On the Informativeness of Supervision Signals

Ilia Sucholutsky[*]
Dept. of Computer Science
Princeton University

Raja Marjieh
Dept. of Psychology
Princeton University

Thomas L. Griffiths
Dept. of Psychology
Dept. of Computer Science
Princeton University

Abstract

Learning transferable representations by training a classifier is a well-established technique in deep learning (e.g. ImageNet pretraining), but there is a lack of theory to explain why this kind of task-specific pre-training should result in 'good' representations. We conduct an information-theoretic analysis of several commonly-used supervision signals to determine how they contribute to representation learning performance and how the dynamics are affected by training parameters like the number of labels, classes, and dimensions in the training dataset. We confirm these results empirically in a series of simulations and conduct a cost-benefit analysis to establish a tradeoff curve allowing users to optimize the cost of supervising representation learning.

1 Introduction

The success of modern deep learning methods can perhaps best be understood through the lens of deep representation learning. Representation learning aims to learn latent embeddings of input stimuli. Generally, training a neural network means learning successive layers of representations that, typically, will be used in the final layer 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 for the model by quantifying the similarity between stimuli. Significant work has gone into the design of supervision signals for deep representation learning resulting in a plethora of tasks using 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]. In addition, deep representation learning techniques provide exciting new avenues for cognitive science. Specifically, uncovering the structure of human representations by directly training on behavioral data such as similarity judgments can provide a modern alternative to classic methods such as multi-dimensional scaling [Shepard, 1980].

Contrastive learning has been a particularly productive framework for achieving impressive results in learning useful, semantically-meaningful representations. Most contrastive learning frameworks involve training a model on pairs of stimuli with a loss function that rewards pushing together embeddings of similar stimuli and pushing apart the embeddings of dissimilar ones. However, we observe that something akin to contrastive learning also implicitly occurs when a model is trained with a classification objective, since embeddings from the same classes are pushed together in deeper layers while embeddings from different classes are pushed apart. Recent work has shown that models trained 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 conventional contrastive learning [Peterson et al., 2018, Marjieh et al., 2022]. Thus, different representation learning objectives can be applied to many overlapping use-cases, and it is not clear when one objective should be preferred over another.

[*] Correspondence to `is2961@princeton.edu`

Information-Theoretic Principles in Cognitive Systems Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

In this paper, we perform an information-theoretic analysis of two popular supervision signals from the classification literature – hard labels and soft labels – and quantify their relative (representational) information content by comparing them to similarity triplets (i.e., 'Is x more similar to y than to z?'), a popular supervision signal from contrastive learning and cognitive science literature [Jamieson and Nowak, 2011, Hoffer and Ailon, 2015]. We relate the quantity of information each signal provides to three common features of machine learning datasets: number of labels, number of classes, and dimensionality. We find that while both hard labels and soft labels provide information about hidden representations, the dynamics in terms of those three variables are very different between the signal types. We conduct a series of simulations to empirically confirm these results and to determine how additional information provided by each label translates into better representation learning performance. Finally, we conduct a cost-benefit analysis on these results to establish a tradeoff curve that allows users to optimize the cost of supervising representation learning.

2 Relative Informativeness of Supervision Signals

Problem definition. 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 hidden latent vectors. Our goal is to determine which supervision signal is most effective (in terms of both performance and cost) for representation learning.

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).

Third-order 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 both cases, neural networks are optimizing constraints of the form 'x is closer to y than to z', which we call 'third-order 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 third-order constraint is an inequality of the form $||z_a - z_b|| < ||z_a - z_c||$. This can be rewritten as the query $r_{i,j,k} = \{x \in \mathbb{R}^d : \|f(x_j) - f(x_i)\| < \|f(x_k) - f(x_i)\|\}$, 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. There are two types of third-order queries that we can extract from hard labels. 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 type 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 in each class. Then n hard labels give us $n(k-1)$

Table 1: Properties of supervision signals

Signal	# of constraints	Ratio	Asymptotic behavior
Hard labels	$n(k-1)+n^2(1-1/k)$	$\frac{2n(n+k-1-n/k)}{(n+k)(n+k-1)(n+k-2)}$	$\begin{cases} O(\frac{1}{n}) & n>>k \\ O(\frac{1}{n}) & n=k \\ O(\frac{1}{k^2}) & n<<k \end{cases}$
Soft labels	$nk(k-1)/2+kn(n-1)/2$	$\frac{kn}{(n+k)(n+k-1)}$	$\begin{cases} O(\frac{1}{n}) & n>>k \\ O(1) & n=k \\ O(\frac{1}{k}) & n<<k \end{cases}$

constraints of the second type and $k(n/k)(n-n/k)=n^2(1-1/k)$ constraints of the first type – a total of $n(k-1)+n^2(1-1/k)$ 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 third-order 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.

