Towards Better Understanding of Self-Supervised Representations

Neha Kalibhat [1] **Kanika Narang** [2] **Hamed Firooz** [2] **Maziar Sanjabi** [2] **Soheil Feizi** [1]

Abstract

Self-supervised learning methods have shown impressive results in downstream classification tasks. However, there is limited work in understanding and interpreting their learned representations. In this paper, we study the representation space of several state-of-the-art self-supervised models including SimCLR, SwaV, MoCo V2 and BYOL. Without the use of class label information, we first discover *discriminative features* that are highly active for various subsets of samples and correspond to unique physical attributes in images. We show that, using such discriminative features, one can compress the representation space of self-supervised models up to 50% without affecting downstream linear classification significantly. Next, we propose a sample-wise Self-Supervised Representation Quality Score (or, Q-Score) that can be computed without access to any label information. Q-Score, utilizes discriminative features to reliably predict if a given sample is likely to be mis-classified in the downstream classification task achieving AUPRC of 0.91 on SimCLR and BYOL trained on ImageNet-100. Q-Score can also be used as a regularization term to remedy low-quality representations leading up to 8% relative improvement in accuracy on all 4 self-supervised baselines on ImageNet-100, CIFAR-10, CIFAR-100 and STL-10. Moreover, through heatmap analysis, we show that Q-Score regularization enhances discriminative features and reduces feature noise, thus improving model interpretability.

1. Introduction

Self-supervised models learn to extract useful representations from data without relying on human supervision. These models (Chen et al., 2020a; Caron et al., 2020; Chen et al., 2020b; Grill et al., 2020; Chen & He, 2021; Caron et al., 2018; Khosla et al., 2020) have shown comparable results to supervised models in downstream classification tasks. By means of data augmentation, these models are trained to encode semantically relevant information from images while ignoring *nuisance* aspects. Therefore, the representations ultimately learned should only contain the information required to define a given sample. However, in practice, learned representations are often quite noisy and not interpretable, causing difficulties in understanding and debugging their failure modes (Jing et al., 2022; Huang et al., 2021; Ericsson et al., 2021).

In this paper, our goal is to study the representation space of self-supervised models such as SimCLR (Chen et al., 2020a), SwaV (Caron et al., 2020), MoCo (Chen et al., 2020b) and BYOL (Grill et al., 2020) and discover their informative features in an unsupervised manner. This can help us debug models better, improve their representation spaces and make them more interpretable. Understanding these learned representations is relatively less explored. (Jing et al., 2022), observes that self-supervised representations collapse to a lower dimensional space instead of the entire embedding space. Other methods (von Kügelgen* et al., 2021; Xiao et al., 2021), propose to separate the representation space into variant and invariant information so that augmentations are not task-specific. (Grigg et al., 2021), observe representations across layers of the encoder and compare it to supervised setups. In this work, we focus more on thoroughly studying the representation space of SSL methods and their properties. We investigate the connections between the unsupervised properties in the representation space and mis-classifications when the representations are used in downstream classification tasks. We summarize our contributions as follows:

- We study the representation space of self-supervised models and discover *discriminative features*, in an unsupervised fashion.
- We show that discriminative features often have unique

[1] Department of Computer Science, University of Maryland, College Park, United States [2] Meta AI, Menlo Park, United States. Correspondence to: Neha Kalibhat <nehamk@umd.edu>.

Published at the ICML 2022 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA. Copyright 2022 by the author(s).

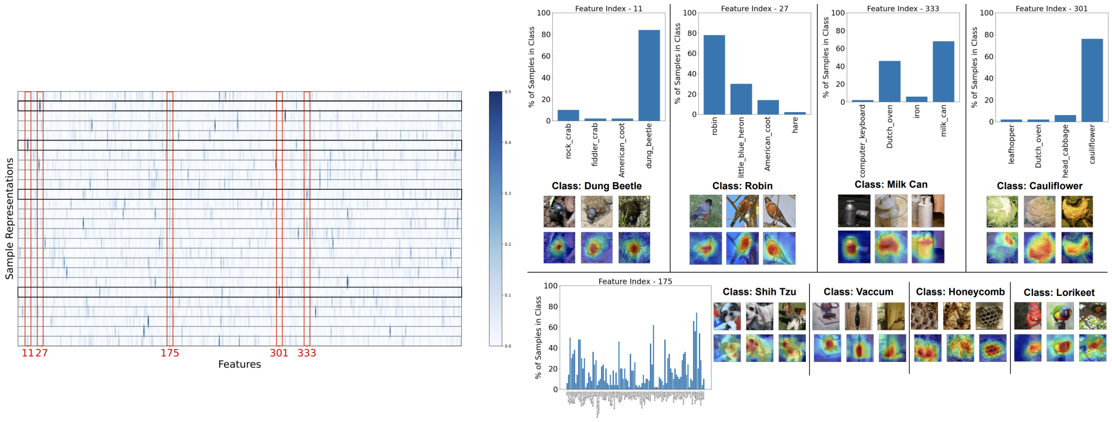

Figure 1. We plot SimCLR representations of randomly selected ImageNet-100 samples (left) and the percentage of samples within each class where a given feature is included in the set of dominant features (right).

physical meanings and although are discovered without any label information (given that SSL models are trained without labels), show strong correlation to class labels. Moreover, with discriminative features, representations can be compressed by up to 50% reliably.

- We introduce **Self-Supervised Quality Score (Q-Score)** to measure the quality of each learned representation. We empirically observe that the higher the Q-Score, the more likely that the sample will be correctly classified, achieving an AUPRC of up to 0.91.
- We apply Q-Score as a regularizer to the self-supervised loss and show that, by improving the quality of low-score samples, we can improve downstream classification accuracy by 8%. Moreover, we show that using the Q-Score regularization improves the interpretability of self-supervised representations by removing noise and highlighting useful information.

2. Self-Supervised Representations

Let us consider a SimCLR model with a ResNet (He et al., 2016) base encoder $f(.)$ and an MLP projection head $g(.)$. We define $\mathbf{x}_i \in \mathbb{R}^n$ and $\tilde{\mathbf{x}}_i \in \mathbb{R}^n$ as two transformed views of the i^{th} sample in a given input dataset containing N samples. Our setup is identical to SimCLR. We apply data transformations, *random crop*, *random horizontal flip*, *random color distortion* and *random Gaussian blur*. We pass the input samples through the base encoder to get self-supervised representations denoted by $f(\mathbf{x}_i) = \mathbf{h}_i \in \mathbb{R}^r$ and $f(\tilde{\mathbf{x}}_i) = \tilde{\mathbf{h}}_i \in \mathbb{R}^r$ where r is the size of the representation space. For contrastive training, we use the output of the projection head $g(\mathbf{h}_i) = \mathbf{z}_i \in \mathbb{R}^p$ and $g(\tilde{\mathbf{h}}_i) = \tilde{\mathbf{z}}_i \in \mathbb{R}^p$ where p is the size of the projection space. The SimCLR optimization for the set of model parameters θ, is as follows,

$$\max_{\theta} \frac{1}{2N} \sum_{i=1}^{2N} \frac{\exp(sim(\mathbf{z}_i, \tilde{\mathbf{z}}_i))}{\sum_{j=1}^{2N} \mathbb{1}_{j \neq i} \exp(sim(\mathbf{z}_i, \mathbf{z}_j))} \tag{1}$$

where $sim(\mathbf{z}_i, \mathbf{z}_j) = \frac{1}{\tau} \frac{\mathbf{z}_i^T \mathbf{z}_j}{\|\mathbf{z}_i\| \|\mathbf{z}_j\|}$.

In Figure 1, we visualize the representations of SimCLR pretrained on ImageNet-100 (Russakovsky et al., 2015). Each row denotes the latent vector ($\mathbf{h}_i$) of a random sample at index i of the ImageNet-100 test set. There are 512 columns corresponding to the representation size of a ResNet-18 (He et al., 2016) encoder. First, we study properties of sample representations ($\mathbf{h}_i$). We observe that each representation is *nearly* sparse, i.e., most feature values are close to zero (Jing et al., 2022). However, there exists a select few features that are strongly deviated from the remaining features in any given representation. For ease of visualization, we have highlighted 4 different representations (in black), which show at least one dominant feature. For the i^{th} sample whose latent representation is $\mathbf{h}_i \in \mathbb{R}^r$, let μ_i denote the mean of $\mathbf{h}_i$ and σ_i denote the standard deviation of $\mathbf{h}_i$. We formally define the set of *dominant features* for the i^{th} sample as, $L_i := \{j : h_{ij} > \mu_i + \epsilon\sigma_i\}$.

where ϵ is a hyperparameter that is empirically selected. In practice, we find that $\epsilon = 4$ works best for our experiments. We observe that dominant features of a sample may be unique to that particular representation or may be shared with other samples in the population. For example, in Figure 1, features 11 and 301 (highlighted in red) are strongly activated for a single sample, whereas, features 27 and 333 are strongly activated for more than one sample. In Figure 1, we plot all the samples where the given feature is strongly activated (i.e., dominant) and group them by their class labels (note that the selection of dominant features are done without using the label information). If we take feature 11, 27, 301 and 333, we observe that they are dominant for over 80% of a particular class and significantly higher than the remaining classes. The gradient heatmaps also correspond to informative visual attributes necessary for identifying these objects. We will refer to these features as *discriminative features* defined formally in the next section. On the other hand, some dominant features may be strongly activated for a large number of samples in the population. For example, in Figure 1, we illustrate the range of classes that feature 175 is dominant for, varying from birds, animals, vegetables, household items etc. If we examine the gradient heatmaps of this feature across various classes, we observe that this feature does not have a clear physical interpretation.

