

Abstract

The ability to detect OOD data is a crucial aspect of practical machine learning applications. In this work, we show that cosine similarity between the test feature and the typical ID feature is a good indicator of OOD data. We propose Class Typical Matching (CTM), a post hoc OOD detection algorithm that uses a cosine similarity scoring function. Extensive experiments on multiple benchmarks show that CTM outperforms existing post hoc OOD detection methods.

Contributions

- Our main contribution is CTM, a post-hoc method that leverages angular information to enhance OOD detection performance.
- We conduct comprehensive experiments and ablation studies to demonstrate the effectiveness of our method across 3 ID datasets and 10 OOD datasets.
- We also provide theoretical insight into the relationship between CTM and influence measures.

Problem statement

One common approach to out-of-distribution (OOD) detection is to construct a scoring function $S: \mathcal{X} \mapsto \mathbb{R}$ that assigns lower scores to inputs drawn from an out-distribution. The detector, denoted as g , is then constructed based on the level set obtained from the score function

$$g(\mathbf{x}) = \begin{cases} \text{ID}, & \text{if } S(\mathbf{x}) \geq \lambda \\ \text{OOD}, & \text{if } S(\mathbf{x}) < \lambda \end{cases},$$

where $S(\mathbf{x})$ denotes a scoring function and λ is commonly set so that g correctly classifies a high proportion (e.g., 95%) of in-distribution (ID) data.

Motivation

Angular view from simple MaxLogit baseline

$$\begin{aligned} & \max_k \langle \mathbf{w}_k, \mathbf{z} \rangle + b_k \\ &= \max_k \|\mathbf{w}_k\| \|\mathbf{z}\| \cos(\mathbf{w}_k, \mathbf{z}) + b_k \end{aligned}$$

- The cosine similarity term $\cos(\mathbf{w}_k, \mathbf{z})$ carries the most information for the model's prediction.
- Using only angular information from of feature embedding can retain the performance on the OOD detection task.
- Replace \mathbf{w}_k by the within-class mean $\boldsymbol{\mu}_k$.

Table 1. Cosine similarity is effective. Test accuracy and OOD Detection performance (AUROC) of models before and after modification.

Model & Dataset	Test Accuracy		AUROC	
	Standard/CW/CM	Standard/CW/CM	Standard/CW/CM	Standard/CW/CM
WideResNet-40-2 + CIFAR-10	94.84/94.82/95.02	91.29/92.49/92.49		
WideResNet-40-2 + CIFAR-100	75.95/75.93/75.03	77.39/79.77/86.95		
DenseNet + CIFAR-10	94.52/94.55/94.40	94.62/94.40/96.40		
DenseNet + CIFAR-100	75.08/74.69/71.66	80.28/75.01/89.11		

