Masters, Philip J Marshall, et al. Integrating human and machine intelligence in galaxy morphology classification tasks. *Monthly Notices of the Royal Astronomical Society*, 476(4):5516–5534, 2018.

[19] Yashesh Gaur, Florian Metze, Yajie Miao, and Jeffrey P Bigham. Using keyword spotting to help humans correct captioning faster. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[20] P Jonathon Phillips, Amy N Yates, Ying Hu, Carina A Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, et al. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018.

[21] Nicholas Bien, Pranav Rajpurkar, Robyn L Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N Patel, Kristen W Yeom, Katie Shpanskaya, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS Medicine*, 15(11):e1002699, 2018.

[22] Bhavik N Patel, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, et al. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ Digital Medicine*, 2 (1):1–10, 2019.

[23] Pranav Rajpurkar, Chloe O'Connell, Amit Schechter, Nishit Asnani, Jason Li, Amirhossein Kiani, Robyn L Ball, Marc Mendelson, Gary Maartens, Daniël J van Hoving, et al. Chexaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with hiv. *NPJ Digital Medicine*, 3(1):1–8, 2020.

[24] Gideon Keren. Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, 77(3):217–273, 1991.

[25] Daniel Kahneman and Amos Tversky. On the reality of cognitive illusions. *Psychological Review*, pages 582–591, 1996.

[26] Joshua Klayman, Jack B Soll, Claudia Gonzalez-Vallejo, and Sema Barlas. Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, 79(3):216–247, 1999.

[27] 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.

[28] Anonymous. Imagenet-16H: a large-scale behavioral data set of noisy image classifications. To be publicly released on Open Science Foundation, 2021.

[29] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[30] Christian Genest, James V Zidek, et al. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135, 1986.

[31] Robert A Jacobs. Methods for combining experts' probability assessments. *Neural Computation*, 7(5):867–888, 1995.

[32] Ludmila I Kuncheva. On the optimality of naive Bayes with dependent binary features. *Pattern Recognition Letters*, 27(7):830–837, 2006.

[33] Lei Xu, Adam Krzyzak, and Ching Y Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(3):418–435, 1992.

[34] Hyun-Chul Kim and Zoubin Ghahramani. Bayesian classifier combination. In *Artificial Intelligence and Statistics*, pages 619–627. PMLR, 2012.

[35] Clintin P Davis-Stober, David V Budescu, Stephen B Broomell, and Jason Dana. The composition of optimally wise crowds. *Decision Analysis*, 12(3):130–143, 2015.

[36] Michael D Lee and Megan N Lee. The relationship between crowd majority and accuracy for binary decisions. *Judgment and Decision Making*, 12(4):328, 2017.

[37] Darryl E Wright, Chris J Lintott, Stephen J Smartt, Ken W Smith, Lucy Fortson, Laura Trouille, Campbell R Allen, Melanie Beck, Mark C Bouslog, Amy Boyer, et al. A transient search using combined human and machine classifications. *Monthly Notices of the Royal Astronomical Society*, 472(2):1315–1323, 2017.

[38] Olga Russakovsky, Li-Jia Li, and Li Fei-Fei. Best of both worlds: human-machine collaboration for object annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2121–2131, 2015.

[39] David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *arXiv preprint arXiv:1711.06664*, 2017.

[40] Maithra Raghu, Katy Blumer, Greg Corrado, Jon M. Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. *CoRR*, abs/1903.12220, 2019. URL `http://arxiv.org/abs/1903.12220`.

[41] Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, pages 7076–7087. PMLR, 2020.

[42] Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. *arXiv preprint arXiv:2005.00582*, 2020.

[43] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Dan Weld. Is the most accurate AI the best teammate? optimizing AI for teamwork. In *AAAI 2021*, February 2021.

[44] Nastaran Okati, Abir De, and Manuel Gomez-Rodriguez. Differentiable learning under triage. *arXiv preprint arXiv:2103.08902*, 2021.

[45] Jize Zhang, Bhavya Kailkhura, and T Yong-Jin Han. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International Conference on Machine Learning*, pages 11117–11128. PMLR, 2020.

[46] Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Hao Song, Peter Flach, et al. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. *arXiv preprint arXiv:1910.12656*, 2019.

[47] Matthew D Hoffman and Andrew Gelman. The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[48] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[49] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

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

[51] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[52] 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.

[53] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[54] 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.

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

[56] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[57] Kanil Patel, William Beluch, Bin Yang, Michael Pfeiffer, and Dan Zhang. Multi-class uncertainty calibration via mutual information maximization-based binning. *arXiv preprint arXiv:2006.13092*, 2020.

[58] Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. *CoRR*, abs/1909.10155, 2019. URL `http://arxiv.org/abs/1909.10155`.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 7.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 7.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Included in the Supplemental Results.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Anonymized URL provided in introduction.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5. Additionally, all code is made available.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 5.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Reported in the Supplemental Results, see Section E.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] Reported in the Supplemental Results, see Section E.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] URL to code assets provided in the Introduction (Section 1)
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] All data used is or will be publicly available. The CIFAR-10H dataset has been available publicly for a number of years and the ImageNet-16H dataset will be publicly released later this year.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] As far as we are aware, there is no personally identifiable information or offensive content in CIFAR-10H dataset or the subset of ImageNet classes (the ImageNet-16H dataset) that we are using. The class labels are for objects and animals in both datasets (this of course doesn't guarantee. that there is no personally identifiable information or offensive content, but makes it significantly less likely). The 10 classes in CIFAR-10 are 'airplane', 'automobile', 'bird', 'cat', 'deer', 'dog', 'frog', 'horse', 'ship', 'truck'. The 16 ImageNet classes in ImageNet-16H are: 'airplane', 'bear', 'bicycle', 'bird', 'boat', 'bottle', 'car', 'cat', 'chair', 'clock', 'dog', 'elephant', 'keyboard', 'knife', 'oven', 'truck'.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We used pre-existing annotated datasets in this paper. Details of the original crowdsourcing experiments can be found in the cited references.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]