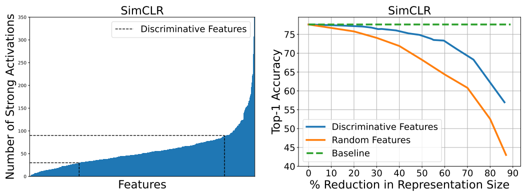

Figure 2. On the left, we plot each feature in ascending order of the number of times they are dominant in the population (y axis) and select the middle portion as discriminative features. On the right, we plot the linear accuracy of various subsets of discriminative features.

3. Selecting Discriminative Features

In Figure 2 (left), we plot the number of samples where a given feature at index j is part of the set of dominant features (i.e., $j \in L_i$). Intuitively, this is the number of times a given feature is strongly activated for the samples in the population. The features are visualized in the ascending order of the number of strong activations in the population (containing 5000 samples).

We define three broad categories of dominant features: (i) Features that are strongly activated across a very small fraction of the population, corresponding to the lower tail features in Figure 2. These features are image-specific and are unlikely to have class-relevance. (ii) Features dominant across a large number of samples in the population i.e, the upper tail features in Figure 2. Like feature 175 in Figure 1), such features are likely to encode very broad and general characteristics (like texture, color etc.) common to most samples and therefore, are not class-discriminative. The third category includes, (iii) features that are active across a moderate number of samples in the population (i.e. the middle parts in Figure 2). These features are most likely to encode unique physical attributes associated with particular classes, similar to those illustrated in the top panel in Figure 1. We refer to this subset of dominant features as *discriminative features* (denoted by the dotted lines in Figure 2). We emphasize that these discriminative features are selected without the use of any label information.

We justify the described method of selection in Figure 2 (right), where we plot the top-1 accuracy of a linear classifier on ImageNet-100 using subsets of discriminative features of varying sizes. We also plot the top-1 accuracy when random subsets of features are selected. We observe that discriminative features perform significantly better than randomly features and can reduce the representation size up to 50%, with minimal reduction in performance. See Appendix Figures A.1 and A.2 for results on other self-supervised models.

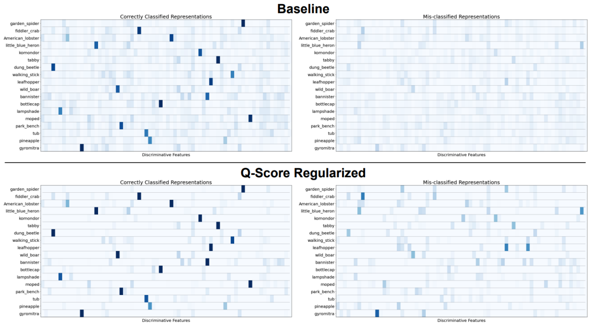

Figure 3. We visualize discriminative features of average representations (correct and incorrect classifications) of several ImageNet-100 classes. In the top panel, we display the SimCLR (baseline) and in the bottom panel, we visualize the same after using Q-Score regularization.

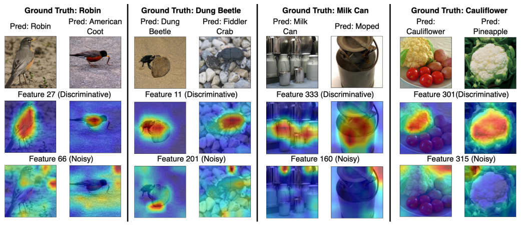

Figure 4. We visualize heatmaps of discriminative and noisy features of SimCLR on 4 classes in ImageNet-100.

4. Interpreting Representations

In this section, we study some properties of the latent space that drive representations to be correctly classified by a downstream linear classifier. In Figure 3, we visualize the discriminative features selected in an unsupervised manner using the approach outlined in Section 3 for several ImageNet-100 classes. Each row corresponds to the class averaged representation. On the left, we show the average representations of correctly classified samples in each class while on the right, we show the same for mis-classified representations in each class.

As we can see, in Figure 3, in the first panel, there is a clear difference between representations of correctly and incorrectly classified examples. Both correct and mis-classified representations are *nearly* sparse, however, the discriminative features are strongly deviated only in correct classifications. This is especially interesting because the label information has not been used in the selection of discriminative features. In Figure 1, we observe that discriminative features show strong correlation to ground truth, which suggests that their presence may be useful in correctly classifying representations. In Figure 3, our claim is confirmed as we observe that mis-classified representations do not show high activation on discriminative features. For every representation, apart from the discriminative features, there are a large number of features that have very low activation (close to zero). We name such features as *noisy features* for each representation. Correctly classified representations contain few noisy features and are more sparse while mis-classified representations are often more noisy. Our visual observations hint that we can potentially classify representations in an unsupervised manner by leveraging these structural properties of representations.

We observe that the heatmaps of discriminative features (in Figure 4) of any given class, capture relevant and defining characteristics of the images, therefore are highly correlated with the ground-truth. The physical attribute associated with a discriminative feature is consistent between the correct and incorrect classifications. Noisy features lead to noisy heatmaps which focus on aspects of the inputs that are uninformative (See Section A.10). For any given sample, having a large number of noisy features and low activations on discriminative features are strong signals indicating its potential mis-classification in the downstream task. We would like to emphasize that our results only indicate an *association* between these structural properties and classification accuracy and we do not claim any causal relationship between the two.

5. Self-Supervised Q-Score

Our study of learned representation patterns help us discover discriminative features in an unsupervised manner. We combine our observations to design a sample-wise quality score for self-supervised representations.

Let us define D, such that $|D| < r$, as the set of discriminative features for a given self-supervised model trained on a given dataset. For the i^{th} sample, we have $\mathbf{h}_i \in \mathbb{R}^r$ (representation), μ_i (mean of $\mathbf{h}_i$), σ_i (standard deviation of $\mathbf{h}_i$) and the set of dominant features $L_i = \{j : h_{ij} > \mu_i + \epsilon\sigma_i\}, |L_i| < r$. We define our Self-Supervised Quality Score for sample i as,

$$Q_i := \frac{\sum_{j \in L_i \cap D} h_{ij}}{|L_i \cap D| \|\mathbf{h}_i\|_1} \tag{2}$$

where, $L_i \cap D$ is the set of discriminative features specific to the i^{th} sample. We also use the L_1 norm $\|h_i\|_1$ to promote sparsity in representations consequently ensuring that noisy features are penalized. Intuitively, higher Q_i implies that the representation is sparse with strongly activated discriminative features. It combines two favorable properties of representations i.e., sparsity (or feature noise, computed by the L_1 norm) and strongly activated discriminative features. Our objective with this metric is to compute a sample-specific score in an unsupervised manner indicating the quality of its representations. Ideally, we would like to argue that samples with higher Q-score have improved representations and thus are more likely to be classified correctly in the downstream task. We confirm this in Section A.3.

Table 1. We compute the linear accuracy before and after Q-Score regularization and show that Q-Score regularization improves the baselines by up to 8%.

Dataset	SimCLR		SwaV		MoCo		BYOL	
	Baseline	Q-Score Regularized	Baseline	Q-Score Regularized	Baseline	Q-Score Regularized	Baseline	Q-Score Regularized
CIFAR-10	90.83	**92.31**	89.17	**90.03**	86.91	**92.77**	86.72	**92.25**
CIFAR-100	65.91	**71.90**	62.89	**66.52**	63.47	**68.16**	60.97	**67.71**
STL-10	76.42	**79.83**	73.94	**75.03**	73.21	**74.29**	70.59	**74.47**
ImageNet-100	77.62	**80.79**	74.09	**78.90**	78.32	**85.16**	80.10	**86.72**

We can also use Q-Score as an intervention by further training state-of-the-art self-supervised models with Q-Score regularization. For example, we can apply this regularizer to the SimCLR optimization as follows,

$$\max_{\theta} \frac{1}{2N} \sum_{i=1}^{2N} \Big[\frac{\exp(sim(\mathbf{z}_i, \tilde{\mathbf{z}}_i))}{\sum_{j=1}^{2N} \mathbb{1}_{j \neq i} \exp(sim(\mathbf{z}_i, \mathbf{z}_j))} + \lambda \mathbb{1}_{Q_i < \alpha}(Q_i) \Big] \tag{3}$$

where, α is a threshold with which we select the samples whose Q-Scores should be maximized and λ is the regularization coefficient. In other words the goal of this regularization is to improve low-quality representations, similar to the ones shown in Figure 3, by maximizing their Q-Score to improve their quality for downstream classification. As shown in in Table 1, Q-Score regularization improves the top-1 accuracy on each dataset on all of the self-supervised state-of-the-art models. On ImageNet-100 we observe up to an 8% relative improvement in accuracy, most significant on MoCo.

