Understanding the properties and limitations of contrastive learning for Out-of-Distribution detection

Nawid Keshtmand
Dept. Aerospace Engineering
University of Bristol
Email: yl18410@bristol.ac.uk

Raul Santos-Rodriguez
Dept. Engineering Mathematics
University of Bristol
Email:enrsr@bris.ac.uk

Jonathan Lawry
Dept. Engineering Mathematics
University of Bristol
Email: j.lawry@bris.ac.uk

Abstract—A recent popular approach to out-of-distribution (OOD) detection is based on a self-supervised learning technique referred to as *contrastive learning*. There are two main variants of contrastive learning, namely instance and class discrimination, targeting features that can discriminate between different instances for the former, and different classes for the latter.
In this paper, we aim to understand the effectiveness and limitation of existing contrastive learning methods for OOD detection. We approach this in 3 ways. First, we systematically study the performance difference between the instance discrimination and supervised contrastive learning variants in different OOD detection settings. Second, we study which in-distribution (ID) classes OOD data tend to be classified into. Finally, we study the spectral decay property of the different contrastive learning approaches and examine how it correlates with OOD detection performance. In scenarios where the ID and OOD datasets are sufficiently different from one another, we see that instance discrimination, in the absence of fine-tuning, is competitive with supervised approaches in OOD detection. We see that OOD samples tend to be classified into classes that have a distribution similar to the distribution of the entire dataset. Furthermore, we show that contrastive learning learns a feature space that contains singular vectors containing several directions with a high variance which can be detrimental or beneficial to OOD detection depending on the inference approach used

I. INTRODUCTION

In recent years, neural networks have been increasingly deployed for prediction tasks. However, they still make erroneous predictions when exposed to inputs from an unfamiliar distribution (1). This poses a significant obstacle to the deployment of neural networks in safety-critical applications. Therefore, for applications in these domains, it is necessary to be able to distinguish in-distribution (ID) data from data belonging to a different distribution on which the neural network was trained (OOD data). The problem of detecting such inputs is generally referred to as *out-of-distribution* (OOD) detection, outlier detection, or anomaly detection (2).

Recent studies have shown that a type of self-supervised learning referred to as *contrastive learning* can learn a feature space that is able to capture the features of an input x that enable effective OOD detection (3; 4; 5). The two main types of contrastive learning approaches are referred to as unsupervised instance discrimination and supervised contrastive learning/supervised class discrimination. The former works by bringing together the features of the different views of the same data point close together in the feature

space enabling discrimination between different instances (3). Instead, supervised contrastive learning brings data points belonging to the same class close together in feature space, which enables discrimination between data points belonging to different classes (6). In light of the variety of contrastive learning approaches used in practice, it is difficult to ascertain which properties of contrastive learning are beneficial to OOD detection. In this work, we aim to understand the effectiveness and limitations of existing contrastive learning methods for OOD detection. We approach this in 3 ways.

1) We systematically study the performance difference of the instance discrimination and supervised contrastive learning variants in different OOD detection settings (Section IV).
2) We investigate which ID classes OOD data tend to be classified into.
3) We analyze the spectral decay property of the different contrastive learning approaches and how it correlates with OOD detection performance. The spectral decay indicates how many different directions of high variance are present in the feature space and is related to the feature space compression (7).

This involves designing a fair training setup, categorizing the difficulty of OOD detection based on whether ID and OOD datasets are semantically similar (near-OOD) or unrelated (far-OOD), as well as comparing the OOD detection performance using established metrics (Section III). Our main findings and contributions are summarized below: The key contributions of the paper are as follows:

- We perform experiments that show unsupervised instance discrimination, in the absence of fine-tuning, is well-suited to far-OOD detection, where it can be competitive with supervised approaches but ineffective at near-OOD detection (Section IV).
- We analyze the classification of OOD samples and see that they tend to be classified into classes that have a distribution similar to the distribution of the entire dataset. This suggests that similar to the phenomena seen in deep generative models, discriminative models assign OOD samples to classes where the latent features capture non-class discriminatory information rather than class-discriminatory information (8) (Section V).

- We show that contrastive learning learns a feature space that contains singular vectors containing several directions with a high variance which can be detrimental or beneficial to OOD detection depending on the inference approach used (Section VI).

