On the Informativeness of Supervision Signals

Abstract

Supervised learning typically focuses on learning transferable representations from training examples annotated by humans. While rich annotations (like soft labels) carry more information than sparse annotations (like hard labels), they are also more expensive to collect. We use information theory to compare how a number of commonly used supervision signals contribute to representation-learning performance, as well as how their capacity is affected by factors such as the number of labels, classes, dimensions, and noise. Our framework provides theoretical justification for using hard labels in the big-data regime, but richer supervision signals for few-shot learning and out-of-distribution generalization. We validate these results empirically in a series of experiments with over 1 million crowdsourced image annotations and conduct a cost-benefit analysis to establish a tradeoff curve that enables users to optimize the cost of supervising representation learning on their own datasets.

1 INTRODUCTION

Modern machine learning relies heavily on using large amounts of labeled data, and those labels typically come from annotations generated by humans. This raises an important question: how can we most efficiently use human annotations to create objective functions for machine learning systems? This is not just a matter of designing good interfaces and algorithms for collecting annotations—it involves a subtle interplay between the choices we make about the supervision signals we use to train our models and the difficulty of collecting the relevant annotations. For example, soft labels for images (indicating uncertainty via a distribution over classes) are more expensive to collect than hard labels (indicating a single class), but are also potentially more informative to the learner [Peterson et al., 2019, Sucholutsky and Schonlau, 2021b,a, Collins et al., 2022]. Making good choices about what questions to ask humans about our data requires understanding the informativeness of different supervision signals.

In this paper, we explore this question for the case of *representation learning*, where the aim is to learn useful latent embeddings of input stimuli. Generally, training a neural network means learning successive layers of representations that will be used to perform some sort of task (e.g., classification). The key decision in implementing a representation-learning framework often revolves around designing a supervision signal by quantifying the similarity between stimuli. Significant work has gone into the design of supervision signals for deep representation learning resulting in contrastive objectives [Chen et al., 2020, Khosla et al., 2020], classification objectives [Huh et al., 2016, Ridnik et al., 2021], reconstruction objectives [Devlin et al., 2018, Kingma and Welling, 2013], and many others [Guo et al., 2019]. For example, recent work has shown that models trained with hard labels on classification tasks can approximate the structure of human latent representations at a fraction of the cost of exhaustively collecting the pairwise-similarity judgments required for soft contrastive learning [Peterson et al., 2018, Marjieh et al., 2022]. Thus, in many cases, multiple representation learning objectives could work, and it is not clear when one objective should be preferred.

Our goal in this paper is to compare the informativeness of various supervision signals (i.e. types of annotations) to **empower researchers to optimize data annotation for their supervised representation learning tasks** (see Figure 1). We **develop an information-theoretic framework (Section 3)** for analyzing supervision signals and use it to compare two popular supervision signals from the classification literature—hard labels and soft labels. We quantify their relative (representational) information content by comparing them to similarity triplets (i.e., queries of the form "Is x more similar to y than to z?"), a supervision signal used in contrastive learning and cognitive science [Jamieson

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

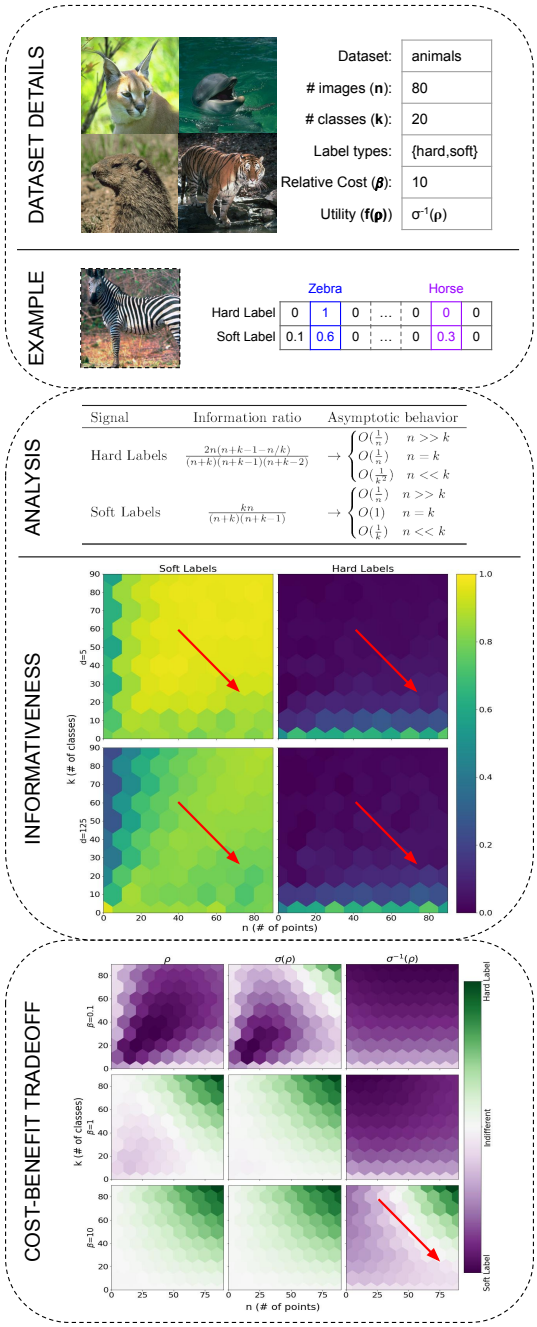

Figure 1: An example workflow for using our framework to decide what label type(s) to collect when annotating a dataset. **Top**: User specifies the task. **Middle**: Information-theoretic characterization of the task (e.g., many classes, but few examples). Heatmaps show correlation (ρ) of ground-truth similarities with pairwise similarities recovered from simulated labels when varying latent dimensionality (d), number of points (n), and number of classes (k). **Bottom**: Cost-benefit analysis of signal type based on subjective utility ($u(\rho)$), cost (β), number of points (n), and number of classes (k). Red arrows point to cells corresponding to the user-specified dataset.