In addition to top-1 accuracy, Q-Score also shows significant improvement in representation quality and interpretability. In Figure 3, we compare the representation space before and after Q-Score regularization. We observe that discriminative features become more enhanced after Q-Score regularization on both correct and mis-classified representations. Our regularization also greatly reduces noisy features and produces cleaner representations with clear discriminative features that are easier to classify (See Section A.4). Therefore, we attribute the improvement in performance to improved representation quality. Our motivation is to produce better quality representations that are more distinguishable across classes and therefore, easier to classify. Although Q-Score improves accuracy, it does not entirely prevent mis-classifications as mis-classifications may occur due to a variety of reasons such as, hardness of samples, encoder capacity, dataset imbalance etc.

References

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Caron, M., Bojanowski, P., Joulin, A., and Douze, M. Deep clustering for unsupervised learning of visual features. *Lecture Notes in Computer Science*, pp. 139–156, 2018. ISSN 1611-3349. doi: 10.1007/978-3-030-01264-9_9. URL http://dx.doi.org/10.1007/978-3-030-01264-9_9.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9912–9924. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/70feb62b69f16e0238f741fab228fec2-Paper.pdf.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020a. URL http://proceedings.mlr.press/v119/chen20j.html.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15750–15758, June 2021.

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning, 2020b.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., , and Vedaldi, A. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Coates, A., Lee, H., and Ng, A. Y. Stanford stl-10 image dataset. URL https://cs.stanford.edu/~acoates/stl10/.

Ericsson, L., Gouk, H., and Hospedales, T. M. Why do self-supervised models transfer? investigating the impact of invariance on downstream tasks, 2021.

Grigg, T. G., Busbridge, D., Ramapuram, J., and Webb, R. Do self-supervised and supervised methods learn similar visual representations?, 2021.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., and Valko, M. Bootstrap your own latent - a new approach to self-supervised learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Huang, W., Yi, M., and Zhao, X. Towards the generalization of contrastive self-supervised learning, 2021.

Jing, L., Vincent, P., LeCun, Y., and Tian, Y. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=YevsQ05DEN7.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning, 2020.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-10 (canadian institute for advanced research). a. URL http://www.cs.toronto.edu/~kriz/cifar.html.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-100 (canadian institute for advanced research). b. URL http://www.cs.toronto.edu/~kriz/cifar.html.

Maji, S., Kannala, J., Rahtu, E., Blaschko, M., and Vedaldi, A. Fine-grained visual classification of aircraft. Technical report, 2013.

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

von Kügelgen*, J., Sharma*, Y., Gresele*, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, December 2021. *equal contribution.

Xiao, T., Wang, X., Efros, A. A., and Darrell, T. What should not be contrastive in contrastive learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=CZ8Y3NzuVzO.

A. Appendix

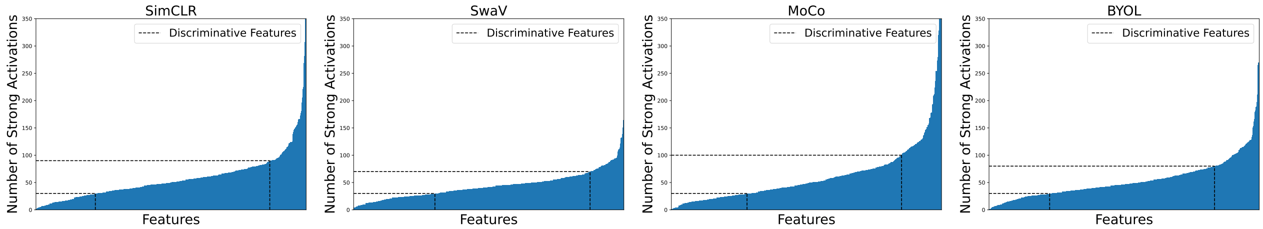

Figure A.1. **Selecting discriminative features:** We plot each feature in ascending order of the number of times they are dominant in the population (y axis). We show this for SimCLR (Chen et al., 2020a), SwaV (Caron et al., 2020), MoCo (Chen et al., 2020b) and BYOL (Grill et al., 2020). Discriminative features are selected such that they are dominant for a range of samples, indicating that they may have strong class-correlation, therefore, are useful for downstream classification. Features that are activated for a very large number of samples may not be discriminative as they often encode information common to several classes (e.g. Feature 175 in Figure 1).

A.1. Discussion

In this paper, we studied the representation space of self-supervised models in downstream classification tasks and discovered discriminative features. Discriminative features are a subset of features that show strong activations in smaller but sizable sets of samples, wherever relevant, and correspond to unique physical attributes. With discriminative features, we can reduce the representation size by 50% without affecting linear accuracy. We also observe significant differences in representations between correctly and incorrectly classified samples. Building on these observations, we define Self-Supervised Quality Score (Q-Score) that is effective in determining how likely samples are to be correctly or incorrectly classified. Our proposed score can be computed per sample in an unsupervised manner (without label information). With the help of Q-Score regularization, we remedied these low-quality samples by improving their Q-Scores, thereby, improving the overall accuracy of state-of-the-art self-supervised models by up to 8%. We also observed that regularization improves model interpretability by enhancing discriminative features and reducing feature noise. Our paper poses important questions for future studies such as: 1) why do self-supervised models trained without any labels, produce axis-aligned, sparse representations, 2) what are the causes of mis-classifications, apart from representation quality, 3) how can we utilize the better representations space for other tasks besides classification.

A.2. Experimental Setup

Our setup consists of pre-trained self-supervised encoders ($f(.)$) SimCLR (Chen et al., 2020a), SwaV (Caron et al., 2020), MoCo (Chen et al., 2020b) and BYOL (Grill et al., 2020) on ImageNet-100 (Russakovsky et al., 2015), CIFAR-10 (Krizhevsky et al., a), CIFAR-100 (Krizhevsky et al., b) and STL-10 (Coates et al.). We maintain the same encoder optimization, training parameters and optimizers as the respective papers. We train the pre-trained encoders on their self-supervised objectives with our Q-Score regularizer on the latent representations. We train with $\lambda = 0.1$ for 100 epochs until convergence on a single NVIDIA RTX A4000 GPU.

A.3. ROC Plots of Q-Score

We measure how effective our score is in differentiating between correctly and incorrectly classified representations in an unsupervised manner. In Figure A.3, we plot the Precision-Recall (PR) curve and the Receiver Operating Characteristic (ROC) curve of Q-Score for SimCLR, SwaV, MoCo and BYOL, for the validation set of ImageNet-100 containing 5000 samples. We also compute the AUROC (area under receiver operating characteristic curve) and AUPRC (area under precision-recall curve) in differentiating between correct and incorrect classifications. We observe up to 0.91 AUPRC and 0.74 AUROC among our baselines. Based on these results we can conclude that, Q-Score is a reliable metric in assessing the quality and representations with lower Q-Score (quality), are more likely to be mis-classified.

A.4. Gradient Heatmaps after Q-Score Regularization

In Figure A.4, we visualize heatmaps of several samples of 4 classes in the Q-Score regularized model. The discriminative features under each class, activate relevant attributes for each example within that class. The noisy features, in contrast

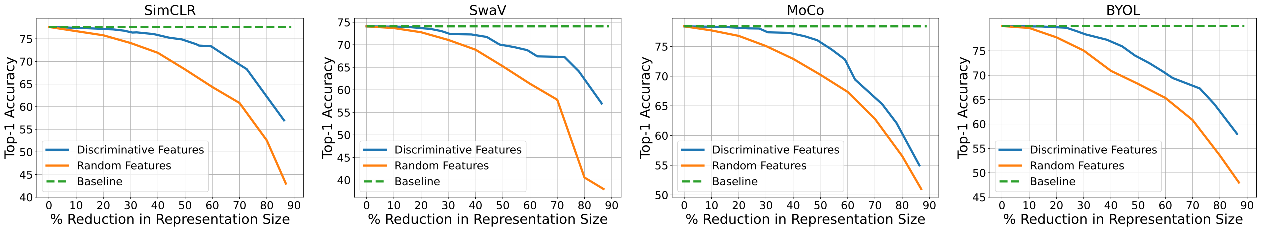

Figure A.2. **Linear classification accuracy on discriminative features:** We train linear classifiers after selecting subsets of discriminative features of various sizes (middle portion of Figure 2) and plot their top-1 accuracy for SimCLR, SwaV, MoCo and BYOL. We compare these results to the baseline and the accuracy on randomly selected features where models trained using discriminative features consistently outperform those of randomly selected features. We can achieve up to 50% reduction in representations size using discriminative features without significantly affecting the top-1 accuracy.