II. BACKGROUND

Most OOD detection methods use the features obtained from a classification network trained with the supervised cross-entropy loss. These classification networks learn decision boundaries that discriminate between the classes in the training set. However, using a conventional cross-entropy loss does not incentivize a network to learn features beyond the minimum necessary to discriminate between the different classes and therefore may not be well suited for OOD detection. An alternative to using classification networks for learning is contrastive learning (9). The idea of contrastive learning is to learn an encoder f_θ, where θ denotes the parameters of the encoder, to extract the necessary information to distinguish between similar and other samples. Let x be a query, whilst $\{x_+\}$ and $\{x_-\}$ denote the sets of positive samples of x and negative samples of x respectively. The goal of contrastive loss is to learn features of x, denoted $z = f_\theta(x)$, similar to the features of $\{x_+\}$ denoted $\{z_+\} = f_\theta(\{x_+\})$ whilst also being dissimilar to $\{z_-\} = f_\theta(\{x_-\})$. The most common form of similarity measure between features z and z' is given by the cosine similarity as given in Eqn. 1:

$$sim(z, z') = \frac{z \cdot z'}{\|z\| \, \|z'\|} \qquad (1)$$

The first type of contrastive loss considered in this paper is the Momentum Contrast (Moco) loss (10), where the objective relates to the task of instance discrimination and utilizes two different encoders, a query encoder f_{θ_q} and a key encoder f_{θ_k}. The parameters of the query encoder are updated using traditional backpropagation, whilst the parameters of the key encoder are an exponentially weighted average of the parameters of the query encoder, shown by Eqn. 2:

$$\theta'_k = m\theta_k + (1 - m)\theta_q \qquad (2)$$

where m is an encoder momentum hyperparameter, θ_q are parameters of the query encoder, and θ_k and θ'_k are the parameters of the key encoder before and after the update respectively.

The setup for the training process is as follows. Let $\widetilde{x}_i^1$ and $\widetilde{x}_i^2$ be two independent augmentations of x_i, namely $\widetilde{x}_i^1 = T_1(x_i)$ and $\widetilde{x}_i^2 = T_2(x_i)$ where T_1, T_2 are separate augmentation operators sampled from the same family of augmentations $\mathcal{T}$. The Moco objective treats each $(\widetilde{x}_i^1, \widetilde{x}_i^2)$ and $(\widetilde{x}_i^2, \widetilde{x}_i^1)$ as a query-positive key pair $x^{query}, x^{key_{pos}}$. Query data points are passed through a query encoder to obtain query features $q = f_{\theta_q}(x^{query})$ and key data points are passed through the key encoder to obtain the key features $k_{pos} = f_{\theta_k}(x^{key_{pos}})$. The contrastive loss for Moco, $\mathcal{L}_{moco}$ is then given by Eqn. 3:

$$\mathcal{L}_{moco} = -log\frac{q \cdot k_{pos}/\tau}{\sum_{i=0}^{K} exp(q \cdot k_i/\tau)} \qquad (3)$$

where k_{pos} is the positive key for each query q, whilst k_i are previous outputs of the key encoder which were saved in a queue and τ refers to a temperature hyperparameter.

The second type of contrastive loss considered in this paper is the Supervised Contrastive (SupCLR) loss (6). The SupCLR loss is a generalization of the instance discrimination loss which can deal with an arbitrary number of positives samples to enable contrasting samples in a class-wise manner instead of an instance-wise manner. In this case, all the data points in the same class as the query x^{query} are treated as positive samples $x^{key_{pos}}$ whilst data points in a different class as the query are treated as negative samples $x^{key_{neg}}$. The SupCLR loss only requires using a single encoder f_θ to obtain features for the queries q, positive keys k_{pos} and negative keys k_{neg} where $q = f_\theta(x^{query})$, $k_{pos} = f_\theta(x^{key_{pos}})$ and $k_{neg} = f_\theta(x^{key_{neg}})$. The SupCLR contrastive loss is then given by Eqn. 4:

$$\mathcal{L}_{SupCLR} = \frac{-1}{|P(q)|} \sum_{pos \in P(q)} log\frac{exp(q \cdot k_{pos}/\tau)}{\sum_{i \in A(q)} exp(q \cdot k_i/\tau)} \qquad (4)$$

Here, $P(q)$ is the set of indices of all positives keys of q which are present in the batch, and $|P(q)|$ is its cardinality. $A(q)$ is the set of all indices in the batch excluding q itself, this includes both k_{pos} and k_{neg}. Intuitively, the instance discrimination loss learns the features in each instance which makes it different from other instances whilst also being invariant to data augmentation, whilst the SupCLR loss learns the features in common between different instances of the same class.