and Nowak, 2011, Hoffer and Ailon, 2015]. We relate the information each signal provides to three common features of machine learning datasets: number of labels, number of classes, and dimensionality, and find that while both hard and soft labels provide information about hidden representations, their responses to these three variables are very different. **Simulations confirm these results (Section 4)** , showing how the marginal information provided by each label translates into better representation learning performance. A **cost-benefit analysis (Section 5)** comparing soft and hard labels shows there are meaningful differences between them: soft labels are expensive to collect but provide a considerable amount of representational information while hard labels are cheap but fairly uninformative (Figure 1). To close this gap, we consider several types of sparse representations that enable selective interpolation between the soft and hard label regimes (Figures 2, 3). We extend our analysis with these sparse supervision signals to **establish a tradeoff curve allowing users to optimize the cost of labeling their datasets for supervised representation learning (Section 6)**. Finally, we confirm the theoretical and simulation results by **human experiments (Section 7)** with crowdsourced similarity judgments and various types of soft and hard labels collected for `CIFAR-10` [Krizhevsky et al., 2009].

2 BACKGROUND

Throughout this paper, we discuss a variety of supervision signals. In this section, we identify the practical settings these signals correspond to and summarize related work.

Pairwise similarity judgments: For every pair of points in the dataset labelers are asked to rate the similarity on a fixed, bounded scale (e.g., a Likert scale). This signal has a long history of being employed by cognitive scientists to learn about (hidden) human representations of stimuli, typically by using an embedding method like multi-dimensional scaling (MDS) [Shepard, 1980], and more recently can be found in state-of-the-art contrastive learning objectives [Chen et al., 2020, Khosla et al., 2020].

Triplet similarity judgments: For every set of three points x, y, z in the dataset, labelers are asked to respond to queries of the form "Is x closer to y than to z?" This non-metric signal has also been used as a method for learning human representations, often by applying embedding techniques like non-metric MDS [Jamieson and Nowak, 2011, Agarwal et al., 2007, Davenport, 2013]; it is considered to be a more accurate alternative to pairwise similarity judgments as it is easier to compare two lengths than to provide consistent judgments over un-normalized scales. This signal is used in machine learning (typically with the triplet loss) and cognitive science [Hoffer and Ailon, 2015, Hermans et al., 2017, Hebart et al., 2020, Roads and Love, 2021].

Hard labels: For each point, labelers pick a single, most relevant class out of the fixed set of all classes in the dataset. This signal is typically used for classification, though it can also be used for representation learning via pre-training [Huh et al., 2016, Ridnik et al., 2021].

Soft labels: For each point, labelers assign proximity or probability to each of the fixed set of all classes in the dataset. This signal is also used for classification and representation learning via pre-training, and can be more effective than hard labels, particularly in settings with small data [Xie et al., 2020, Sucholutsky and Schonlau, 2021a, Sucholutskv et al., 2021, Liu et al., 2021, Malaviya et al., 2022].

Top-class soft labels: The researcher picks a subset of size $\hat{k}$ of the classes in the dataset that maximize mutual information (estimated based on the already collected subset). For each point, labelers assign proximity or probability to each of the fixed set of classes in the subset. The number of classes can be reduced by simply taking an arbitrary subset, or more systematically via methods like label coarsening [Hanneman and Lavie, 2011].

Sparse soft labels: For each point, labelers assign proximity or probability to each of the exactly $\hat{k}$ most relevant classes out of the fixed set of all classes in the dataset [Collins et al., 2022]. In addition to the connections discussed above for the other soft label variants, this particular variant is also connected to top-k classification [Lapin et al., 2017] and soft vector quantization [Seo and Obermayer, 2003].

3 RELATIVE INFORMATIVENESS OF SUPERVISION SIGNALS

Setup: We consider the scenario where a researcher collects an initial set of labels for a small subset of their dataset and wants to use it to optimize the labelling of their entire dataset (Figure 1, top). The researcher has a number of options for what kind of labels to collect and wishes to maximize representation learning performance while minimizing labelling cost. We formalize representation learning as the process of recovering a hidden (low-dimensional) latent structure from a set of (high-dimensional) stimuli (e.g. images). In particular, we focus on the non-metric setting where we want to recover the correct rank order of pairwise distances between all latent vectors. Our goal is to determine which supervision signal is most effective (in terms of both performance and cost) for representation learning.