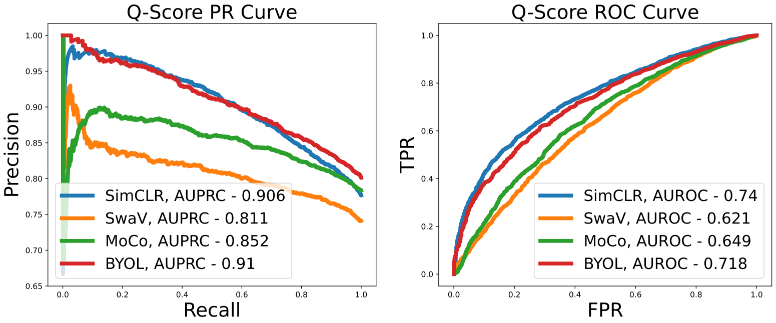

Figure A.3. **Precision-Recall and ROC curves of Q-Score:** We compute the precision-recall and ROC curve of Q-Score for correct and mis-classified representations on ImageNet-100 on SimCLR, SwaV, MoCo and BYOL. We achieve an AUPRC of up to 0.91 and AUROC of 0.71 in distinguishing between correct and mis-classified representations using Q-Score.

to Figure 4, are now not activated for the majority of the examples. Strongly activated discriminative features, sparse representations (Figure 3) and reduced noise in heatmaps (Figure A.4), are indicators that Q-Score regularized representations are more interpretable for downstream classification.

A.5. Linear Head

In Figure A.5, we plot the magnitude of linear head weights of the top classes for each feature. We observe that the linear head weight at the given feature index is strongly correlated with a particular class. The class mappings are identical to those observed in Figure 1. This indicates that the linear head implicitly learns to assign higher weights for discriminative features at their respective classes.

A.6. Selecting Discriminative Features

In Figure 2, we select features based on the number of times they are dominant in the population. Discriminative features are selected in the middle portion, illustrated by dotted black lines. In this section, we discuss the performance when the selection range is either to the left or the right of the curve. In Figure A.6, we add two additional baselines on top of Figure A.2. The red curve shows the top-1 accuracy of features selected from the lower tail (left) of Figure 2. The purple curve shows the top-1 accuracy of features selected from the upper tail (right) of Figure 2. We observe that discriminative features outperform all baselines, indicating that selecting discriminative features from the middle portion in Figure 2 is optimal for best linear evaluation performance.

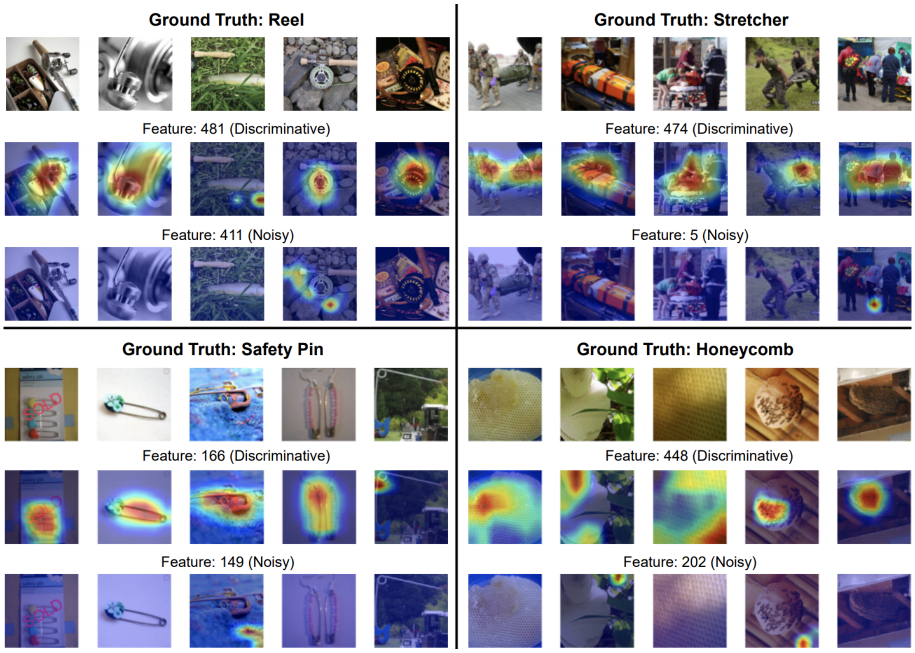

Figure A.4. **Heatmaps of discriminative and noisy features after Q-Score Regularization:** We visualize heatmaps of discriminative and noisy features of 4 classes in Q-score regularized SimCLR. We observe that discriminative features capture meaningful attributes of each image. The noisy features (unlike Figure 4), do not get activated in most cases indicating that the regularization reduces noise in the representation space, thereby improving model interpretability. These observations are consistent between correct and incorrect classifications.

A.7. Deconstructing Self-Supervised Q-Score

The self-supervised Q-Score contains two components - 1) The mean of the discriminative features within a representation and 2) The L1-norm of the representation. In Figure A.8, we visualize the AUC curves for each component separately to study their effectiveness in discriminating between correct and mis-classified representations. We observe that the AUPRC and AUROC values, for each individual component is significantly lower compared to Q-Score in Figure A.3, for each self-supervised model baseline. Combining both the components allows us to promote the magnitude of discriminative features, and at the same time ensure that representations remain sparse. The improved AUC numbers in Figure A.3, confirm that combining both properties in Q-Score is more effective in distinguishing between correct and incorrect classifications.

In Figure A.7, we also visualize histograms of Q-Score between correct and mis-classified samples for 4 self-supervised models. We observe that mis-classified representations are clearly shifted in their Q-Score distribution on each model. This confirms that Q-Score, without the use of any label information, can predict the likelihood of being correctly classified for any given representation.

A.8. Class-wise Accuracy and Sparsity

In Figure A.9, we plot the class-wise accuracy of ImageNet-100 classes before and after Q-Score regularization. We observe that Q-Score regularization, maintains or improves that performance of 83 classes and does not significantly degrade the performance of any class. In Figure A.10, we plot the sparsity of the representations in ImageNet-100 before and after regularization. Due to the use of L1-norm, we observe a significant increase in sparsity.

A.9. Transfer Performance of Q-Score Regularization

In Table A.1, we tabulate the transfer learning performance (linear evaluation) of various unseen datasets (Krizhevsky et al., a;b; Coates et al.; Maji et al., 2013; Nilsback & Zisserman, 2008; Bossard et al., 2014; Krause et al., 2013; Cimpoi et al., 2014) on 4 self-supervised models trained on ImageNet-100 with Q-Score regularization. We observe that the average accuracy of unseen datasets improves on all setups, especially on SwaV and BYOL. We would like to mention that these

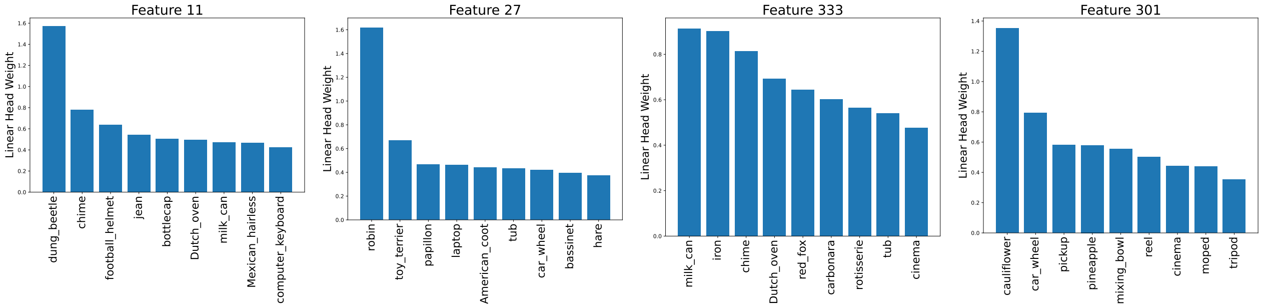

Figure A.5. **Linear head weights at various features:** We plot the weight magnitude of the linear head for various features and show that the weights are significantly high for select classes in ImageNet-100. These classes exactly match with Figure 1.

