Is it all a cluster game? - Exploring Out-of-Distribution Detection based on Clustering in the Embedding Space

Poulami Sinhamahapatra,[1] **Rajat Koner,** [2] **Karsten Roscher,**[1] **Stephan Günnemann** [3]

[1]Fraunhofer-Institut für Kognitive Systeme IKS, [2] Ludwig Maximilian University of Munich
[3] Technical University of Munich

Abstract

It is essential for safety-critical applications of deep neural networks to determine when new inputs are significantly different from the training distribution. In this paper, we explore this out-of-distribution (OOD) detection problem for image classification using clusters of semantically similar embeddings of the training data and exploit the differences in distance relationships to these clusters between in- and out-of-distribution data. We study the structure and separation of clusters in the embedding space and find that the supervised contrastive learning leads to well separated clusters while its self-supervised counterpart fails to do so. In our extensive analysis of different training methods, clustering strategies, distance metrics and thresholding approaches, we observe that there is no clear winner. The optimal approach depends on the model architecture and selected datasets for in- and out-of-distribution. While we could reproduce the outstanding results for contrastive training on CIFAR-10 as in-distribution data, we find standard cross-entropy paired with cosine similarity outperforms all contrastive training methods when training on CIFAR-100 instead. Cross-entropy provides competitive results as compared to expensive contrastive training methods.

1 Introduction

The recent success of Deep Neural Networks (DNN) has motivated their application in an varity of tasks. While DNNs have demonstrated remarkable performance, they cannot be expected to work reliably on inputs that are not represented by the training distribution. Such out-of-distribution (OOD) samples can lead to unpredictable behaviour and overconfident predictions [Nguyen et al. 2015, Guo et al. 2017, Hendrycks & Gimpel 2018], with severe consequences in case of safety-critical applications like autonomous driving or automated medical diagnoses. Therefore, it is crucial to detect such inputs when applied to the model to allow for additional fallback measures to be triggered [Henne et al. 2019] or to abstain from automated decisions in rare or unseen situations [Zhou et al. 2021, Prabhu et al. 2018].

One promising research direction for out-of-distribution detection - especially in image classification - is to exploit the distribution of training samples in the learnt embedding space assuming that related images exhibit similar features and are therefore in close proximity according to their latent representation [Lee et al. 2018]. Since contrastive learning (CL) methods [Khosla et al. 2020, Chen et al. 2020a] are supposed to improve the separability of instances or samples in the embedding space by pulling similar inputs together and pushing dissimilar ones apart, it is only natural that their use for OOD detection based on latent representations has demonstrated state-of-the-art results recently [Sehwag et al. 2021]. Figure 1 illustrates the intuition behind those approaches.

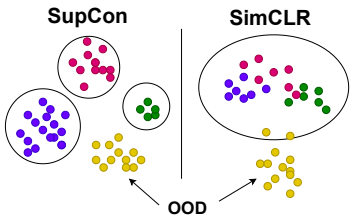

Figure 1: Simplified illustration of the expected latent space clusters for supervised (SupCon) and unsupervised (SimCLR) contrastive training methods

However, while the results are promising, several aspects are left unexplored. On the one hand, there is the question to which extent different training methods really influence the formation of well-defined clusters of in-distribution (ID) samples in the embedding space where different distance metrics may be applied to measure sample similarity. On the other hand, clustering strategies and the optimal number of clusters have barely been touched in existing literature, with the odd choice of a single cluster representing all the ID data apparently leading to the best results [Sehwag et al. 2021]. Since the use of machine learning in safety-critical contexts depends on a sound understanding of the insufficiencies and expected failure modes of the deployed models [Burton et al. 2021], we conduct an extensive study on the performance of supervised and self-supervised contrastive learning methods for OOD detection focussing on the following contributions:

- **Structure of the embedding space:** In Section 3, we provide detailed insights into cluster formations across

supervised (SupCon) and unsupervised (SimCLR) contrastive learning methods, by using *Global Separation* and *Cluster Purity* metrics to analyse cluster quality. We find that supervised training leads to well-separated clusters, while unsupervised training leads to closely overlapping clusters.

- **OOD detection based on clustering:** In Section 4, we provide a modular OOD detection approach based on the similarity of an input sample to a set of clusters allowing the comparison of different distance metrics, clustering methods and thresholding strategies. We further investigate, whether observations are consistent for different models and data sets. Our results indicate that there is no clear winner: the optimal combination indeed depends on the model size, distance metrics and training data.