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 'good' representations. 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 label to the total number of possible queries. 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 Table 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, 2021]). Our results predict three scaling phases for soft labels and two scaling phases for hard labels.

3 Experiments

Simulations. We have posited that information ratios are a proxy for representation learning performance and have shown that if this is the case, then soft labels should lead to better performance than hard labels, particularly when there are few labels and many classes. However, we still need to understand how information ratio actually translates to representation learning performance. To that end, we conduct simulations to see the effect of four variables (and their interactions) on representation learning 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 randomly sampled centers $C_1, ..., C_k \in \mathbb{R}^d$. We compute hard and soft labels for these points using the equations defined above and then mine all third-order 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 third-order constraints. The Gram matrix outputted by GNMDS can be interpreted as the 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 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 results of the simulations

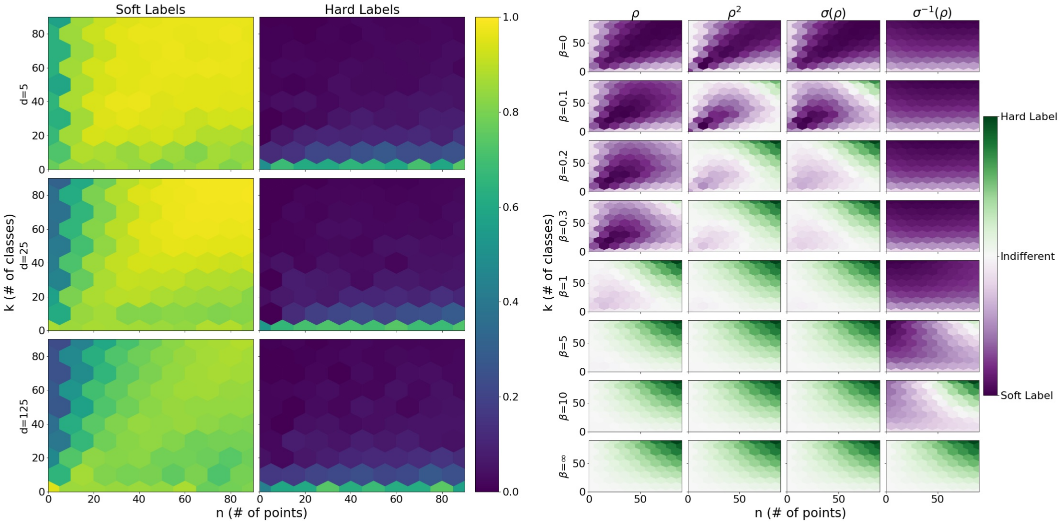

Figure 1: **Left**: Spearman rank correlation (ρ) of pairwise similarities recovered from running GNMDS on each set of labels when varying latent dimensionality (d), number of points (n), and number of classes (k). **Right**: Supervision signal preference based on subjective utility function ($u(\rho)$), cost weighting parameter (β), number of points (n), and number of classes (k).

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 representation learning performance.

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 $b > 0$, etc.). 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 cost 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 β, $\hat{u}$ and visualize the results for several combinations of these two variables 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.

4 Conclusion

In this paper, we have provided theoretical grounding for how hidden representations can be recovered through supervised classification, and we have related the quality of these recovered representations to training parameters like number of labels, classes, and dimensions. We find 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]. Finally, we enable researchers to analyze cost-benefit tradeoffs when determining which supervision signals to collect.

References

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.

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.

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

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

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.

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.

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

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

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, 2021.

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.

A Appendix

A.1 Limitations

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, assumptions about stability or invariance, often called ‘inductive biases’, 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.