Problem Definition: Consider a set of stimuli $\{x_i\}_{i=1}^n \in \mathbb{R}^d$ with some associated latent representations $\{z_i\}_{i=1}^n \in \mathbb{R}^h$. The distance between each pair of latent vectors induces a relational order over latent pairs and our goal is to find a function $f : \mathbb{R}^d \to \mathbb{R}^h$ (where typically $h << d$) such that it preserves this relational order, that is, $||f(x_i) - f(x_j)|| \leq ||f(x_i) - f(x_k)||$ iff $||z_i - z_j|| \leq ||z_i - z_k||$. Crucially, the latent vectors are accessible only implicitly via different supervision signals (or queries) such as hard and soft labels. We operationalize the informativeness of different supervision signals as the number of relational constraints that a naive learner can recover based on them (i.e., a learner that attempts to follow the signals as is without applying other geometric constraints such as triangle inequalities).

Triplet Constraints: Conceptually, when training a neural network for classification, providing a label for a point roughly corresponds to requiring that the network weights should be updated such that the embedding of this point will be closer to one class than to other classes. For this analysis, we assume that each class can be represented by its centroid (e.g., each class is unimodal), and so classification labels provide information about proximity of latent vectors to these centroids. When training with batches, providing labels for a batch additionally corresponds to requiring that the centroid of each class be closer to its associated set of embeddings than to the other embeddings in the batch. In both cases, neural networks are optimizing constraints of the form "x is closer to y than to z," which we call "triplet constraints". We now formalize this concept to use it as a measure of information content in labels.

Suppose we have a system with n labels, k classes with centroids $C_1, ..., C_k$, and stimuli $\{x_i\}_{i=1}^n$ with latent representations $\{z_i\}_{i=1}^n$. A triplet constraint is an inequality of the form $||z_a - z_b|| < ||z_a - z_c||$. This can be rewritten as the query $r_{i,j,k} = \left\{x \in \mathbb{R}^d : \|f(x_j) - f(x_i)\| < \|f(x_k) - f(x_i)\|\right\}$, and each such query provides at most one bit of information [Jamieson and Nowak, 2011]. For any set of three stimuli, there are three unique queries: $r_{i,j,k}, r_{j,i,k}, r_{k,i,j}$. Thus, the total number of unique queries for n stimuli is $3\binom{n}{3}$. However, in the case of hard and soft labels, we make queries not only in terms of the n objects but also in terms of the k class centroids. In other words, we are seeking to recover embeddings not only for the n points of interest, but also k additional reference embeddings. As a result, the total number of unique queries in these cases is $3\binom{n+k}{3}$.

Hard Labels: We define the hard label for stimulus x as a vector l of length k, such that $l_i = 1$ if $i = \arg\min_j(||f(x) - f(C_j)||)$ and 0 otherwise. We can now extract two types of triplet queries. The first is a triplet consisting of a class centroid C_i, a stimulus x_P that is a "positive" example of this class, and a negative example stimulus x_N. The query has the form $||f(x_P) - f(C_i)|| < ||f(x_N) - f(C_i)||$. The second is a triplet consisting of a stimulus x_i, the class centroid that is closest to it (C_P), and another class centroid further away (C_N). This query has the form $||f(C_P) - f(x_i)|| < ||f(C_N) - f(x_i)||$. If the hard labels are distributed evenly between k classes, then on average there are n/k stimuli per class. Then n hard labels give us $n(k-1)$ constraints of the second type and $k(n/k)(n - n/k) = n^2(1 - 1/k)$ of the first type—a total of $T_H(n, k) = n(k-1) + n^2(1 - 1/k)$ constraints.

Soft Labels: To produce a probability distribution over classes, neural networks often have a softmax activation function after the output layer [Bridle, 1989, Martins and Astudillo, 2016, Krizhevsky et al., 2017]. Accordingly, we define the soft label for a stimulus x as a vector l of length k, such that $l_i = \frac{e^{-||f(x)-f(C_i)||}}{\sum_j e^{-||f(x)-f(C_j)||}}$. There are again two types of triplet queries that we can extract from soft labels. The first is a triplet of the form $||f(x_P) - f(C_i)|| < ||f(x_N) - f(C_i)||$ where C_i is the centroid of class i and x_P, x_N are two training set points with corresponding soft labels l^P, l^N such that $l_i^P > l_i^N$. The second is a triplet consisting of the form $||f(x) - f(C_i)|| < ||f(x) - f(C_j)||$ where x is a training set point corresponding to label l and C_i, C_j are the centroids of classes i, j such that $l_i > l_j$. Our n soft labels thus give us $nk(k-1)/2$ constraints of the second type and $kn(n-1)/2$ of the first—a total of $T_S(n,k) = kn(k+n-2)/2$ triplet constraints.

Information Ratio: While we now have a measure of how much information each label provides, it is unclear how much information is actually needed to recover human-aligned representations for all the objects. Intuitively, we would expect that more information is required when more objects are being embedded (i.e. when $n+k$ increases). We can normalize our results from the previous section to account for this by taking the ratio of the number of constraints we can recover from a set of labels to the total number of possible queries (i.e. $IR(n,k) = \frac{T(n,k)}{3\binom{n+k}{3}}$). This "information ratio" may be a proxy for how much information we are recovering about the latent representations. We present the information ratios for hard and soft labels in the "Analysis" portion of Figure 1 along with their asymptotic behavior in three regimes: the many-shot case (where there are many more points than classes), the one-shot case (where there is one point per class; [Fei-Fei et al., 2006]), and the less-than-one-shot case (where there are fewer points than classes; [Sucholutsky and Schonlau, 2021a]).