III. OOD EXPERIMENTAL PRELIMINARIES

In this section, we discuss the experimental setup that we use throughout this work.

Datasets We study OOD detection for the following ID dataset (D_{in}) and OOD dataset (D_{out}) pairs. We use both simple grayscale dataset pairs and more complex RGB dataset pairs for the task. For the grayscale dataset case which uses MNIST, FashionMNIST, and KMNIST, one of the datasets is defined as D_{in} whilst the other datasets are defined as D_{out} (11; 12; 13). For the RGB datasets, we use CIFAR-10, CIFAR100 as well as SVHN and follows the same procedure where one dataset is defined as D_{in} whilst the others are defined as D_{out} (14; 15).

Similar to Winkens et al, we use the class-wise confusion log probability (CLP) as a measure of the difficulty of an OOD detection task (4). As described in Winkens et al, CLP is based on the probability of a classifier confusing OOD data with ID data. This involves computing an estimate of the expected probability of a test OOD sample x being predicted to be an ID class k by an ensemble of N classifiers as given by:

$$c_k(x) = \frac{1}{N} \sum_{j=1}^{N} p^j(\widehat{y} = k|x) \qquad (5)$$

where $p^j(\widehat{y} = k|x)$ refers to the probability that the jth classifier assigns x as belonging to class k. The class-wise confusion log probability (CLP) of ID class k for $\mathcal{D}_{test}$ becomes:

$$CLP_k(\mathcal{D}_{test}) = log\left(\frac{1}{\mathcal{D}_{test}}\sum_{x\in\mathcal{D}_{test}} c_k(x)\right) \qquad (6)$$

We compute the CLP with $\mathcal{D}_{test}$ being the test samples belonging to the OOD dataset. The distance between the dataset pairs is defined by the min-max bounds on the CLP (lowest CLP value to highest CLP value for the dataset pair). In our work, we use 3 Resnet-18 models to calculate $c_k(x)$. For further details on the Confusion Log probability, see Winkens et al (4), section 4.

Datasets	CLP bounds
ID:MNIST, OOD:FashionMNIST	-7.26 to -7.15
ID:MNIST, OOD:KMNIST	-7.29 to -6.88
ID:FashionMNIST, OOD:MNIST	-7.43 to -7.23
ID:FashionMNIST, OOD:KMNIST	-7.31 to -6.54
ID:KMNIST, OOD:MNIST	-6.87 to -6.03
ID:KMNIST, OOD:FashionMNIST	-7.22 to -6.96
ID:CIFAR10, OOD:SVHN	-7.25 to -6.27
ID:CIFAR10, OOD:CIFAR100	-4.80 to -3.53
ID:CIFAR100, OOD:SVHN	-9.08 to -7.52
ID:CIFAR100, OOD:CIFAR10	-7.74 to -4.94

TABLE I
CLASS-WISE CONFUSION LOG PROBABILITY (CLP) MIN-MAX BOUNDS FOR THE DIFFERENT ID-OOD DATASET PAIRS.

The min-max bounds of the CLP for the different dataset pairs are given in Table I. For the purpose of categorizing the different dataset pairs, we label pairs that have both values for the bounds of the CLP above 6.5 as a far-OOD dataset pair, below 6.5 as near-OOD dataset pair, and dataset pairs which have bounds that have a value above and below 6.5 as a near & far-OOD dataset pair. The CLP values were chosen by hand so as to separate the datasets and enable comparison.

Metrics We measure the quality of OOD detection using the established metrics for this task, which are the AUROC, AUPR and the FPR at 95% TPR (16). For the AUPR, we treat OOD examples as the positive class. Unless otherwise stated, metrics reported in this work are obtained based on 8 repeat readings.
Training Setup Experiments were conducted using 2 Tesla P100-PCIE-16GB with 28 CPUs using Pytorch Version 1.7.1 and Pytorch Lightning Version 1.2.3 (17; 18). We adopt a Resnet-50 as the encoder f_θ for all the different models with a fixed-dimensional output dimensionality of 128-D with the outputs being l2 normalized (19). For the case of the classification network, also referred to as the Cross-Entropy (CE) model, the output of the Resnet-50 encoder is followed by an additional fully connected layer with an output dimensionality equal to the number of classes in the ID training data.