2 Related Work

DNNs are increasingly used in tasks like classification[Dosovitskiy et al. 2020],scene prediction[Koner et al. 2021c, 2020] and other high level tasks such as reasoning [Hildebrandt et al. 2020, Koner et al. 2021a]. However, presence of OOD samples presents an important concern in the successful completion of all such tasks, particularly in safety-critical systems. Thus, reliable OOD detection has become an important direction of research.

Out-of-Distribution Detection The problem of OOD detection has often been formulated as outlier detection [Hodge & Austin 2004, Sehwag et al. 2021], one-class classification [Ruff et al. 2018, Perera et al. 2019], novelty detection [Tack et al. 2020, Pidhorskyi et al. 2018], anomaly detection [Golan & El-Yaniv 2018, Hendrycks et al. 2019a] and open set recognition [Boult et al. 2019, Geng et al. 2020]. Some contemporary ways to approach the problem are: density approximation based generative modelling [Ren et al. 2019, Nalisnick et al. 2019], self-supervision to learn discriminatory features [Hendrycks et al. 2019b, Mohseni et al. 2020, Tack et al. 2020, Sehwag et al. 2021], softmax score based classifier methods [Hendrycks & Gimpel 2018, Liang et al. 2020], detection score based methods [Lee et al. 2018, Winkens et al. 2020, Tack et al. 2020], utilisation of uncertainty quantifications based methods [Schwaiger et al. 2020, Charpentier et al. 2020] as well as methods using self-attention based transformers [Koner et al. 2021b]. Since OOD samples can vary in many different ways, many outlier exposure methods use few known OOD samples, thus inducing a form of prior knowledge of OOD [Lee et al. 2018, Hendrycks et al. 2019a, Liang et al. 2020]. However, this approach could lead to problems when generalising across diverse novel OOD datasets. Many contemporary works have explored multi-class OOD detection settings without inducing prior bias for OOD samples, but often do not perform well with only near-OOD data, i.e. semantically similar from ID data. In such a scenario, instance based discriminatory method like CL can be used to learn useful semantic features.

Contrastive Learning: Discriminative approaches using contrastive loss [Bachman et al. 2019, Hjelm et al. 2019] had shown great promise in the past, however recently CL has found even greater application in multiple application domains [Henaff 2020, Tack et al. 2020, Sehwag et al. 2021] following the success of self-supervised methods like SimCLR [Chen et al. 2020a], MoCov2 [Chen et al. 2020b], etc as well as Supervised CL method [Khosla et al. 2020] and similar. Recent works like [Sehwag et al. 2021, Tack et al. 2020, Winkens et al. 2020] have employed contrastive training for OOD detection by either modifying the contrastive training objective or assuming inherent class-conditioned clusters. As a novel contribution, we present an extensive study into the quality of clusters formed by various contrastive training approaches and their influence on OOD detection.

3 Structure of the embedding space

In this section, we investigate potential clusters in the embedding space and address the question about how to evaluate the quality of clusters as well as their separation in the high-dimensional embedding space.

3.1 Contrastive Learning towards clustering

The key intuition behind any CL method is to preserve a meaningful representation by maximising the agreement between similar instances and at the same time minimising the agreement with dissimilar instances. This means that, given an anchor image and a set of positives and negatives, the positives are pulled closer based on similarity with the anchors while the negatives are pushed apart in the embedding space. In this work, we focus on Supervised Contrastive Learning (SupCon) [Khosla et al. 2020] and the unsupervised approach SimCLR [Chen et al. 2020a] and compare them to a baseline trained with standard cross-entropy (CE) loss.

SimCLR uses strong data augmentation to compare with positive instances of an anchor image to learn without supervision. While this promotes discriminative feature learning, it often leads to disagreement between instances of same classes. *SupCon* tries to address this caveat, by increasing the number of positives by using all the samples of the same class/ same ground-truth (GT) labels for comparison with each anchor during CL. Both methods apply the contrastive loss, based on cosine similarity, at a lower-dimensional non-linear projection layer. Since this layer is trained to be invariant to augmentations, the authors [Chen et al. 2020a] suggest that good quality representations are most likely to be preserved at the last *feature* layer of the encoder.

3.2 Determining Cluster Quality

In our investigation of cluster quality, we want to determine how well clusters are separated from each other. Clusters are either formed based on class labels corresponding to GT classes or by using additional clustering like k-means on the embedding vectors of all training samples. In the case of k-means clustering, we are further interested to understand if these clusters reflect semantically similar samples, e.g. with samples belonging to the same class.

