Contrastive Learning-based Representational Space for Out-of-distribution Detection

Harshith Gundappa, Sai Raghava Mukund Bhamdipati, Sourav Suresh, Omkar Prabhune

[1]University of Wisconsin-Madison

{gundappa,bhamidipati3,sourav.suresh,oprabhune}@wisc.edu

Abstract. *Detecting out-of-distribution (OOD) samples is critical for the reliable deployment of machine learning models. Previous approaches impose a strong distributional assumption of the underlying feature space. Newer models such as [2] attempt to solve this problem and generalize well across different architectures. We propose to use contrastive learning along with this approach so that the latent representations would yield well-defined clusters resulting in better OOD separability.*

Problem Description :

For an AI system to be reliable in a real-world setting, it is crucial that the system is capable of recognizing out-of-distribution data. The model should not only perform well on the test data drawn from the same distribution as training data but also recognize the test data sampled from unknown distributions that the model has never seen during training. In the context of image classification and object detection where the model has been trained on a fixed set of categories, it should be able to identify if an image (or a bounding box) belongs to one of the known categories (in-distribution or ID) or an unknown category (OOD).

With this project, we aim to design a new framework for OOD detection that involves learning a favorable latent representation of samples along with a carefully crafted loss function that encourages the clustering of ID points and separation from OOD points.

Proposed Approach :

With regards to outlier synthesis, previous work relies on a variety of approaches to perform OOD detection - from maintaining real, large outlier datasets to synthesizing newer data [6][3][2]. The former category can be infeasible as it is difficult teach the model each and every scenario it can encounter in the wild. Modern methods, thus, employ generating outlier samples during training time.

In [4], the authors generate new training samples in the high dimensional pixel space during training time. This method, however, slowly falls apart as we increase the size of the image space. Alternatively, in [3], the authors hypothesize that latent representations of the same class form a Gaussian distribution. OOD samples are then synthesized in the latent space based on this distributional assumption. This method showcases good results while being easy to implement. However, the assumption of Gaussian distribution in the latent space may not hold true [2]. More recent methods [2] try to build upon this idea by trying to morph the latent space into a hypersphere to allow easy separability and lesser computation.

With regards to loss functions, existing methods combine latent representations with targeted loss functions [3][2] to cluster the samples. It can also be seen from [6] that the need for such clustering is conducive to using Contrastive Learning based training which inherently tries to group together similar instances while pushing away dissimilar ones.

We propose to combine the best of both worlds in our hypothesis as described below:

- Transform images to a latent space for better separability.
- A contrastive learning approach to learn these latent representations to yield well-defined clusters for each class, facilitating improved OOD detection.
- Using a distance-based score such as the Mahalanobis distance, L2 distance, KNN, etc. to infer whether a particular test sample is ID or OOD.

Evaluation Plan :

Performance metric

The performance of the proposed OOD detection methodologies will be evaluated using the metrics of false positive rate (FPR) and the area under the receiver operating characteristic curve (AUROC) as used in [6]. Further, we will evaluate the effect of OOD detection mechanism on the classification performance of ID samples.

Dataset

We will begin with the CIFAR benchmarks as commonly observed in literature [6]. Common OOD datasets such as Textures [1], SVHN [5], Places365 [9], LSUN-C [8], and iSUN [7] as used in [6] will be used to evaluate the performance of OOD detection.

Project Timeline :

- Phase 1 - (Oct 28): Conceptualization of the loss function and initial training setup
- Phase 2 - (Nov 25): Iterative model optimization with training and validation
- Phase 3 - (Dec 8): Evaluation of the overall framework and compilation of results in report
- Dec 8 - 21 Final Submission/ Project presentation

Referências

[1] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[2] Xuefeng Du, Gabriel Gozum, Yifei Ming, and Yixuan Li. Siren: Shaping representations for detecting out-of-distribution objects. In *Advances in Neural Information Processing Systems*, 2022.

[3] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis, 2022.

[4] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples, 2017.

[5] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[6] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors, 2022.

[7] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

[8] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[9] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.