For the case of the Moco model, both the query encoder and the key encoder have the same architecture. All models are trained for 300 epochs with a batch size of 256, using the SGD optimizer with a learning rate of $3e^{-2}$, optimizer

momentum of 0.9 and weight decay of $1e^{-4}$, queue size of 4096, and an encoder momentum m of 0.999. Furthermore, both the Moco and SupCLR model use a softmax temperature τ of 0.07 in the loss function.

For data augmentation, we use random crop and resize (with random flip), color distortion (for the RGB datasets), and Gaussian blur.

IV. INSTANCE DISCRIMINATION IS EFFECTIVE FOR FAR-OOD DETECTION

A. Hypothesis

We hypothesized that as instance discrimination requires being able to distinguish each individual sample from one another, the model trained using instance discrimination will output features z which retains a large amount of information about the input x. By retaining a large amount of information regarding an input, the Moco model could achieve high OOD performance. This depends on how semantically similar the ID and OOD points are to one another and what information the inference approach focuses on.

B. Procedure

To investigate when the instance discrimination loss is effective, we compare the Moco model with the SupCLR and CE model to see the difference when using training labels in the training process. Furthermore, we look at the effect of supervised labels during inference by examining the OOD detection performance of different models across several different dataset pairs using two different inference approaches. The inference approaches used are the Mahalanobis Distance with class-dependent covariance matrices (4) which use class labels, as well as the kernel density estimation (KDE) approach which does not use class labels.

C. Results

For the case of the KDE approach, Table II shows that the Moco model consistently performed worse. This showed that in the absence of any training labels, the instance discrimination training makes it difficult to group ID data together and therefore leads to poor OOD detection performance. Furthermore, the CE model is generally the highest performing model across the metrics, even outperforming SupCLR. This shows that in the absence of labels during the inference process, the CE model is best able to group the data in the feature space such that the OOD test data are farther than the ID test data from the ID training data points.

For the task of OOD detection using the Mahalanobis Distance inference approach, Table III shows that for the grayscale datasets, the Moco model is competitive with the CE model and the SupCLR model achieves the best results. This shows that for the grayscale datasets, the features learned from contrastive learning are able to compete with or outperform the CE model. However, for the case of the RGB datasets, the Moco approach outperforms the CE model on the CIFAR10-SVHN dataset pair and outperforms SupCLR on the CIFAR100-SVHN dataset pair, both of which are categorized as far-OOD.

ID	OOD	AUROC	AUPR	FPR
		CE/Moco/SupCLR		
CIFAR10	SVHN	0.950 / 0.574 / **0.962***	0.975 / 0.779 / **0.980***	**0.166** / 0.887 / 0.170
	CIFAR100	**0.883** / 0.576 / 0.882	0.863 / 0.591 / **0.864***	**0.399** / 0.888 / 0.415
	MNIST	**0.939** / 0.462 / 0.900	**0.920** / 0.437 / 0.880	**0.219** / 0.875 / 0.346
	FashionMNIST	**0.955** / 0.514 / 0.943	**0.943** / 0.487 / 0.933	**0.162** / 0.887 / 0.232
	KMNIST	**0.941** / 0.472 / 0.935	0.911 / 0.444 / **0.915***	**0.188** / 0.893 / 0.233
CIFAR100	SVHN	**0.853** / 0.567 / 0.828	**0.907** / 0.767 / 0.897	**0.415** / 0.879 / 0.482
	CIFAR10	0.726 / 0.569 / **0.736***	0.680 / 0.562 / **0.693***	0.784 / 0.881 / **0.722***
	MNIST	**0.646** / 0.307 / 0.594	**0.620** / 0.378 / 0.564	**0.719** / 0.964 / 0.775
	FashionMNIST	**0.851** / 0.522 / 0.841	**0.812** / 0.484 / 0.806	**0.458** / 0.881 / 0.485
	KMNIST	0.757 / 0.403 / **0.769***	0.720 / 0.413 / **0.736***	**0.608** / 0.945 / 0.643
MNIST	FashionMNIST	0.984 / 0.587 / **0.992***	0.979 / 0.621 / 0.99*	0.080 / 0.899 / **0.034***
	KMNIST	0.957 / 0.616 / **0.982***	0.958 / 0.694 / 0.98*	0.218 / 0.899 / **0.077***
FashionMNIST	MNIST	**0.807** / 0.624 / 0.754	**0.818** / 0.606 / 0.804	0.614 / 0.821 / 0.820
	KMNIST	**0.821** / 0.638 / 0.791	**0.822** / 0.611 / 0.821	**0.576** / 0.771 / 0.738
KMNIST	MNIST	**0.972** / 0.656 / 0.971	**0.970** / 0.712 / 0.966	0.108 / 0.822 / **0.107***
	FashionMNIST	**0.985** / 0.620 / **0.985**	**0.984** / 0.621 / 0.978	**0.056** / 0.818 / **0.056**