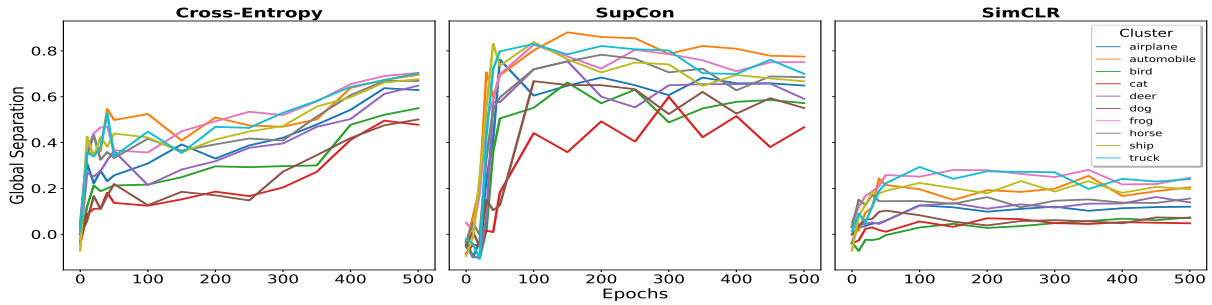

Figure 2: Evolution of Global Separation between clusters based on GT classes over training epochs for different contrastively trained models on CIFAR-10 as compared to baseline CE.

Global Separation (GS) [Bojchevski et al. 2017] generalises and extends the idea of the Silhouette Coefficient [Rousseeuw 1987]. GS utilises the intuition that separability between clusters can be determined by inspecting intra-cluster and inter-cluster distances. Thus, for each cluster c, a list of pairwise distances $P_{c,c}$ is calculated for all samples within the same cluster (intra-cluster distance). In addition, a list of pairwise distances for samples from another cluster c', $P_{c,c'}$ is computed.

Finally, GS for a given cluster, taking smallest $x\%$ samples, is given as the difference between the intra-cluster distances and the distance to the closest different cluster, normalised by maximum of the two values, given as:

$$GS_c(x) = \frac{P_{c,c'}(x) - P_{c,c}(x)}{max(P_{c,c'}(x), P_{c,c}(x))} \quad (1)$$

Cluster Purity is used to determine how many samples in a cluster belong to the same class when k-means clustering is applied:

$$CP_c = \frac{max_j|K_c \cap t_j|}{N_c} \quad (2)$$

where K_c , N_c denote the samples in a given k-means cluster and their total count respectively whereas t_j refers to samples from j^{th} GT class. Say, we assume a k-means cluster with 500 samples that has 490 samples belonging to the same GT class. That would lead to a cluster purity of 98%.

3.3 Experiments and Discussion

Here, we discuss several experiments conducted on the CIFAR-10 dataset [Krizhevsky et al. 2009] using ResNet-50 model [He et al. 2016] to evaluate cluster quality of contrastively trained SupCon, SimCLR with CE as baseline using the metrics presented in Section 3.2.

Evolution of cluster formation over training time: We investigated the evolution of separation of class-based clusters over training epochs as shown in Figure 2. We observe that supervised methods like CE and SupCon show increasing global separation over time starting initially from a negative separation (epoch 0) with SupCon learning much faster and showing better separation. Notably, CE still leads to quite well-defined clusters even without a contrastive loss. The unsupervised SimCLR does not show much further separation after an initial clustering of embeddings, based on GT classes. The unsupervised goal of discriminating between individual samples rather than class, hurts the overall formation of class-based clusters placing it even further below cross-entropy in that domain. Some classes like 'frog ', 'automobile 'or 'truck 'show good separability while 'cat 'and 'dog 'have consistently worse clusters across all methods. This could indicate that 'cat 'and 'dog 'are harder to distinguish, sharing the majority of overall features, than the other more separated classes. On the other hand, it could also be an artefact of the individual distribution of samples for each class in the CIFAR-10 dataset.

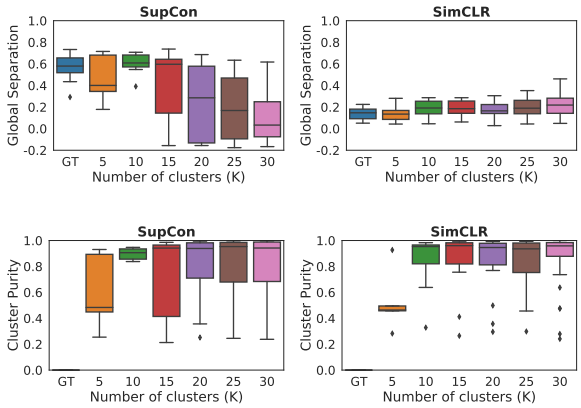

Figure 3: Comparison of cluster quality for SupCon and SimCLR based on GS and CP. Starting with default class-based (GT) clusters, k-means with an increasing number of clusters is performed to determine the cluster quality of underlying feature-based clusters.