Our results predict three scaling phases for soft labels and two scaling phases for hard labels. In particular, hard labels and soft labels are predicted to have similar asymptotic behavior in the many-shot case which may explain why pre-training on very large datasets using hard-label classification is effective (e.g. [Huh et al., 2016, Ridnik et al., 2021]). However, soft labels are predicted to have much better representation learning performance in the one-shot and less-than-one-shot cases. Notably, the results predict that in the one-shot case, the quality of representations learned from soft labels should not degrade (as it does in every other case) when the number of points and classes increases.

Representation Learning as Communication: Our representation learning setup can be seen as a communication problem where we want to recover a fixed number of bits about hidden representations of a black-box model. The model can be queried via multiple channels, each of which has its own encoder and corresponds to a different choice of supervision signal. As established by Jamieson and Nowak [2011], each triplet query is a single bit. In information theory, the efficiency, also known as the normalized entropy, of each channel is defined as $\eta(X) = \frac{H(X)}{H_{max}(X)}$ where H_{max} is max entropy (i.e. the total number of bits) and H is entropy (i.e. the number of bits remaining with unknown states after transmission). We defined information ratio as the ratio of the number of bits recovered to the total number of bits, and thus a stochastic version of information ratio (with data sampled from a distribution instead of being fixed) can be defined as $IR(X) = \frac{H_{max}(X)-H(X)}{H_{max}(x)} = 1 - \eta(X)$.

Signal-to-Noise Ratio: Our triplet analysis implicitly assumed that the noise distribution was the same between each type of label and could thus be ignored when comparing their relative informativeness. However, in practice, eliciting different kinds of annotations may be associated with different levels of noise due to changes in difficulty, user interface, participant pools, etc. Within our information-theoretic framework, noise can be viewed as the probability ϵ_s of a bit flip (i.e. that we get the incorrect response to a query of the form "Is x closer to y than to z?") during communication over channel s . If a set of n labels (with k classes) of type s provides $T_s(n,k)$ triplet constraints under our framework in the noise-free case, then the number of constraints in the noisy case is just $(1-\epsilon_s)T_s(n,k)$ where ϵ_s is the bit flip rate for labels of type s.

4 SIMULATIONS

Our analysis predicts soft labels should generally lead to better performance than hard labels, particularly when there are few labels and many classes. However, we still need to understand how information ratios actually translate to representation learning performance. We conducted simulations to see the effect of four variables (and their interactions) on performance: label type (soft or hard), number of points (n), number of classes (k), and latent dimension (d). We consider values of n and k in the range of $[3, 90]$, and $d \in \{5, 25, 125\}$. For each combination of n, k, d we sample a total of n points from Gaussians centered at k random locations $C_1, ..., C_k \in \mathbb{R}^d$. We computed hard and soft labels for these points using the equations defined above and then mine all triplet constraints of both types from both sets of labels. We apply Generalized Nonmetric Multi-Dimensional Scaling (GNMDS; Davenport [2013]) to both sets of queries to find embeddings that best fit the respective triplet constraints. The Gram matrix outputted by GNMDS can be interpreted as predicted (unnormalized) pairwise cosine similarities between all $n+k$ points and centroids.

To understand how much information we recover from each of these two sets of queries, we construct a matrix of the true pairwise cosine similarities for the set of all $n+k$ points and

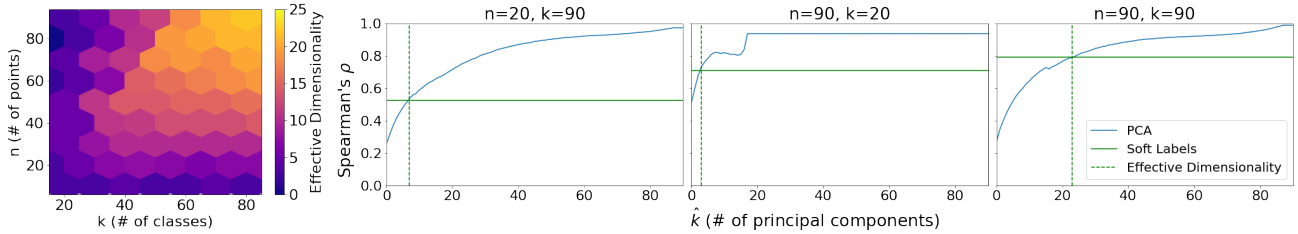

Figure 2: **Left**: Effective dimensionality of soft labels at different combinations of n and k. **Right**: Three examples of PCA curves used to compute effective dimensionality of soft labels for every combination of n and k.

class centroids and compute the Spearman rank correlation (ρ) between the upper triangle of the Gram matrices and the ground truth matrix. Thus, a higher ρ corresponds to better recovery of the underlying latent representations. We visualize the simulation results in Figure 1. The results confirm the theoretical findings from the previous section. Specifically, the three phases for soft labels and two phases for hard labels match our analytical results, and **a higher information ratio translates into better performance**.