TABLE II
AUROC, AUPR AND FPR FOR DIFFERENT MODELS ON DIFFERENT ID-OOD DATASET PAIRS USING THE KDE APPROACH AND * INDICATES THAT AN APPROACH IS BETTER THAN THE CE BASELINE.

ID	OOD	AUROC	AUPR	FPR
		CE/Moco/SupCLR		
CIFAR10	SVHN	0.891 / 0.908* / **0.950***	0.939 / 0.960* / **0.973***	0.283 / 0.408 / **0.187***
	CIFAR100	0.875 / 0.784 / **0.902***	0.851 / 0.778 / **0.892***	0.418 / 0.657 / **0.386***
	MNIST	0.931 / **0.988*** / 0.951*	0.906 / **0.985*** / 0.935*	0.214 / **0.051*** / 0.176*
	FashionMNIST	0.945 / 0.971* / **0.973***	0.926 / **0.971*** / 0.965*	0.180 / 0.143* / **0.111***
	KMNIST	0.929 / 0.952* / **0.956***	0.892 / **0.941*** / 0.933*	0.195 / 0.174* / **0.142***
CIFAR100	SVHN	**0.858** / 0.828 / 0.818	0.900 / **0.908*** / 0.870	0.420 / 0.565 / **0.413***
	CIFAR10	0.731 / 0.614 / **0.746***	0.680 / 0.574 / **0.698***	0.730 / 0.836 / **0.683***
	MNIST	**0.656** / 0.398 / 0.625	**0.623** / 0.467 / 0.604	**0.754** / 0.964 / 0.811
	FashionMNIST	0.886 / 0.895* / **0.902***	0.856 / **0.880*** / 0.877*	0.375 / 0.403 / **0.319***
	KMNIST	0.777 / 0.604 / **0.798***	0.730 / 0.595 / **0.762***	**0.588** / 0.887 / 0.599
MNIST	FashionMNIST	0.988 / **0.997*** / 0.990*	0.985 / **0.995*** / 0.983	0.018 / **0.012*** / 0.028
	KMNIST	0.984 / 0.972 / **0.993***	0.982 / 0.962 / **0.991***	0.061 / 0.108 / **0.030***
FashionMNIST	MNIST	0.921 / **0.980*** / 0.971*	0.933 / **0.980*** / 0.969*	0.456 / **0.105*** / 0.138*
	KMNIST	0.941 / 0.966* / **0.972***	0.950 / 0.965* / **0.969***	0.309 / 0.173* / **0.138***
KMNIST	MNIST	0.973 / 0.957 / **0.991***	0.970 / 0.954 / **0.989***	0.089 / 0.208 / **0.037***
	FashionMNIST	0.962 / 0.982* / **0.984***	0.953 / **0.978*** / 0.973*	0.219 / 0.076* / **0.043***

TABLE III
AUROC, AUPR AND FPR FOR DIFFERENT MODELS ON DIFFERENT ID-OOD DATASET PAIRS USING THE MAHALANOBIS DISTANCE INFERENCE APPROACH AND * INDICATES THAT AN APPROACH IS BETTER THAN THE CE BASELINE

In contrast, for the case of CIFAR10-CIFAR100 (near-OOD) and CIFAR100-CIFAR10 (near & far-OOD), Moco performs significantly worse than both supervised approaches. This suggests that for far-OOD detection tasks, it can be beneficial to use the unsupervised instance discrimination training to perform OOD detection, although, class discriminatory information obtained from supervised training is important for high performance in near-OOD detection settings. Furthermore, it can be seen that using class labels to group the features during the inference process leads to a larger improvement in the contrastive learning models compared to the CE model as shown by a larger improvement in the contrastive models when using the Mahalanobis Distance inference approach rather than KDE.