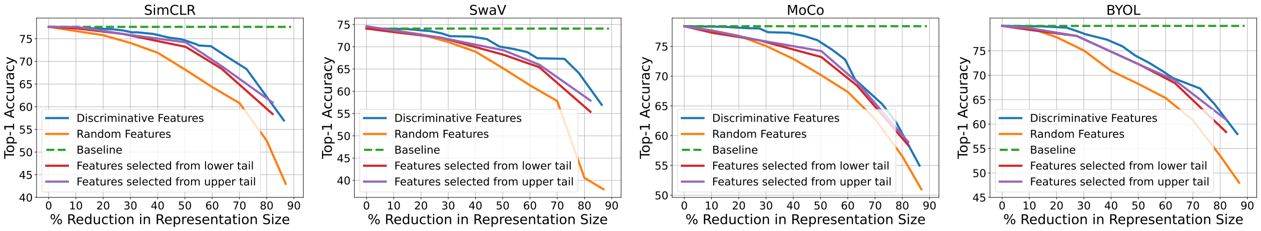

Figure A.6. **Top-1 Accuracy of subsets of features:** Building on Figure A.2, we plot the linear accuracy of features selected from the lower tail in Figure 2 (red) and features selected from the upper tail in Figure 2. We observe that discriminative features outperform all the baselines.

self-supervised models use ResNet-18 and are pre-trained on ImageNet-100 due to limited resource constraints, therefore, the accuracy numbers may be lower than those reported in the baselines.

In Figure A.11, we visualize the gradient heatmaps of the discriminative features discovered on SimCLR on ImageNet-100 on both ImageNet-100 and unseen datasets, Aircraft (Maji et al., 2013), Food (Bossard et al., 2014) and Cars (Krause et al., 2013). We observe that the physical meaning associated with each discriminative feature is consistent between both the training and unseen data. The heatmaps also correspond to informative features, strongly correlated with the ground truth. These gradients indicate that discriminative features are transferable across unseen datasets, which support the improvement we observe in Table A.1.

We also visualize the representations of correct and incorrect classifications of the Flowers (Nilsback & Zisserman, 2008) dataset in Figure A.12. We use SimCLR pre-trained on ImageNet-100 (top panel) and with Q-Score regularization (bottom panel). We observe that the same properties as Figure 3 on ImageNet-100 (train dataset) transfer at test time to Flowers, an unseen dataset. Before regularization, representations, especially the mis-classified ones are more noisy. We observe dominant features on each sample which get more enhanced after Q-Score regularization. The representations also become more sparse with reduced noise leading to improved top-1 accuracy as shown in Table A.1.

A.10. More Gradient Heatmaps of SimCLR

In Figures A.13, A.14, A.15 and A.16, we plot more heatmaps of discriminative and noisy features of SimCLR. We observe that discriminative features are highly correlated with the ground truth, whereas, noisy features, map to spurious portions that do not contribute to useful information.

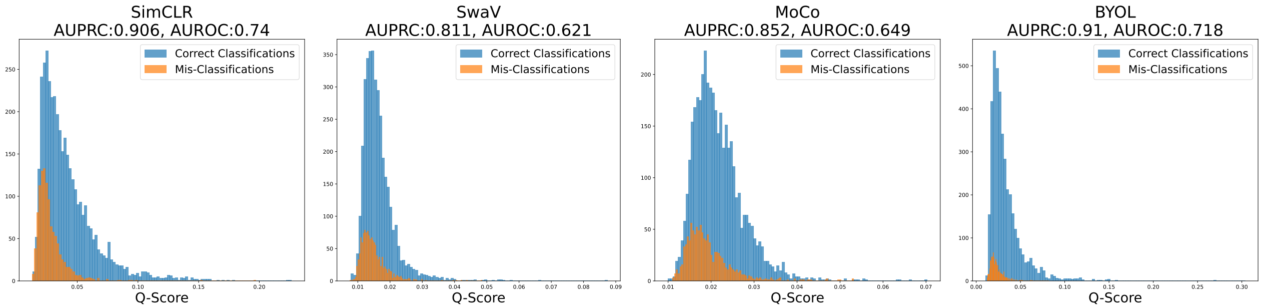

Figure A.7. **Histograms of Q-Score of correct and incorrect classifications:** We plot the distributions of Q-Score between correct and mis-classified samples in SimCLR, SwaV, MoCo and BYOL. We observe that the distributions are clearly shifted, indicating that Q-Score is effective predicting whether any given sample would be correctly or incorrectly classified.

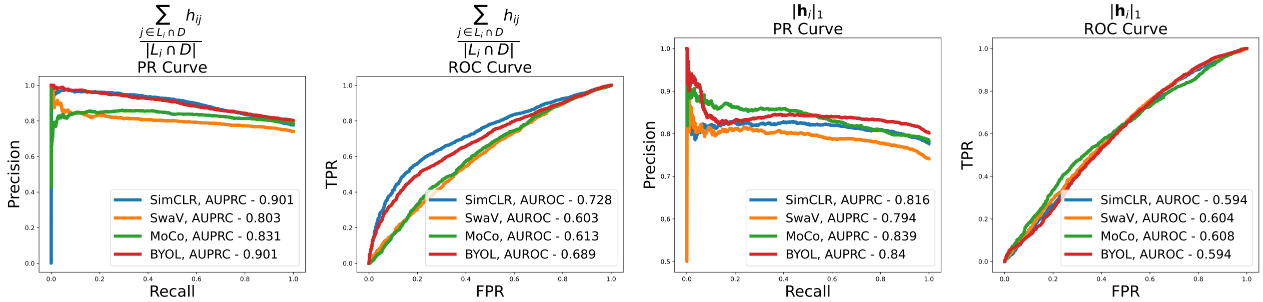

Figure A.8. **AUC plots of Q-Score components:** We plot the ROC and PR curves of the two components of Q-Score. We observe that, on all self-supervised models, the AUROC and AUPRC scores are better with Q-Score (Figure A.3) which combines both properties.

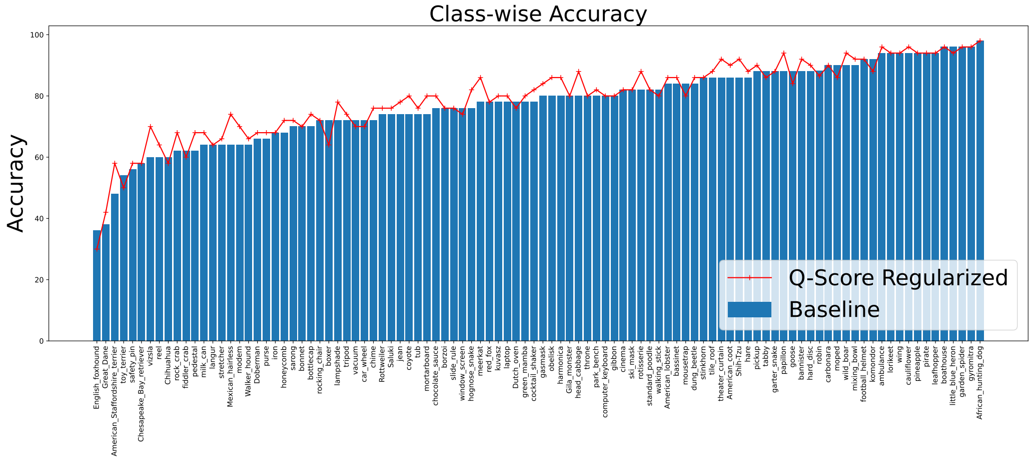

Figure A.9. **Class-wise accuracy of SimCLR with and without Q-Score regularization):** We observe that Q-Score regularization improves the accuracy of 83 classes in ImageNet-100. Q-Score regularization, does not degrade the performance significantly for any class.

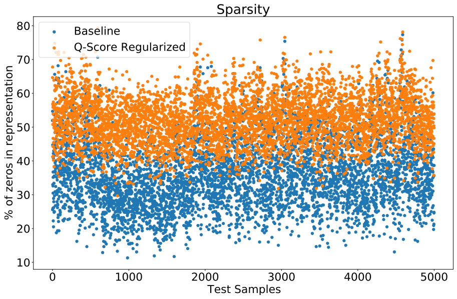

Figure A.10. **Sparsity of representations:** In this scatter plot, we show that Q-Score regularization significantly increases the sparsity of representations compared to the SimCLR baseline.

Table A.1. **Transfer learning performance of various state-of-the-art self-supervised models trained on ImageNet-100 with Q-Score regularization:** We observe that Q-Score regularization improves the average transfer accuracy on all self-supervised models.