5 COST-BENEFIT TRADEOFFS

We can now construct cost-benefit tradeoff curves to **determine when a user would prefer to use one signal over the other**. Suppose we define ρ as above, and subjective utility as $U(\rho)$. This utility function can take many forms (e.g., $U(\rho) = b\rho$ or $b\sigma(\rho)$ where σ is the sigmoid function and $b > 0$). If we assume that the cost of collecting a soft label over k classes is about k times more expensive than collecting a hard label, we can define the subjective loss function as $L_s = C(s) - U(\rho)$, where $C(s) = cn$ if $s \in S_{hard}$ and cnk if $s \in S_{soft}$. This is equivalent to optimizing $\hat{L} = \frac{c}{b}\hat{c}(s) - \hat{u}(\rho)$ which we can re-parametrize to a form reminiscent of the information bottleneck [Tishby et al., 2000]: $\hat{L} = \beta\hat{c}(s) - \hat{u}(\rho)$. Since we have shown that the information ratio, which we define as $\hat{\rho}$, can provide us with an estimate for ρ, we can also replace $U(\rho)$ by $U(\hat{\rho})$.

We investigate cost-benefit tradeoffs by varying $(\beta, \hat{u})$, showing the results for several combinations in Figure 1. While the results depend greatly on choice of $\hat{u}$, a few regularities emerge. First, in all cases, regardless of β, when the number of classes (k) or the number of points (n) is low, soft labels are roughly as preferred as, or more preferred, than hard labels. Second, when there is an emphasis on cost (i.e. high β), hard labels become preferable as n and k both increase, but when the emphasis is on performance (i.e. low β), soft labels remain preferable as n and k increase.

6 LABEL OPTIMIZATION

Effective Dimensionality: While soft labels appear to be more informative than hard labels, they are not necessarily

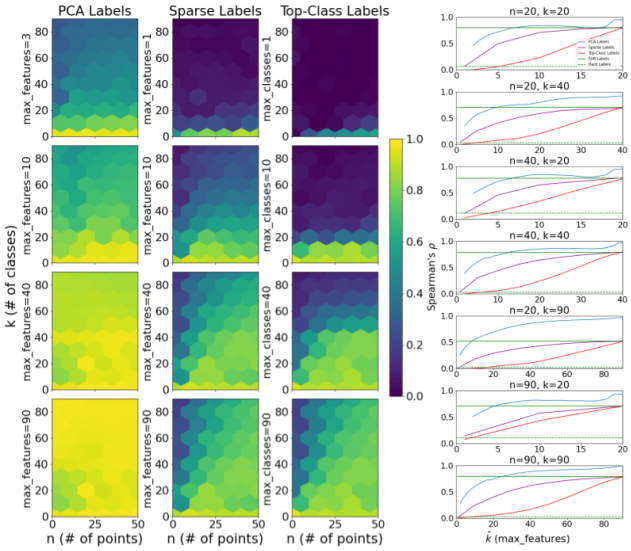

Figure 3: Label sparsity. **Left**: Spearman rank correlation (ρ) of pairwise similarities recovered from PCA labels, sparse labels, and top-class labels when varying the maximum number of features/classes in the labels ($\hat{k}$), number of points in the dataset (n), and number of classes in the dataset (k). **Right**: Comparison of sparsity curves for PCA labels (blue), sparse labels (purple), and top-class labels (red) for several combinations of n and k. Straight lines correspond to soft label (solid green) and hard label (dashed green) performance.

an optimal encoding for efficiently communicating information about representations. In order to study how an optimal encoding might perform, we run principal component analysis (PCA) on our simulated datasets and observe the effects of varying the number of PCs that are retained ($\hat{k}$). The resulting performance curves provide an approximate upper bound on how well any set of vectors of length $\hat{k}$ (e.g. a set of soft labels) can communicate information about representations. We can also use these curves to understand how efficient soft labels are at this task. Specifically, we define the "effective dimensionality" of a set of soft labels as the minimum number of PCs ($\hat{k}$) necessary to achieve the same representation learning performance as when using the soft labels in the way we described in the previous section. In Figure 2, we show the effective dimensionality of soft labels for different combinations of n and k along with several examples of the PCA curves. We found a strong

positive correlation between information ratio and effective dimensionality ($r = 0.734, p < 10^{-15}$) providing further evidence information ratio is a useful metric for evaluating representation learning signals and predicting performance.

Label Sparsity: Since the effective dimensionality of soft labels appears to generally be much lower than the number of components of each soft label (k), this suggests that soft labels are an inefficient encoding for representational information, potentially due to redundancy. We now investigate two methods for remedying this inefficiency by introducing sparsity into soft labels in a disciplined manner.

The first method, which we call "top-class soft labels", introduces sparsity globally by ignoring classes that are the least informative across the entire dataset. Formally, we construct a matrix X where the i-th row correspond to the i-th soft label and the j-th column corresponds to the j-th class, and estimate the mutual information between each column and the ground-truth similarity matrix. We then keep only the top $\hat{k}$ most informative columns and set the rest to 0.

The second method, which we call "sparse soft labels", is to introduce sparsity locally by ignoring classes that are least informative for each point. Formally, we again construct a matrix X as above, but now for each row, we individually keep only the $\hat{k}$ largest components and set the rest to 0. These two methods provide a way to reduce the cost of collecting soft labels, while retaining much of the information.

In Figure 3, we visualize representation learning performance when using PCA, sparse labels, or top-class labels at different levels of sparsity. While sparse labels consistently outperform top-class labels (which is to be expected since sparse labels provide finer control over how sparsity is induced), we note that there is also a gap between PCA and sparse labels. This again suggests that it may be possible to design more effective labeling methods than soft labels, potentially by defining a search space over possible classes and applying a procedure like PCA to optimize over it. We leave this as a promising direction for future work.