V. OOD SAMPLES ARE CLASSIFIED INTO CLASSES THAT LEARN NON-CLASS DISCRIMINATORY STATISTICS

A. Hypothesis

We hypothesized that the more similar a class distribution is to the distribution of the entire dataset, which we refer to as the overall distribution, the more likely it is that OOD data will be misclassified as that class. The intuition here is that OOD data tends to be misclassified as belonging to the least distinctive class. We believe that the similarity of a class distribution to the overall distribution indicates how much the features of a particular class contain non-class discriminatory information, which indicates how distinctive the class distribution is.

B. Procedure

To quantify the dissimilarity of the class distribution to the overall distribution, we approximated the class distribution as well as the overall distribution by Gaussian distributions and

calculated the KL divergence between the overall distribution and class distribution, $KL\left(Overall\|Class\right)$. Furthermore, the expected Overall-Class KL as shown by Eqn. 7 was calculated for the contrastive models to see whether there was any difference in this metric between the two approaches.

$$\mathbb{E}\left[KL\left(Overall\|Class\right)\right] = \sum_{k}^{K} KL\left(Overall\|Class_k\right) \quad (7)$$
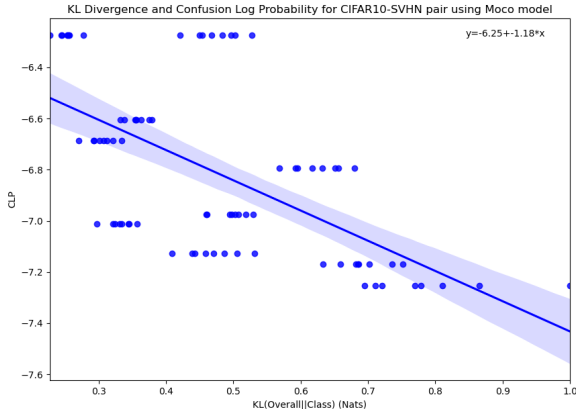
C. Results

Fig. 1. Scatter plot of the normalized KL divergence between the overall distribution and class distribution, $KL\left(Overall\|Class\right)$ (normalized by dividing by the maximum $KL\left(Overall\|Class\right)$), and the class-wise confusion log probability (CLP).

From Fig. 1, it can be seen that there is a negative correlation between the KL divergence and the class-wise confusion log probability. This supports our hypothesis that the more dissimilar the class distribution is to the overall distribution, the more likely OOD data points will be misclassified into that ID class. This behavior agrees with the phenomena seen in deep generative models where OOD data points are given high likelihoods by arbitrarily having low compressed lengths and being similar to the non-class discriminatory statistics of the data (8).

Datasets	Moco	SupCLR
MNIST	173.790	567.539
FashionMNIST	249.147	1484.179
KMNIST	122.158	468.819
CIFAR10	44.395	193.645
CIFAR100	96.943	1603.609

TABLE IV
EXPECTED OVERALL-CLASS KL BETWEEN THE OVERALL DISTRIBUTION AND CLASS DISTRIBUTION, $\mathbb{E}\left[KL\left(Overall\|Class\right)\right]$, WHEN USING DIFFERENT LEARNING APPROACHES.

From Table IV, it can be seen that the Moco model had expected Overall-Class KL values smaller than the SupCLR model. We believe this explains why the Moco model performs poorly on the near-OOD detection pairs such as CIFAR10-CIFAR100 and CIFAR100-CIFAR10. For the case of Moco where there is a small KL divergence between the overall and class distributions, it indicates that the features from Moco capture general non-class discriminatory information which is common in the ID dataset. This can make near-OOD detection difficult as the OOD dataset is likely to have similar non-class discriminatory statistics in common with the ID dataset. However, in the case of a far-OOD detection situation, the non-class discriminatory statistics are sufficiently different between the two datasets to enable Moco to effectively detect OOD data.

VI. NUMBER OF DIRECTIONS OF SIGNIFICANT VARIANCE

A. Hypothesis

From the OOD detection results (Section IV), contrastive learning results in poorer performance in OOD detection than the traditional CE model when using KDE than when using the Mahalanobis Distance approach. We claim that this is, at least in part, due to contrastive learning trained models learn a rich feature space with several different directions of high variance (4).

In the situation where KDE is used for inference, having a feature space with several directions of variance is likely to result in the OOD data having similar features to the ID data, leading to the misclassification of OOD data as ID. However, in the situation where the Mahalanobis Distance inference approach is used, using class labeled data can help to identify which factors of variation are important for class discrimination which can be used to aid in discriminating the ID data from the OOD data.

