Comprehensive OOD Detection Improvements

Anish Lakkapragada[1, 2], Amol Khanna[1], Edward Raff[1,3], Nathan Inkawhich[4]

[1]Booz Allen Hamilton, [2]Yale University, [3]University of Maryland, Baltimore County, [4]Air Force Research Laboratory

Introduction

Goal: We aim to improve the state of out-of-distribution (OOD) detection.

Motivation: As machine learning becomes increasingly prevalent in impactful decisions, recognizing when inference data is outside the model's expected input distribution through OOD detection is paramount for giving context to predictions.

Contributions: We address the entire OOD detection landscape. We employ dimensionality reduction on feature embeddings in representation-based methods. Additionally, we propose DICE-COL, a modification of the popular DICE method.

Conclusion: Our methods achieve comparable, if not higher, performance as SOTA OOD detectors.

Background

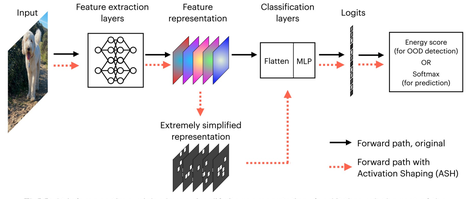
OOD Detection

$$G(x) = \begin{cases} ID \ if \ S(x) \geq \lambda \\ OOD \ if \ S(x) \geq \lambda \end{cases}$$

Decision function $G(x)$ determines if sample x is out-of-distribution (OOD) or in-distribution (ID) based on scoring function $S(x)$ and threshold λ.

Background

Logit-Based Methods



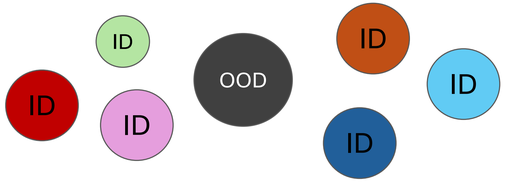
TL;DR: At inference time, pick a layer, simplify its representation, feed it through the rest of the network. Accuracy is not affected and OOD detection is much better!

From the ASH paper (Djurisic 2022)

Logit-based methods apply some adjustment to the the model exclusively during inference and utilize model's **prediction** in their scoring function $S(x)$. These adjustments form distinguishable distributions of $S(x)$ scores between OOD and ID samples.

Representation-based Methods

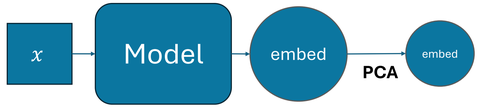
Representation-based methods utilize the model's **embedding** of a sample x in their scoring function $S(x)$ to understand how OOD a given sample is. **Visualization below:**



Contribution

Representation-based Method Contributions

- **Dimensionality reduction** to reduce embedding space of model's has been found to improve performance for some representation-based methods (Woodland 2023)
- **Our contribution**: Test dimensionality reduction more rigorously across representation-based methods: MDS, KNN, RMDS
- We use PCA to perform dimensionality reduction



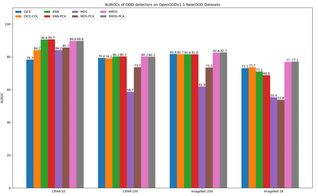
Logit-based Method Contributions

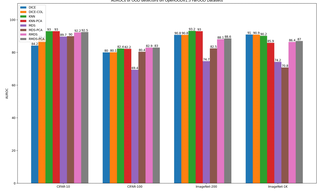
- **DICE** (Sun and Li 2022) is a method that zeroes the 90% of the weights of the network's final layer and uses Energy scores as scoring function
- **DICE** computes which weights to remove across the *entire* weight matrix, which can zero entire weight class column
- **Our contribution**: Proposing **DICE-COL**, which computes masks for the weights on a column-by-column basis to prevent zero-age of the entire weight class column
- We intuitively expect DICE-COL to achieve, at worst, equal performance to DICE

Results

Evaluation Framework: OpenOODv1.5

ID Dataset	Model	NearOOD	FarOOD
CIFAR-10	ResNet-18	CIFAR-100, TIN	MNIST, SVHN , Textures, Places365
CIFAR-100	ResNet-18	CIFAR-10, TIN	MNIST, SVHN, Textures, Places365
ImageNet-200	ResNet-18	SSB-Hard, NINCO	iNaturalist, Textures, OpenImage-O
ImageNet-1K	ResNet-50	SSB-Hard, NINCO	iNaturalist, Textures, OpenImage-O





Conclusion

- Feature transformations have value for OOD representation-based detection methods
- We set new SOTA records on the OpenOODv1.5 benchmark and create methods with comparable, if not higher, performance