Class-based or feature-based clusters? In the previous section, we found no prominent GT *class-based cluster* separation for the unsupervised CL approach. However, it is still possible that *feature-based clusters* form based on common features than class semantics. Thus, we further investigated whether applying k-means clustering on top of such contrastively trained embeddings can regroup them into better separated distinct clusters using the GS and CP metrics along with cosine similarity, as shown in Figure 3. Class-based clusters $K = GT$ are shown for comparison. The boxplots indicate the range of global separation for all clusters

for a given configuration. CP is only applied to the k-means clusters since the class-based clusters always have a purity of 1.

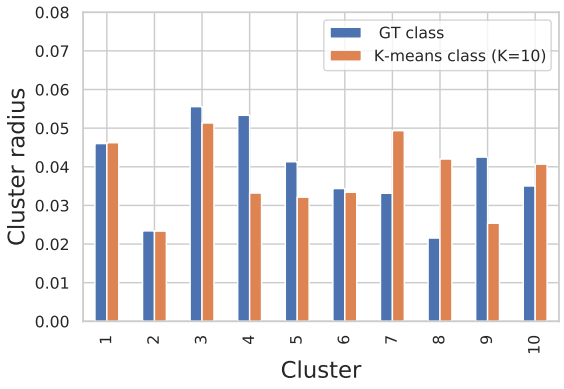

Figure 4: Variation in sizes of clusters in embedding space

SupCon shows already good cluster separation for GT class-based clusters due to supervised contrastive training. Subsequently, with 10 k-means clusters, separation gets even slightly better while cluster purity around 90% indicates a significant overlap with the GT clusters. But, for SimCLR we observe GT class-based clusters are not well separated. This is expected since the contrastive loss is applied at the projection layer maximising its effect there, it is difficult for feature layer to show well separated GT class-based clusters for an unsupervisedly learnt embeddings.

What do clusters look like? Many works on OOD detection using the distance-based scores use a single global threshold to distinguish between ID and OOD samples [Hsu et al. 2020, Lee et al. 2018, Tack et al. 2020, Sehwag et al. 2021, Hendrycks et al. 2019b], i.e. if a sample is farther away from the closest cluster than this threshold, it is classified as OOD. However, this implies the assumption that all clusters have an equal size or even hyperspheres with the same diameter depending on distance metrics used. Since this is a very strong assumption, we compare the size of the individual clusters. Figure 4 shows the cosine distance from the cluster centre encompassing 95% of the training samples of this cluster using the feature layer embeddings from SupCon as an example. We interpret this figure as an approximation of the cluster radius and observe that the largest cluster has almost 2.5 times the smallest cluster. This is true for GT clusters as well as k-means clusters. We therefore conclude that the use of a single global threshold does not reflect the true nature of the clusters motivating the use of individual thresholds for each cluster. Thus, in the next section, we present a study on the OOD detection performance for different models and metrics by evaluating with both per cluster-based as well as global threshold based metrics.

4 OOD detection based on clustering

Here, we describe our OOD detection method using cluster-based thresholds and compare them to global thresholds for different distance-based scoring metrics and clustering methods across diverse ID/ OOD datasets and model architecture. Further, we investigate the relationship between cluster quality and OOD detection performance.

4.1 Method

From our analysis in Section 3, we can find well-separated clusters of ID samples at the last layer of the feature extractor once a model is sufficiently trained. We use the distance to the mean of a cluster as an indication of how similar a test sample is to one of the samples in ID cluster. The farther the sample, the less likely it is to be related to ID clusters. If a sample is far enough away from all clusters, it is considered to be an OOD sample. Since we discovered in the previous section that clusters have varying sizes, we propose to employ *cluster-based thresholds* as compared to global thresholds for distance-based scoring metrics.

Our *approach* consists of the following steps: (1) During training, the mean of all clusters with respective training samples is calculated. This is for clusters based on GT class labels. In case of k-means clustering or using Guassian Mixture Models (GMM), mean is calculated based on samples assigned to respective clusters by these methods. (2) Using mean and distance metrics, distance scores are calculated for all train and test samples. (3) During inference, for each cluster, the set of distance scores for the respective reference distribution (train/test) are taken. For each new test sample, a probability score is assigned depending upon where the distance score of the given test sample can fit in the overall distribution of distance scores of the given cluster. These probability scores of test samples are finally used for calculating evaluation metrics. The global threshold based probability scores can also be similarly calculated by taking the entire reference distribution of distance scores into account.