B. Procedure

To quantify how rich a feature space is, we investigated the number of factors of variation in the features by following the procedure taken in Section 5 of Roth et al which involves calculating the spectral decay ρ (7). This involves computing the singular value decomposition (SVD) of the features of the training data, normalizing the sorted spectrum of singular values, and computing the KL-divergence between a D-dimensional discrete uniform distribution and the sorted spectrum of singular values. Using this metric, lower values of ρ indicate more directions of significant variance.

C. Results

From Fig. 2, it can be seen that both contrastive models generally have lower values of ρ than the CE model, with the exception of the CIFAR100 dataset where the CE model has a lower ρ than the SupCLR approach. As the supervised SupCLR model was able to outperform the supervised CE model on all the datasets in which the SupCLR model had a lower ρ value when using the Mahalanobis Distance, this suggests that having a large number of significant directions of variance is an important property for OOD detection. The Moco model consistently has the lowest ρ which indicates its features have the largest number of factors of variation. This is likely due to the difficult task of instance discrimination requiring learning a large number of factors of variation to
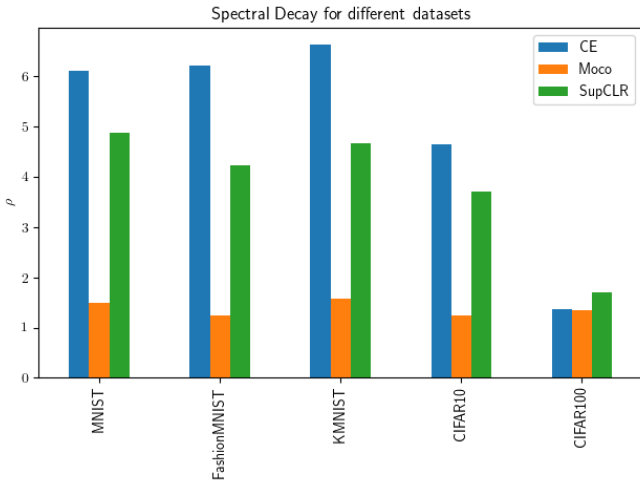
Fig. 2. Spectral decay for the different training datasets using different learning approaches.

perform the task. Whereas, the CE and SupCLR models only require learning features needed to discriminate data between different classes resulting in a smaller number of significant factors of variation.

VII. RELATED WORK

Recent approaches in OOD detection can be categorized as calibration-based (20; 21), density estimation-based (22), and self-supervised learning-based. Self-supervised learning-based approaches can be split into auxiliary pretext task-based (23), contrastive based (4) or approaches that utilize aspects of both auxiliary pretext tasks and contrastive learning (5; 24). The contrastive learning approaches vary significantly and our work also falls in the category of investigating contrastive learning for OOD detection. Some approaches use instance discrimination-based pretraining followed by fine-tuning with a supervised cross-entropy loss (4; 25). Both these papers and ours focus on the instance discrimination contrastive learning task, however, our work focuses on instance discrimination in the absence of fine-tuning with a supervised loss. Other approaches do not use the pretraining followed by fine-tuning paradigm and instead focus on using a supervised contrastive loss for OOD detection when labels are available (26; 5; 27). Our work also examines the performance of the supervised contrastive loss for OOD detection, although we use the original supervised contrastive loss for the purpose of analyzing the properties present in contrastive learning which makes it effective for OOD detection. Furthermore, our work differs from all the aforementioned work as it is the first to examine the differences between using a supervised contrastive compared to an instance discrimination approach as well as the effect of different inference methods on OOD detection performance.

VIII. CONCLUSION

In this work, we investigated the effectiveness and limitations of two popular contrastive learning approaches for the problem of OOD detection. In our work, we saw that:

- Contrastive learning learns a larger number of factors of variation compared to classification networks. This can be detrimental to OOD detection performance when using KDE (unsupervised inference) but beneficial when using Mahalanobis distance (supervised inference).
- Unsupervised instance discrimination learns a richer feature space with a wide variety of factors of variation compared to supervised contrastive learning as shown by analyzing the spectral decay.
- By having a large number of factors of variation, instance discrimination can be more effective than the Cross-Entropy baseline at far-OOD detection.
- Despite having a larger number of factors of variation than the supervised contrastive Learning, instance discrimination is ineffective at near-OOD detection.
- Instance discrimination is ineffective at near-OOD detection as it leads to a poorer separation between the different classes in the feature space which was shown by having a lower $\mathbb{E}[KL(Overall\|Class)]$. This is a result of learning factors of variation which contain non-class discriminatory information.