Transfer Dataset	**SimCLR**		**SwaV**		**MoCo**		**BYOL**	
	Baseline	**Q-Score Regularized**	**Baseline**	**Q-Score Regularized**	**Baseline**	**Q-Score Regularized**	**Baseline**	**Q-Score Regularized**
CIFAR-10	70.13	**70.55**	71.27	**72.42**	**73.26**	72.39	71.36	**72.99**
CIFAR-100	40.23	**40.70**	42.52	**42.69**	**45.70**	44.11	**45.92**	45.36
STL-10	65.74	**65.77**	65.81	**65.89**	66.87	**67.03**	85.45	**86.07**
Aircraft	**11.94**	11.79	11.97	**17.31**	12.06	**13.08**	**11.91**	11.73
Flowers	49.52	**51.63**	49.53	**55.03**	50.20	**50.82**	50.23	**51.26**
Food	**48.47**	48.01	48.38	**51.73**	48.36	**49.74**	48.35	**50.22**
Cars	**10.67**	10.59	10.63	**16.17**	10.68	**12.47**	10.72	**13.09**
DTD	55.69	**56.06**	55.63	**57.18**	55.90	**57.12**	55.63	**56.06**
Average	44.05	**44.39**	44.47	**47.30**	45.38	**45.85**	47.45	**48.35**

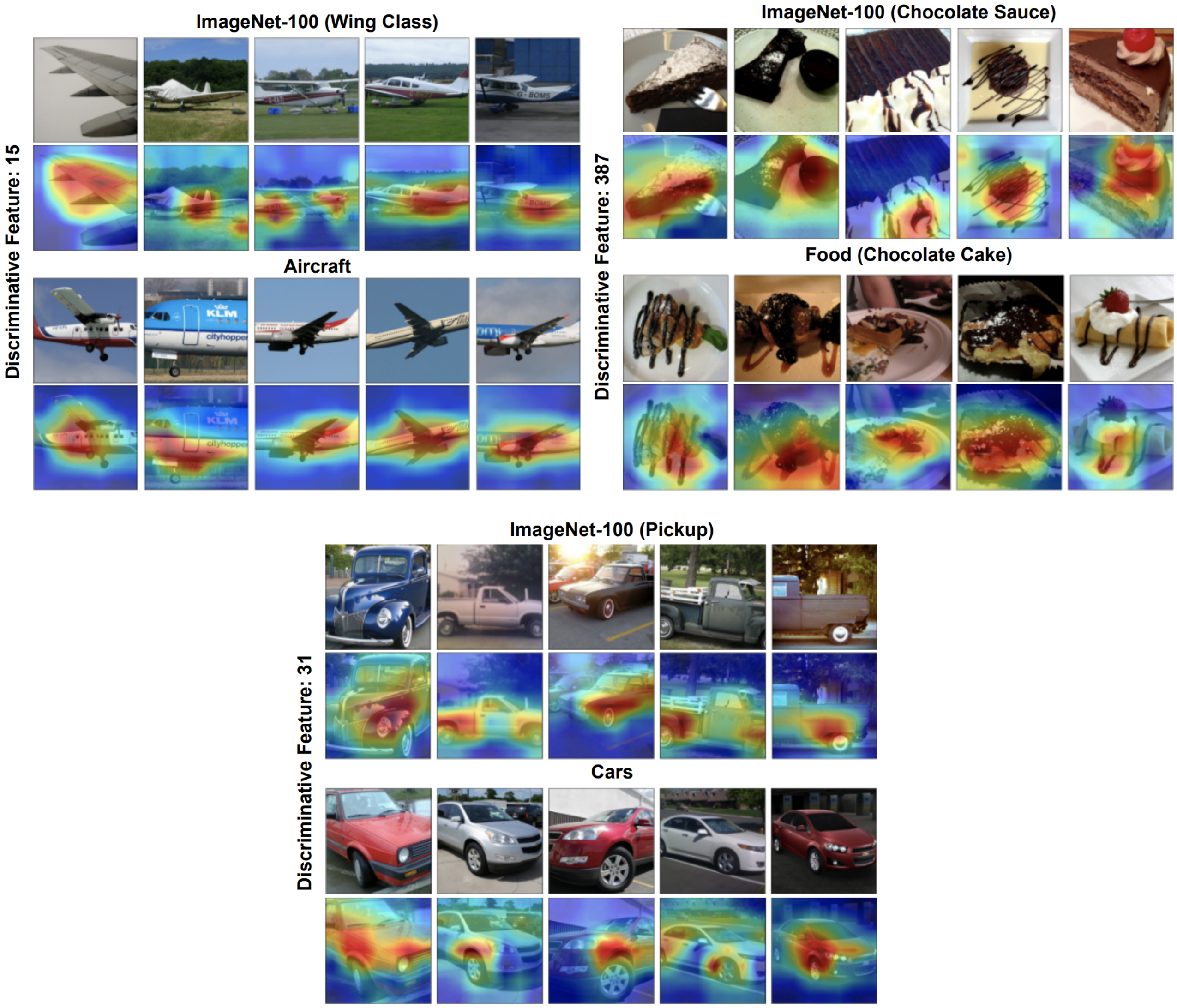
ImageNet-100 (Wing Class)
Aircraft
Discriminative Feature: 15
ImageNet-100 (Chocolate Sauce)
Food (Chocolate Cake)
Discriminative Feature: 387
ImageNet-100 (Pickup)
Cars
Discriminative Feature: 31

Figure A.11. **Discriminative features on unseen datasets:** We visualize the discriminative features discovered on ImageNet-100 on unseen datasets like Aircraft (Maji et al., 2013), Food (Bossard et al., 2014) and Cars (Krause et al., 2013). We observe that discriminative features correspond to the same physical attribute as the training data and are core and informative.

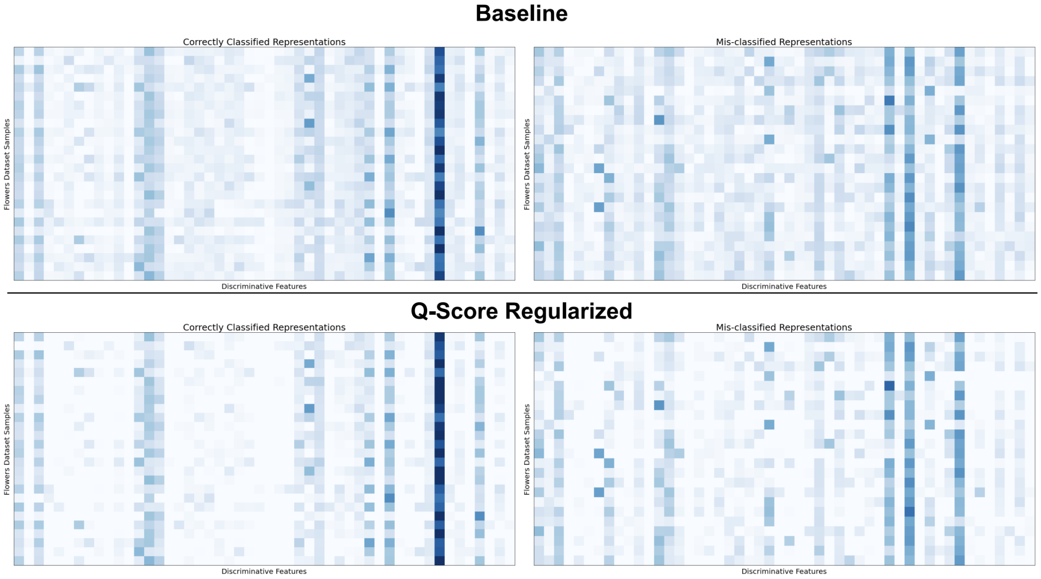

Figure A.12. **Comparing correct and mis-classified representations in Flowers dataset:** In these heatmaps, we visualize the top features several Flowers (Nilsback & Zisserman, 2008) dataset samples. In the top panel, we display the correct (left) and incorrect (right) classifications of SimCLR (trained on ImageNet-100) and in the bottom panel, we visualize the same after using Q-Score regularization. Similar to the observations in Figure 3, we observe that the regularization makes representations more sparse with discriminative features enhanced, thereby leading to an improvement in performance.