Optimizing Label Collection: Since top-class and sparse labels provide a way to selectively interpolate between the soft label and hard label regime, we can now use sparsity to optimize the cost-benefit tradeoff curves discussed above beyond the binary preference optimization shown at the bottom of Figure 1. We use the same loss functions as above with several combinations of cost parameter β and utility function μ, but we now apply them to top-class and sparse labels at various levels of sparsity ($\hat{k}$). We visualize a number of these cost-benefit tradeoff curves in the Supplement. By picking the $\hat{k}$ that corresponds to minimal loss, we can now optimize our label collection to minimize cost while maximizing performance. The results suggest that **using sparse labels and picking the right level of sparsity can often provide big gains** as opposed to using either hard labels or regular (dense) soft labels.

7 EXPERIMENTS

In this section, we investigate how our theory applies to real soft labels crowdsourced from human annotators. First, we present several methods for collecting soft labels—including new sets of annotations that we crowdsource for this study—and place them into the framework described above. Second, we use a large dataset of similarity judgements to assess the informativeness of the different label types. Finally, we assess how the inductive biases conferred by different label types affect the classification performance of a range of convolutional neural networks (CNNs).

7.1 EXPERIMENTAL SETUP

We consider a range of supervision signals over the *testing* subset of `CIFAR-10` images [Krizhevsky et al., 2009], including hard labels, smoothed hard labels, soft labels, and similarity judgments. Each label type represents a different supervision signal presented in Section 2. Each type of soft label was collected using different experimental interfaces, details of which are in the Supplement.

CIFAR-10H: The `CIFAR-10H` labels, originally collected by Peterson et al. [2019], Battleday et al. [2020], are derived by averaging over 500,000 crowdsourced hard labels (roughly 50 per image). These are then normalized at the image level to return probability distributions.

CIFAR-10DS: The `CIFAR-10DS` labels, made up of over 500,000 judgments we crowdsourced for this study, are *dense* soft labels (i.e., assigned over all classes in `CIFAR-10`). Annotators provided numerical judgments on a 0 (not at all) to 1 (completely) scale using sliders depending on how well each category described a certain image.

CIFAR-10S: The `CIFAR-10S` labels from Collins et al. are *sparse soft labels*. Annotators provided around 20,000 judgments about the likelihood of the top two categories for each image and any categories which they believed were definitely wrong (referred to as a "clamp", and assigned zero probability). As the authors only collected such labels over 1,000 examples from the test set, we in-fill the remaining 9,000 with either: 1) hard labels (`CIFAR-10S+hard`) or 2) simulated top-2 soft labels (`CIFAR-10S+dense`).

CIFAR-10LS: We derive two more sets of soft labels by applying label smoothing (LS) to the `CIFAR-10` hard labels. Unlike human-derived soft labels, the "softness" here is applied uniformly and independently of the associated image. This allows us to control for the previously observed regularization effects of label smoothing [Müller et al., 2019]. We pick the smoothing rate, ϵ, to roughly match the distributions of our crowdsourced labels. The "low" level, $\epsilon \approx 0.05$, is the average probability mass per soft label for the 9 non-maximal categories in the `CIFAR-10H` dataset. The "high" level, $\epsilon \approx 0.2$, matches the `CIFAR-10DS` dataset.

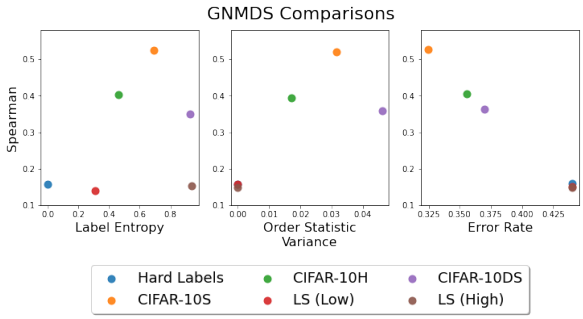

Figure 4: Correspondence between various metrics and Spearman rank correlation for each label variant.

Similarity Judgments: We elicited a total of 49,500 pairwise similarity judgments over two subsets each consisting of 100 `CIFAR-10` test set images by having human annotators rate the similarity between unlabeled image pairs on a Likert-scale ranging from 0 (completely dissimilar) to 6 (completely similar). The images were deliberately chosen to be ambiguous about class (see Supplement).

7.2 LABEL INFORMATIVENESS

GNMDS Results: We first repeat our GNMDS analyses from the theory and simulation sections on all the `CIFAR-10` label variants for 200 high-entropy images (see Supplement). We use GNMDS to get triplet-respecting embeddings for each label type and then compute several metrics. As before, we compute Spearman correlations (ρ) between the `CIFAR-10` GNMDS and elicited similarity judgments. Our previous analyses assumed a consistent signal-to-noise ratio, or error rate, across all the label types. However, with our set of CIFAR-10 label variants coming from different participant pools and elicitation pipelines, it is likely that error rates will vary between label types. We approximate error rates of each label type by counting the proportion of triplet queries derived from the GNMDS embedding whose binary responses do not match the responses from the corresponding triplets computed from the human similarity judgments (i.e. the bit flip rate). Additionally, we compute label entropy and the variance of the first order statistic (i.e. variance of the probability mass assigned to the class with the highest assigned probability) for each label variant.