Distance metrics: In order to test the proximity of a sample to a cluster, we need to either approximate the underlying distribution function or use a simple distance metric such as *Cosine similarity* or *Euclidean* distance. The former can be analysed by mapping to cluster conditioned linear multivariate Gaussians. This can be expressed as *Mahalanobis distance* based score [Mahalanobis 1936], by calculating cluster-wise mean (μ_c) and co-variance (Σ_c) corresponding to the features $f(x)$ of a test sample x, as given in Equation 3. Here $S_c(x)$, represents the distance of a sample x from centre of cluster c.

$$S_c(x) = (f(x) - \mu_c)^T \Sigma_c^{-1} (f(x) - \mu_c) \tag{3}$$

Although Mahalanobis distance is a reasonable choice for highly correlated data it is prone to the *curse of dimensionality* with increasing dimensions of the embeddings [Ververidis & Kotropoulos 2009] as well as increasing components. Cosine similarity, is often deemed to be better suited for computing distances with high dimensional inputs. Since, during contrastive training, instance based comparisons utilise cosine similarity, we employ this score for cluster based distances as well. In our experiments, we have also used GMM to map the embedding space into a mixture

potential Gaussian components, thus assigning each sample to one of these clusters/components based on Expectation Maximisation rather than simply taking the GT class labels. Subsequently, Mahalanobis distance has been used to calculate the distance scores with respect to this cluster assignment. We have incorporated this an alternate to k-means clustering which simply utilises Euclidean distance between the feature vectors to assign samples to different clusters. Finally, we compare all the cluster-based and global evaluation metrics with a global evaluation of the probability scores assigned by applying GMM irrespective of which clusters/-components each sample belong to.

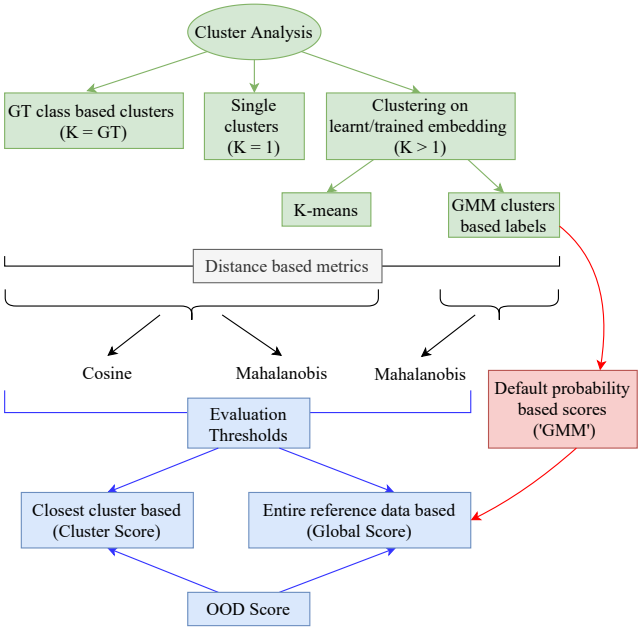

Figure 5: Schematic showing our OOD evaluation pipeline

4.2 Experiments and Discussion

In this section, we investigate the OOD performance for models trained on CIFAR-10, CIFAR-100 [Krizhevsky et al. 2009] with SupCon, SimCLR and baseline CE using ResNet architecture. We present results with SVHN [Netzer et al. 2011] and CIFAR-100 /CIFAR-10 as OOD data sets due to limited space, however we have conducted experiments with other OOD datasets like resized ImageNet, LSUN [Liang et al. 2020] and our observations are also extended to these datasets. The results of our study is presented in Figure 6.

Given 'K' represents the number of clusters, we base our investigation on following type of clusters in the embedding space - GT class based clusters ($K = GT$) where features are taken from embeddings trained based on GT class labels, single cluster ($K = 1$) where all the embeddings are taken as one cluster and finally similar feature-based clusters ($K = 5, 10, 15, ..$) mapped by using either k-means clustering or GMM to re-group the embeddings into distinctive clusters. This process of the OOD evalaution pipeline has been represented in the schematic given in Figure 5. While $K = 1$ case is similar to the results presented in [Sehwag et al. 2021] for SimCLR, however by investigating further clusters we show that the mentioned setting is not always the best case scenario. It largely depends on choice of different variables as shown in our study. As mentioned in previous Section 4.1, for all cluster based analysis using k-means, we employ Cosine similarity ('KM+Cos')and Mahalanobis distance ('KM+Maha') except for clusters based on GMM based target labels ('GMM+Maha ') we use only Mahalanobis distance only due to inherent assumption of Gaussian components. Finally, we employ the Area under ROC curve (AUROC) as the main evaluation metric for OOD detection where we present 'AUROC cluster ', 'AUROC global 'as the respective scores for cluster-based and global thresholds. We also compare with default 'GMM'scores in global evaluation.