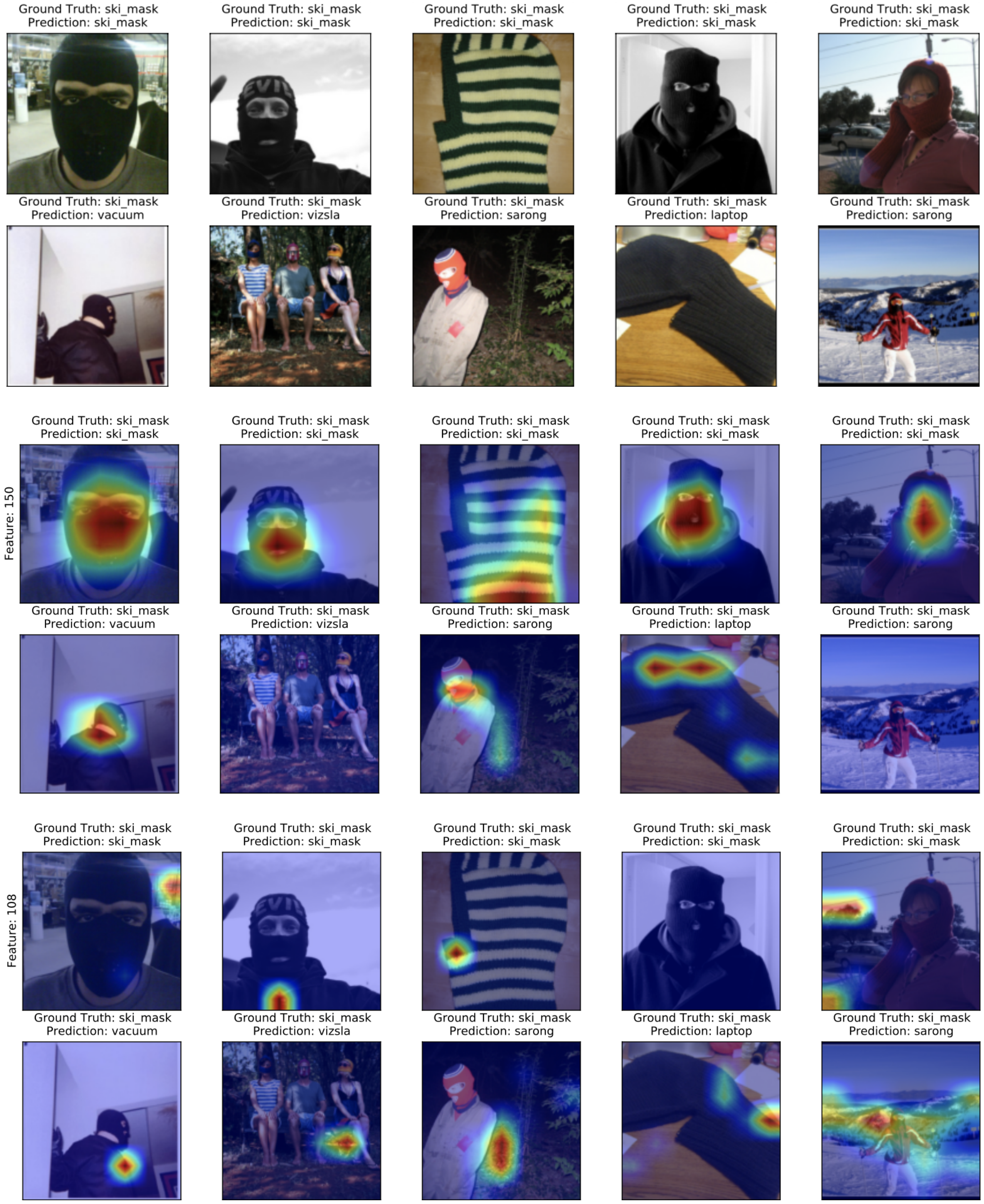

Figure A.13. **Heatmaps of Discriminative and Noisy Features of SimCLR (Class - Ski Mask)**: We plot the gradient heat maps of the most discriminative feature (by magnitude) of the given class and a noisy feature of the same class. We observe that discriminative features are more correlated with ground truth labels in correct classifications but are not correlated in misclassifications. Noisy features map to spurious portions of the images.

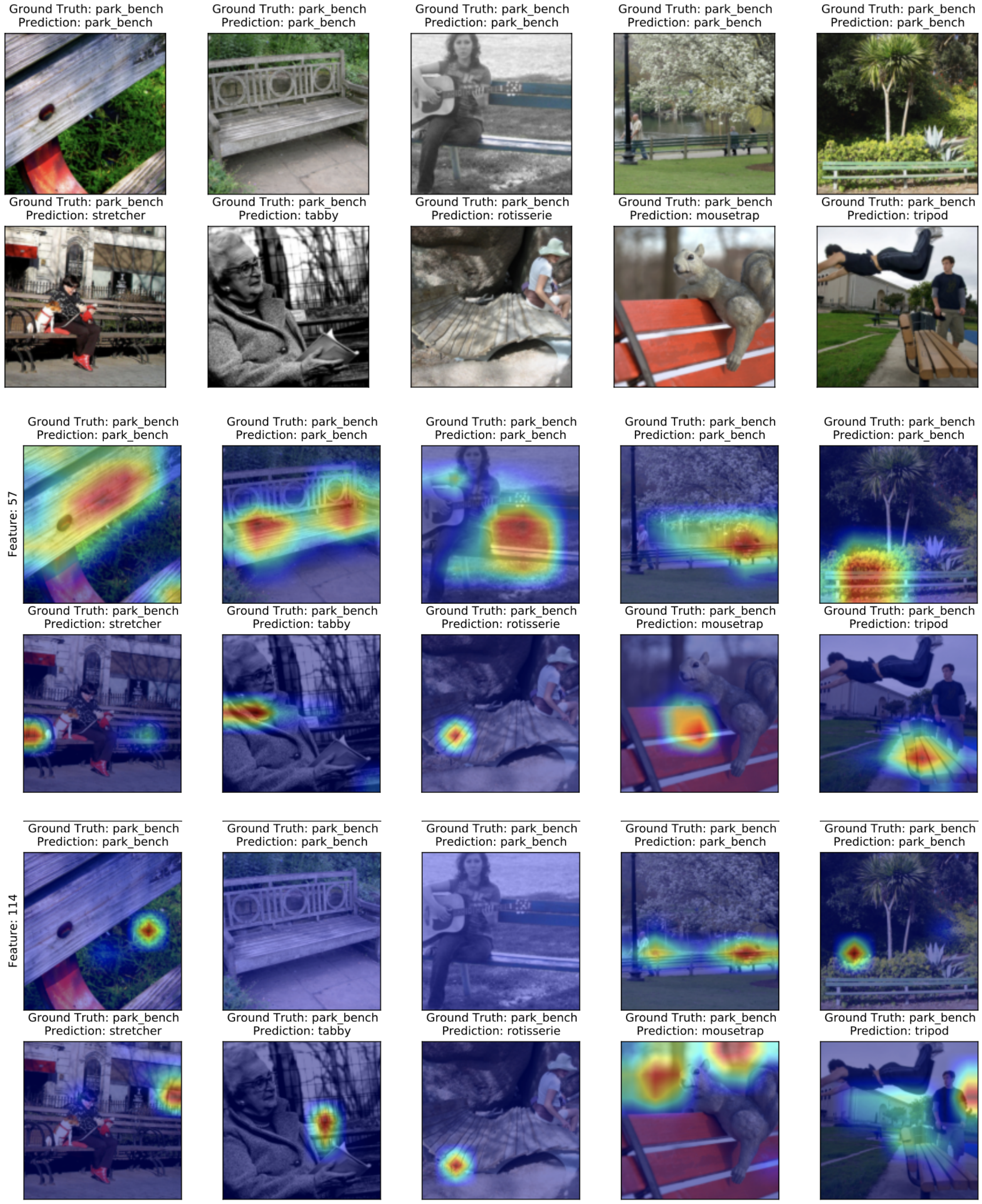

Figure A.14. **Heatmaps of Discriminative and Noisy Features of SimCLR (Class - Park Bench)**: We plot the gradient heat maps of the most discriminative feature (by magnitude) of the given class and a noisy feature of the same class. We observe that discriminative features are more correlated with ground truth labels in correct classifications but are not correlated in misclassifications. Noisy features map to spurious portions of the images.