We visualize these results in Figure 4. We find that representation learning performance (measured by Spearman correlation) generally increases as softness (measured by entropy and variance) increases, but does not increase when smoothing hard labels. However, **we find there is a "sweet spot" for softness** after which performance begins to decrease. We also find the expected linear relationship between error rate and performance that our framework predicts. We hypothesize that the U-shaped relationship between softness and performance may partially be caused by increasing error rates, and partially by a resonance effect (see Supplement).

Models: We train diverse image classifiers on each of our supervision sets. These models have distinct architectural features and reflect seminal developments in the progression of natural-image classification (see Supplement).

Cross-Label Performance: We first assess how training image classifiers on one set of soft labels impacts their validation-set performance when testing on other label types. Our primary measure of performance is cross-entropy between the models' predictive distributions and soft labels. Consistent with the GNMDS experiments, we observe a U-shaped relationship between label entropy and model performance for nearly all soft-label types (Figure 5, top row). The one exception is when testing on `CIFAR-10DS` labels, where training on higher levels of softness is preferred. These results suggest that, for image classification, ***sparse soft labels***—of the kind collected in Collins et al. [2022] or formed via averaging [Peterson et al., 2019]—best capture the representational information expressed across different soft label sets. We also find that this relationship is preserved across different model architectures, with the `Shake Shake` architecture [Gastaldi, 2017] performing best across all label sets.

7.3 ZERO-SHOT GENERALIZATION

To further test how well image classifiers extract representational information from soft labels, we assess zero-shot generalization performance on increasingly out-of-distribution image sets (Figure 5, bottom row; see Supplement). We find that for near-distribution datasets (`CIFAR-10 50K` [Krizhevsky et al., 2009]; `CIFAR10.1v6,v4` [Recht et al., 2018]), classifier performance follows the U-shaped relationship described above. However, for far-distribution datasets (`CINIC` [Darlow et al., 2018], `ImageNet-Far` [Peterson et al., 2019]) a different pattern emerges. As found in Peterson et al. [2019], **soft labels increasingly outperform hard labels as distribution shift increases**. The improved relative performance of the `CIFAR-10DS`-trained networks supports this – although these labels are the most noisy, they also contain the richest supervision signal. It may be that more expressive classifiers are needed to fully capitalize on this richness [Battleday et al., 2020, Singh et al., 2020].

7.4 LEARNING FROM SMALL DATA

We also empirically investigate our theoretical observation that softness is beneficial in the small data regime. Table 1 presents model performance in this regime (80 training examples per class). We find that soft labels outperform hard labels, and, as expected, we see an improvement in the relative performance of models trained on dense soft labels.

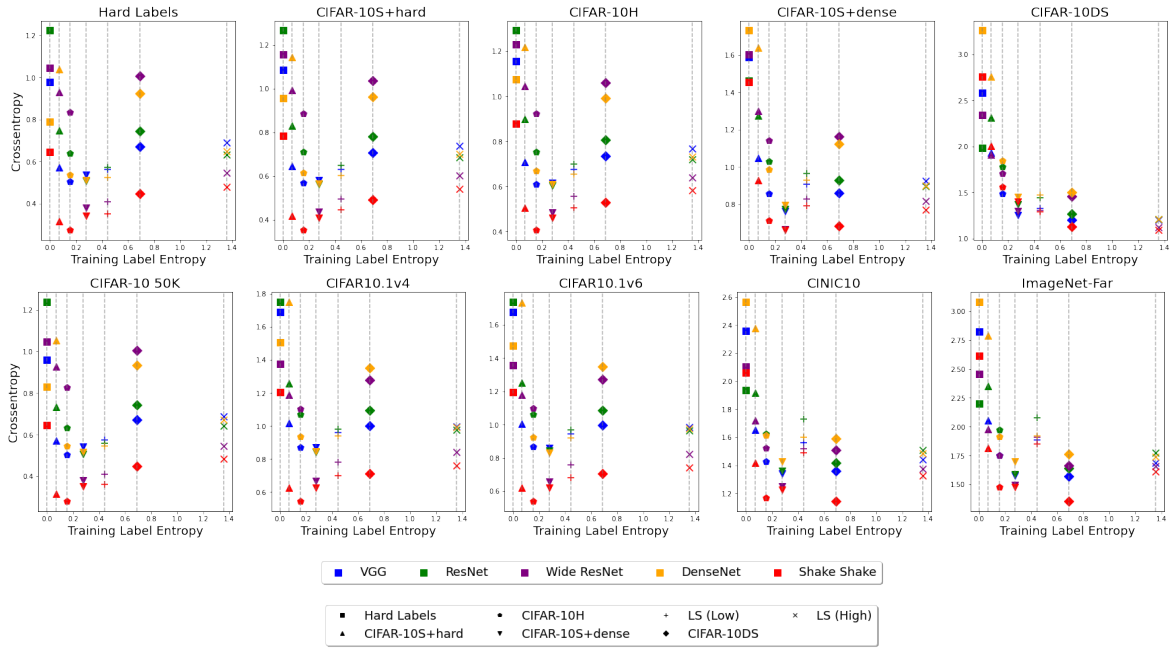

Figure 5: **Top:** Model performance on different label types at test time. **Bottom:** Generalization performance under increasing distributional shift. Each point represents the average score for a single model architecture (specified by color), trained on a particular label type (indicated via shape). Vertical lines represent points for a given label type.