Performance across GT class-based and feature-based clusters: From Figure 6, using class-based clusters ($K = GT$), we observe comparable AUROC scores across all the methods with SupCon quite similar to CE and slightly better than SimCLR. For feature clusters with $K = 10$, we achieve almost similar performance as with class-based clusters for CIFAR-10. This is expected since we see a strong overlap of the clusters generated by k-means/ GMM and the GT classes. With further increasing clusters ($K > 10$), almost similar trend exists as $K = 10$. For CIFAR-100 as ID, we see an overall decreasing trend across all distance metrics except for Cosine ($K > 1$). This could be due to fewer clusters than GT cluster(100). However, for SimCLR it appears to be beneficial to treat all ID samples as one single cluster ($K = 1$), as illustrated in Figure 1. This result is in line with our observation, that SimCLR does not lead to well-separated clusters and therefore making a distinction between clusters is superfluous. Nonetheless, it is a bit surprising that this single clusters also leads to competitive OOD detection performance across all datasets.

Performance across distance metrics: Taking cue on SimCLR performing best at $K = 1$ using Mahalanobis distance in almost all cases in Figure 6 indicates the possibility of high covariance between clusters when taken as a whole. Notably, for GT clusters cosine similarity always performs better than Mahalanobis [Ververidis & Kotropoulos 2009]. However, this is not the case for $K > 1$ as Mahalanobis shows a decreasing trend in general from maximum at single cluster (for SimCLR) while the AUROC for cosine similarity drops for $K = 1$ and then continues an upward trend until optimum cluster at $K = 10$ for SupCon and slightly further for SimCLR in case of CIFAR-10 as ID dataset. For CIFAR-100, it continues with an upward trend for both SupCon and SimCLR. CE with cosine similarity achieves best performance across all methods in CIFAR-100. Also for clustering cases, performance of GMM and k-means cluster based Mahalanobis has been comparably same with slightly greater performance by GMM based clusters. For supervised cases, the AUROCs remain comparable with increase in clusters but for SimCLR they show usual declining trend. For most global cases, the default GMM probability remain steady and similarly high to other distance metrics across different

Figure 6: Comparison of OOD detection performance using cluster-based and global thresholds for AUROC evaluation metric on CIFAR-10 (top 3 rows), CIFAR-100 (bottom 3 rows) as ID data across different methods(CE, supCon, SimCLR) for different distance metrics. x axis shows number of clusters (K), where GT implies GT number of clusters, $K = 1$ implies single cluster and $K > 1$ indicates number of clusters after k-means/ GMM clustering.

clusters.

Performance across cluster-based and global thresholds Although cluster-based thresholds seek to represent the embedding clusters better than global thresholds, however global AUROCs tend to perform slightly better than cluster AUROCs in most cases. Although the clusters are of inequal sizes, however depending on features of the given ID dataset they tend to be quite overlapping, so that global thresholds seem to be good enough for OOD detection.

Performance across OOD datasets: All the above observations remain consistent across all the OOD datasets. However, we note SVHN being semantically quite different from CIFAR-10 (Far OOD) achieves much better AUROC as compared to CIFAR-100 which is semantically quite close to CIFAR-10.

Performance across ID datasets We note that the AUROCs for CIFAR-100 (ID) vs CIFAR-10 (OOD) are much lower compared to vice-versa, although it follows similar trends mentioned above. This could be due to more classes (100 vs 10) leading to much smaller closely overlapping embedding clusters. This distinction becomes more difficult with similar OOD dataset like CIFAR-10, as when compared to really different SVHN.

Performance across model architectures We conducted similar experiments on ResNet-18 (although not reported here) vs ResNet-50, however we find much lesser overall AUROCs in the former, with SimCLR performing superior than supervised cases. This could potentially indicate requirement for bigger models as supervised cases require higher positives for instance based discrimination in CL.