A. Limitations and Future Work

We would also like to point out some limitations of our study. Firstly, we focus on two different contrastive learning methods which uses explicit negatives (e.g. MoCo and SupCLR). We chose to focus on these two methods as these methods are quite general contrastive learning approaches where the only significant difference is is the use of supervision. Therefore, comparing the results of the two methods should give an indication of the effect of supervision in OOD detection performance. We believe that other methods based on explicit negatives such as the InfoNCE and Triplet losses would also exhibit similar behaviour (3; 9). An avenue of future work could examine the effect of different properties of self-supervised approaches on OOD detection performance. Example of such properties could be whether a self-supervised approach uses explicit negatives or not, as well as the effect of redundancy reduction principle from Barlow Twins (28). Furthermore, after identifying that OOD samples tend to be allocated to classes that have a distribution similar to the overall distribution, future work could involve investigating whether it is possible to regularise the contrastive loss to prevent each class from being too similar to the overall distribution. This could enable the instance discrimination contrastive loss to learn a rich feature space whilst also learning class-specific semantic features which can improve near-OOD detection.

REFERENCES

[1] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE*

conference on computer vision and pattern recognition, 2015, pp. 427–436.

[2] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Müller, "A unifying review of deep and shallow anomaly detection," *Proceedings of the IEEE*, 2021.

[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[4] J. Winkens, R. Bunel, A. G. Roy, R. Stanforth, V. Natarajan, J. R. Ledsam, P. MacWilliams, P. Kohli, A. Karthikesalingam, S. Kohl *et al.*, "Contrastive training for improved out-of-distribution detection," *arXiv preprint arXiv:2007.05566*, 2020.

[5] J. Tack, S. Mo, J. Jeong, and J. Shin, "Csi: Novelty detection via contrastive learning on distributionally shifted instances," *arXiv preprint arXiv:2007.08176*, 2020.

[6] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *arXiv preprint arXiv:2004.11362*, 2020.

[7] K. Roth, T. Milbich, S. Sinha, P. Gupta, B. Ommer, and J. P. Cohen, "Revisiting training strategies and generalization performance in deep metric learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8242–8252.

[8] J. Serrà, D. Álvarez, V. Gómez, O. Slizovskaia, J. F. Núñez, and J. Luque, "Input complexity and out-of-distribution detection with likelihood-based generative models," *arXiv preprint arXiv:1909.11480*, 2019.

[9] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 539–546.

[10] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[11] Y. LeCun, "The mnist database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*, 1998.

[12] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[13] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha, "Deep learning for classical japanese literature," *arXiv preprint arXiv:1812.01718*, 2018.

[14] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[15] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," 2011.

[16] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[17] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[18] e. a. Falcon, WA, "Pytorch lightning," *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, vol. 3, 2019.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," *Advances in neural information processing systems*, vol. 31, 2018.

[21] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," *arXiv preprint arXiv:1706.02690*, 2017.

[22] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," *arXiv preprint arXiv:1802.10501*, 2018.

[23] I. Golan and R. El-Yaniv, "Deep anomaly detection using geometric transformations," *arXiv preprint arXiv:1805.10917*, 2018.

[24] K. Sohn, C.-L. Li, J. Yoon, M. Jin, and T. Pfister, "Learning and evaluating representations for deep one-class classification," *arXiv preprint arXiv:2011.02578*, 2020.

[25] A. G. Roy, J. Ren, S. Azizi, A. Loh, V. Natarajan, B. Mustafa, N. Pawlowski, J. Freyberg, Y. Liu, Z. Beaver *et al.*, "Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions," *arXiv preprint arXiv:2104.03829*, 2021.

[26] V. Sehwag, M. Chiang, and P. Mittal, "Ssd: A unified framework for self-supervised outlier detection," *arXiv preprint arXiv:2103.12051*, 2021.

[27] H. Cho, J. Seol, and S.-g. Lee, "Masked contrastive learning for anomaly detection," *arXiv preprint arXiv:2105.08793*, 2021.

[28] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 310–12 320.