8 DISCUSSION

In this paper, we offer a principled set of theoretical and empirical findings aimed at helping researchers to determine which form of supervisory signal they ought to collect for the task at hand. We have provided theoretical grounding for how hidden representations can be recovered through supervised classification and have related the quality of these recovered representations to training parameters such as number of labels, classes, and dimensions. We found that while hard labels and soft labels provide comparable amounts of information in the many-examples-but-few-classes regime, soft labels become increasingly preferable when the number of classes increases or the number of labels decreases. Our findings explain why, for example, pre-training a classifier on `ImageNet1K` (1,000 classes) or `ImageNet21k` (21,000 classes) using hard labels may lead to decent transfer learning performance [Huh et al., 2016, Ridnik et al., 2021] but pre-training with (a form of) soft labels may lead to even better transfer learning performance [Xie et al., 2020]. We support our theoretical contributions with empirical results on a suite of human-derived soft labels.

Table 1: Small data results with 80 labels/class.

Labels	**Cross-Entropy**
`CIFAR-10`	1.80
`CIFAR-10S+hard`	1.70
`CIFAR-10H`	1.67
`CIFAR-10S+dense`	1.71
`CIFAR-10LS (Low)`	1.78
`CIFAR-10DS`	1.71
`CIFAR-10LS (High)`	1.73

We note that, in our analysis, we made no assumptions about the data distribution in stimulus space, nor the function $f(x_i) = z_i$ that maps from stimulus space to hidden representations, but when training neural networks we often assume some level of stability or invariance (i.e., a small perturbation in pixel space does not lead to drastically different perception of the image). When satisfied, inductive biases like assumptions about stability or invariance allow learners to extract additional information from training examples, sometimes even in an unsupervised way when no labels are present. As a result, our analysis here can be considered as a sort of lower-bound on how much information about hidden representations a labeled training dataset can provide. We also examined each supervision signal in isolation, assuming that only labels of one type are collected. A promising future direction would be to analyze additional sources of information (inductive biases, other supervision signals, etc.) as well as the interactions between them; already, we see promising indications of mixing label types in the case of `CIFAR-10S+hard` and `CIFAR-10S+dense`.

Notwithstanding these limitations, our analysis of hard labels, soft labels, and sparse labels that interpolate between them, already enables researchers to develop cost-benefit tradeoff curves in order to optimize the cost of labeling their datasets for supervised learning—and support the development of data-efficient, generalizable ML systems.

References

Sameer Agarwal, Josh Wills, Lawrence Cayton, Gert Lanckriet, David Kriegman, and Serge Belongie. Generalized non-metric multidimensional scaling. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, pages 11–18, 2007.

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.

John Bridle. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. *Advances in neural information processing systems*, 2, 1989.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607, 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.

Mark A. Davenport. Lost without a compass: Nonmetric triangulation and landmark multidimensional scaling. In *2013 5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 13–16, 2013.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Li Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

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

Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019.

Greg Hanneman and Alon Lavie. Automatic category label coarsening for syntax-based machine translation. In *Proceedings of Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 98–106, 2011.

Martin N Hebart, Charles Y Zheng, Francisco Pereira, and Chris I Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, 4(11):1173–1185, 2020.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition*, pages 84–92, 2015.

Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes ImageNet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016.

Kevin G Jamieson and Robert D Nowak. Low-dimensional embedding using adaptively selected ordinal data. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1077–1084. IEEE, 2011.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33: 18661–18673, 2020.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Maksim Lapin, Matthias Hein, and Bernt Schiele. Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1533–1554, 2017.

Weiyang Liu, Zhen Liu, Hanchen Wang, Liam Paull, Bernhard Schölkopf, and Adrian Weller. Iterative teaching by label synthesis. In *NeurIPS*, 2021.

Maya Malaviya, Ilia Sucholutsky, Kerem Oktar, and Thomas L Griffiths. Can humans do less-than-one-shot learning?, Jun 2022. URL `https://escholarship.org/uc/item/486752zx`.

Raja Marjieh, Pol van Rijn, Ilia Sucholutsky, Theodore R Sumers, Harin Lee, Thomas L Griffiths, and Nori Jacoby. Words are all you need? capturing human sensory similarity with textual descriptors. *arXiv preprint arXiv:2206.04105*, 2022.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, pages 1614–1623, 2016.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

Joshua C Peterson, Joshua T Abbott, and Thomas L Griffiths. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, 42(8):2648–2669, 2018.

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.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

Brett D Roads and Bradley C Love. Enriching ImageNet with human similarity judgments and psychological embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3547–3557, 2021.

Sambu Seo and Klaus Obermayer. Soft learning vector quantization. *Neural computation*, 15(7):1589–1604, 2003.

Roger N Shepard. Multidimensional scaling, tree-fitting, and clustering. *Science*, 210(4468):390–398, 1980.

Pulkit Singh, Joshua C Peterson, Ruairidh M Battleday, and Thomas L Griffiths. End-to-end deep prototype and exemplar models for predicting human behavior. *arXiv preprint arXiv:2007.08723*, 2020.

Ilia Sucholutskv, Nam-Hwui Kim, Ryan P Browne, and Matthias Schonlau. One line to rule them all: Generating LO-shot soft-label prototypes. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

Ilia Sucholutsky and Matthias Schonlau. 'Less than one'-shot learning: Learning n classes from $m < n$ samples. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):9739–9746, 2021a.

Ilia Sucholutsky and Matthias Schonlau. Soft-label dataset distillation and text dataset distillation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021b.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.