5 Conclusion

In this work, we first investigated the nature of clusters in embedding spaces of contrastively trained models for image classification. We found that supervised contrastive training leads to well-separated clusters of in-distribution data in the embedding space and that these clusters correlate strongly with the ground truth classes. Unsupervised contrastive training on the other hand leads to mostly overlapping clusters that cannot be clearly distinguished. To our surprise, standard cross-entropy loss also lead to reasonably distinct clusters. Secondly, we proposed a modular OOD detection method exploiting proximity of similar samples in the embedding space allowing us to compare different distance metrics, clustering methods and thresholding strategies across a selection of model architectures, ID and OOD datasets. While we could reproduce the superior performance of SimCLR with a single cluster and Mahalanobis distance for CIFAR-10 vs. CIFAR-100, cross-entropy with clusters based on ground truth classes and cosine similarity performed best when the in- and out-of-distribution roles are reversed. However, in many cases there exist several combinations that lead to similar detection performance making the results even more ambiguous. We therefore have to conclude, that there is no clear winner and not even a solid trend, yet. For a deeper understanding, future work should extend this investigation to a broader and more diverse set of model architectures. Given recent observations that many OOD methods do not translate well from academic datasets to real world applications [Berger et al. 2021], further research should focus on studies with real image data from medical, automotive or industrial use cases.

References

Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning Representations by Maximizing Mutual Information Across Views. *arXiv:1906.00910 [cs, stat]*, July 2019.

Berger, C., Paschali, M., Glocker, B., and Kamnitsas, K. Confidence-based out-of-distribution detection: A comparative study and analysis. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis*, pp. 122–132. Springer, 2021.

Bojchevski, A., Matkovic, Y., and Günnemann, S. Robust Spectral Clustering for Noisy Data: Modeling Sparse Corruptions Improves Latent Embeddings. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 737–746, Halifax NS Canada, August 2017. ACM. ISBN 978-1-4503-4887-4. doi: 10.1145/3097983.3098156.

Boult, T. E., Cruz, S., Dhamija, A. R., Gunther, M., Henrydoss, J., and Scheirer, W. J. Learning and the Unknown: Surveying Steps toward Open World Recognition. *AAAI*, 33(01):9801–9807, July 2019. ISSN 2374-3468. doi: 10.1609/aaai.v33i01.33019801.

Burton, S., Kurzidem, I., Schwaiger, A., Schleiß, P., Unterreiner, M., Graeber, T., and Becker, P. Safety Assurance of Machine Learning for Chassis Control Functions. In *Computer Safety, Reliability, and Security*, LNCS, Cham, 2021. Springer International Publishing. doi: 10/gjq3ch.

Charpentier, B., Zügner, D., and Günnemann, S. Posterior Network: Uncertainty Estimation without OOD Samples via Density-Based Pseudo-Counts. *arXiv:2006.09239 [cs, stat]*, October 2020.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, November 2020a.

Chen, X., Fan, H., Girshick, R., and He, K. Improved Baselines with Momentum Contrastive Learning. *arXiv:2003.04297 [cs]*, March 2020b.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Geng, C., Huang, S.-J., and Chen, S. Recent Advances in Open Set Recognition: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020. ISSN 1939-3539. doi: 10.1109/TPAMI.2020.2981604.

Golan, I. and El-Yaniv, R. Deep Anomaly Detection Using Geometric Transformations. *arXiv:1805.10917 [cs, stat]*, November 2018.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, July 2017.

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

Henaff, O. Data-Efficient Image Recognition with Contrastive Predictive Coding. In *International Conference on Machine Learning*, pp. 4182–4192. PMLR, November 2020.

Hendrycks, D. and Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *arXiv:1610.02136 [cs]*, October 2018.

Hendrycks, D., Mazeika, M., and Dietterich, T. Deep Anomaly Detection with Outlier Exposure. *arXiv:1812.04606 [cs, stat]*, January 2019a.

Hendrycks, D., Mazeika, M., Kadavath, S., and Song, D. Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty. In *Advances in Neural Information Processing Systems 32*, pp. 15663–15674. October 2019b.

Henne, M., Schwaiger, A., and Weiss, G. Managing Uncertainty of AI-based Perception for Autonomous Systems. In *AISafety@IJCAI 2019, Macao, China, August 11-12, 2019*, volume 2419 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019.

Hildebrandt, M., Li, H., Koner, R., Tresp, V., and Günnemann, S. Scene graph reasoning for visual question answering. *arXiv preprint arXiv:2007.01072*, 2020.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv:1808.06670 [cs, stat]*, February 2019.

Hodge, V. and Austin, J. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, 22(2):85–126, October 2004. ISSN 1573-7462. doi: 10.1023/B:AIRE.0000045502.10941.a9.

Hsu, Y.-C., Shen, Y., Jin, H., and Kira, Z. Generalized ODIN: Detecting Out-of-Distribution Image Without Learning From Out-of-Distribution Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10951–10960, 2020.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised Contrastive Learning. *arXiv:2004.11362 [cs, stat]*, April 2020.