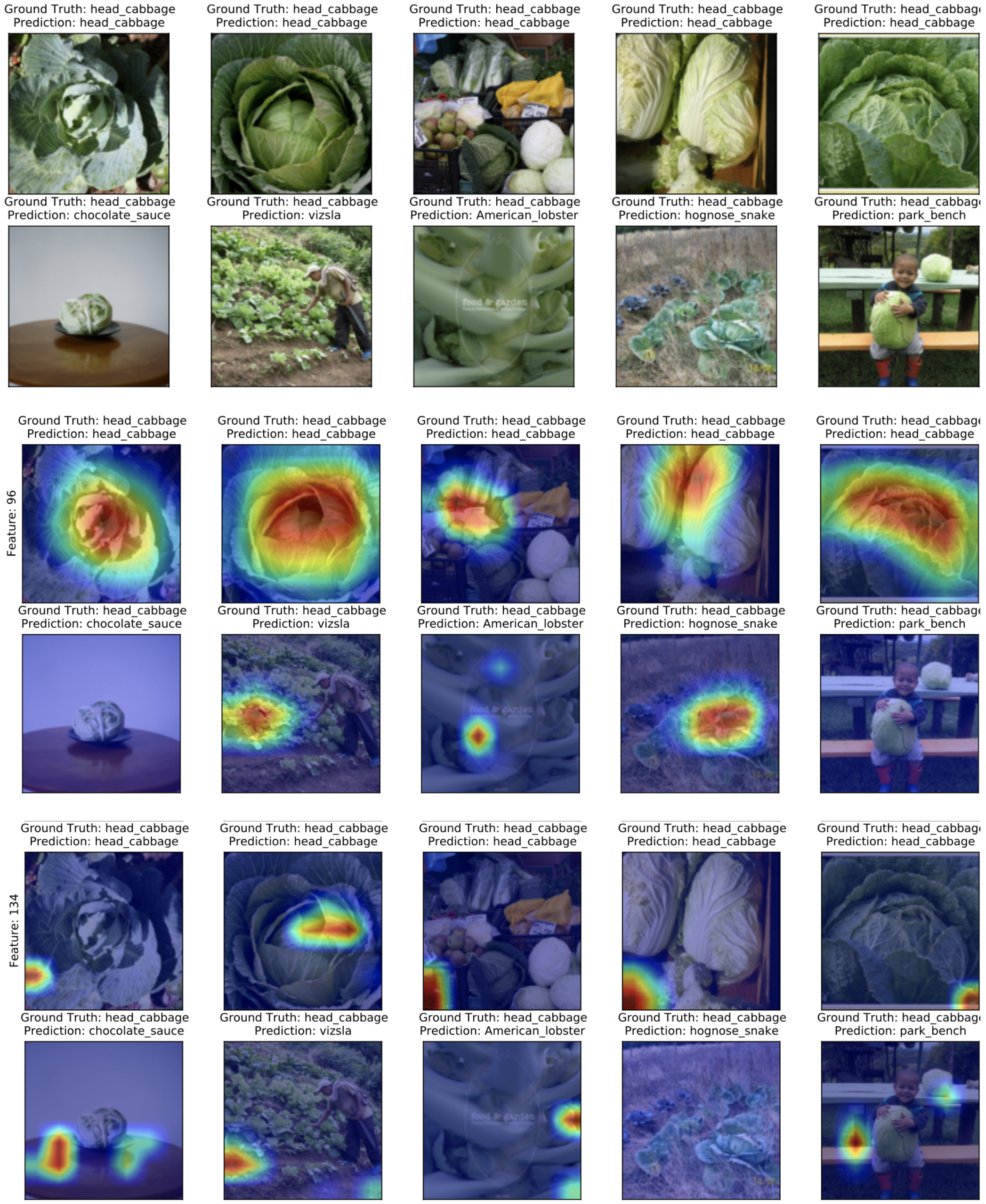
Ground Truth: head_cabbage
Prediction: head_cabbage
Ground Truth: head_cabbage
Prediction: head_cabbage
Ground Truth: head_cabbage
Prediction: head_cabbage
Ground Truth: head_cabbage
Prediction: head_cabbage
Ground Truth: head_cabbag
Prediction: head_cabbage
Ground Truth: head_cabbage
Prediction: chocolate_sauce
Ground Truth: head_cabbage
Prediction: vizsla
Ground Truth: head_cabbage
Prediction: American_lobster
Ground Truth: head_cabbage
Prediction: hognose_snake
Ground Truth: head_cabbag
Prediction: park_bench
Feature: 96
Ground Truth: head_cabbage
Prediction: head_cabbage
Ground Truth: head_cabbage
Prediction: head_cabbage
Ground Truth: head_cabbage
Prediction: head_cabbage
Ground Truth: head_cabbage
Prediction: head_cabbage
Ground Truth: head_cabbag
Prediction: head_cabbage
Ground Truth: head_cabbage
Prediction: chocolate_sauce
Ground Truth: head_cabbage
Prediction: vizsla
Ground Truth: head_cabbage
Prediction: American_lobster
Ground Truth: head_cabbage
Prediction: hognose_snake
Ground Truth: head_cabbag
Prediction: park_bench
Feature: 134
Ground Truth: head_cabbage
Prediction: head_cabbage
Ground Truth: head_cabbage
Prediction: head_cabbage
Ground Truth: head_cabbage
Prediction: head_cabbage
Ground Truth: head_cabbage
Prediction: head_cabbage
Ground Truth: head_cabbag
Prediction: head_cabbage
Ground Truth: head_cabbage
Prediction: chocolate_sauce
Ground Truth: head_cabbage
Prediction: vizsla
Ground Truth: head_cabbage
Prediction: American_lobster
Ground Truth: head_cabbage
Prediction: hognose_snake
Ground Truth: head_cabbag
Prediction: park_bench

Figure A.15. **Heatmaps of Discriminative and Noisy Features of SimCLR (Class - Head Cabbage)**: We plot the gradient heat maps of the most discriminative feature (by magnitude) of the given class and a noisy feature of the same class. We observe that discriminative features are more correlated with ground truth labels in correct classifications but are not correlated in misclassifications. Noisy features map to spurious portions of the images.

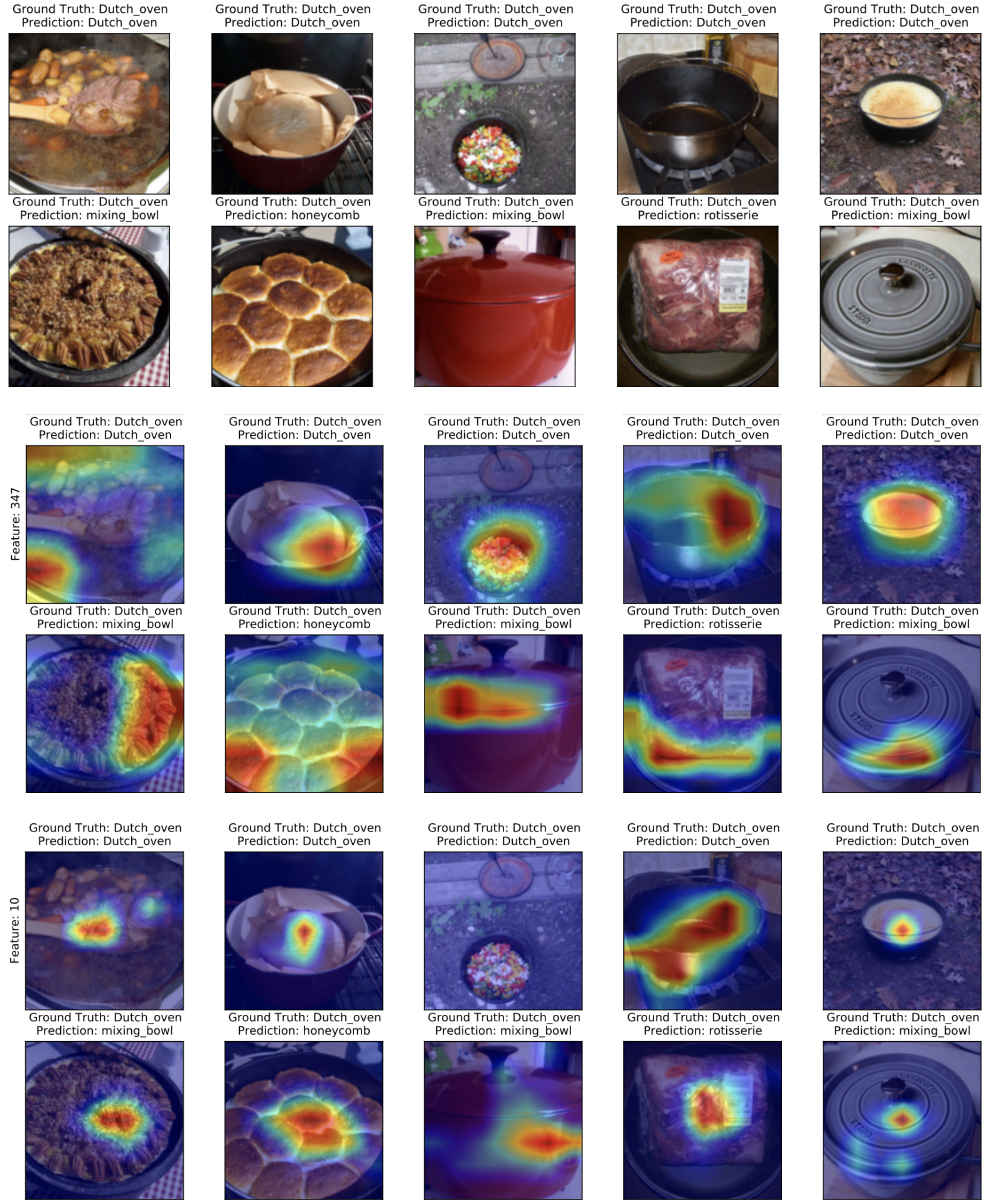

Figure A.16. **Heatmaps of Discriminative and Noisy Features of SimCLR (Class - Dutch Oven)**: We plot the gradient heat maps of the most discriminative feature (by magnitude) of the given class and a noisy feature of the same class. We observe that discriminative features are more correlated with ground truth labels of correct classifications but are not correlated in misclassifications. Noisy features map to spurious portions of the images.