Koner, R., Sinhamahapatra, P., and Tresp, V. Relation transformer network. *arXiv preprint arXiv:2004.06193*, 2020.

Koner, R., Li, H., Hildebrandt, M., Das, D., Tresp, V., and Günnemann, S. Graphhopper: Multi-hop scene graph reasoning for visual question answering. In *International Semantic Web Conference*, pp. 111–127. Springer, 2021a.

Koner, R., Sinhamahapatra, P., Roscher, K., Günnemann, S., and Tresp, V. Oodformer: Out-of-distribution detection transformer. *CoRR*, abs/2107.08976, 2021b. URL https://arxiv.org/abs/2107.08976.

Koner, R., Sinhamahapatra, P., and Tresp, V. Scenes and surroundings: Scene graph generation using relation transformer. *arXiv preprint arXiv:2107.05448*, 2021c.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lee, K., Lee, K., Lee, H., and Shin, J. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. *arXiv:1807.03888 [cs, stat]*, October 2018.

Liang, S., Li, Y., and Srikant, R. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. *arXiv:1706.02690 [cs, stat]*, August 2020.

Mahalanobis, P. C. On the generalized distance in statistics. National Institute of Science of India, 1936.

Mohseni, S., Pitale, M., Yadawa, J., and Wang, Z. Self-Supervised Learning for Generalizable Out-of-Distribution Detection. In *Proc. AAAI 2020*, pp. 8, 2020.

Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., and Lakshminarayanan, B. Do Deep Generative Models Know What They Don't Know? *arXiv:1810.09136 [cs, stat]*, February 2019.

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

Nguyen, A., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Perera, P., Nallapati, R., and Xiang, B. OCGAN: One-Class Novelty Detection Using GANs With Constrained Latent Representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2898–2906, 2019.

Pidhorskyi, S., Almohsen, R., and Doretto, G. Generative Probabilistic Novelty Detection with Adversarial Autoencoders. In *Advances in Neural Information Processing Systems 31*, pp. 6822–6833. Curran Associates, Inc., 2018.

Prabhu, V., Kannan, A., Ravuri, M., Chablani, M., Sontag, D., and Amatriain, X. Prototypical Clustering Networks for Dermatological Disease Diagnosis. *arXiv:1811.03066 [cs]*, November 2018.

Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., DePristo, M. A., Dillon, J. V., and Lakshminarayanan, B. Likelihood ratios for out-of-distribution detection. *arXiv preprint arXiv:1906.02845*, 2019.

Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, November 1987. ISSN 0377-0427. doi: 10.1016/0377-0427(87)90125-7.

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. Deep One-Class Classification. In *International Conference on Machine Learning*, pp. 4393–4402. PMLR, July 2018.

Schwaiger, A., Sinhamahapatra, P., Gansloser, J., and Roscher, K. Is Uncertainty Quantification in Deep Learning Sufficient for Out-of-Distribution Detection? In *Proc. AISafety@IJCAI2020*, volume 2640 of *CEUR Workshop Proceedings*, pp. 8, 2020.

Sehwag, V., Chiang, M., and Mittal, P. SSD: A Unified Framework for Self-Supervised Outlier Detection. *arXiv:2103.12051 [cs]*, March 2021.

Tack, J., Mo, S., Jeong, J., and Shin, J. CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances. *arXiv:2007.08176 [cs, stat]*, July 2020.

Ververidis, D. and Kotropoulos, C. Information loss of the mahalanobis distance in high dimensions: Application to feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2275–2281, 2009. doi: 10.1109/TPAMI.2009.84.

Winkens, J., Bunel, R., Roy, A. G., Stanforth, R., Natarajan, V., Ledsam, J. R., MacWilliams, P., Kohli, P., Karthikesalingam, A., Kohl, S., Cemgil, T., Eslami, S. M. A., and Ronneberger, O. Contrastive Training for Improved Out-of-Distribution Detection. *arXiv:2007.05566 [cs, stat]*, July 2020.

Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., van Ginneken, B., Madabhushi, A., Prince, J. L., Rueckert, D., and Summers, R. M. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proc. IEEE*, 109(5):820–838, May 2021. ISSN 0018-9219, 1558-2256. doi: 10.1109/JPROC.2021.3054390.

6 Acknowledgments

This work was funded by the Bavarian Ministry for Economic Affairs, Regional Development and Energy as part of a project to support the thematic development of the Institute for Cognitive Systems (IKS).