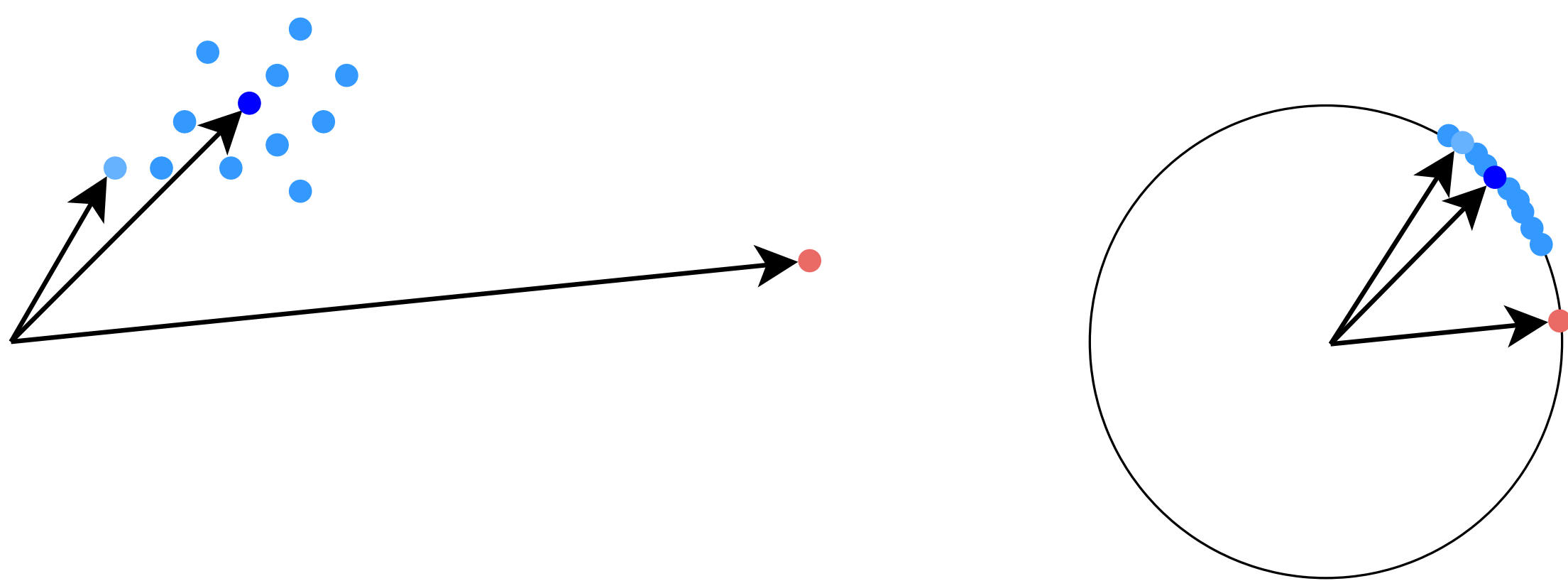


Figure 1. OOD inputs may have a large norm feature embedding and make it hard for MaxLogit to detect.

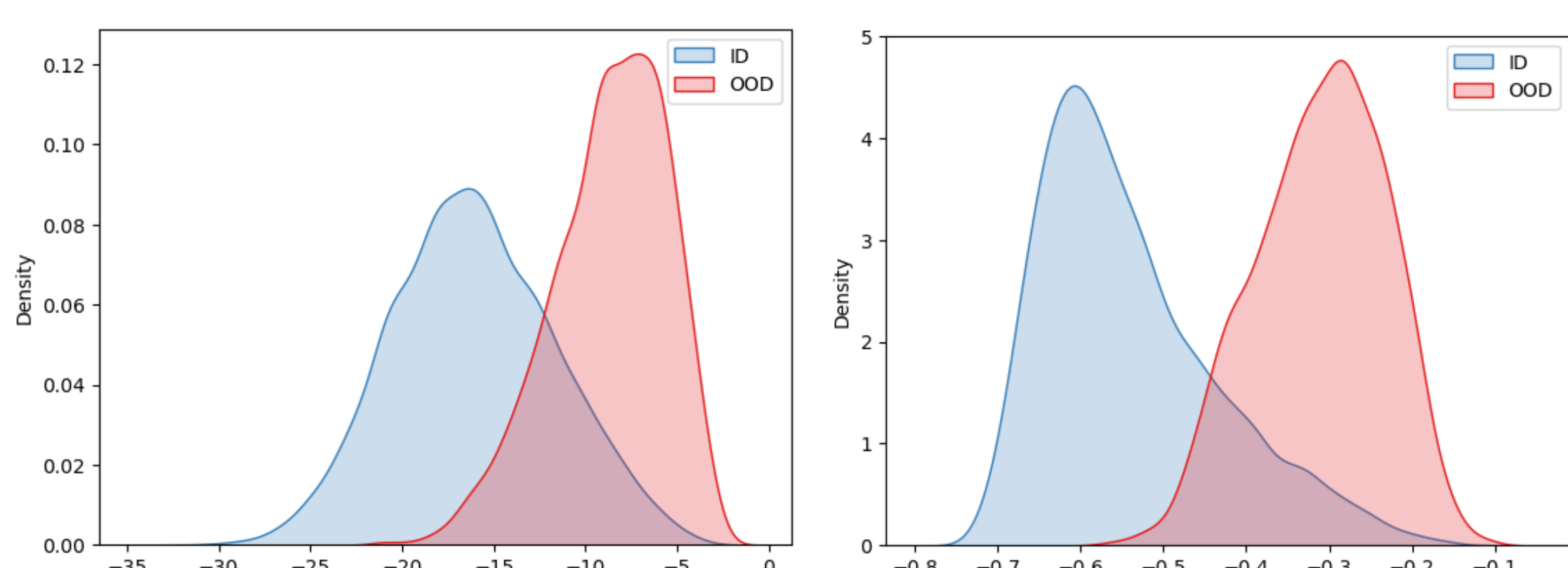


Figure 2. Distribution of ID scores vs OOD scores using MaxLogit (left) and CTM (right)

Proposed method: Class Typical Matching (CTM)

We propose using cosine similarity with within-class feature mean $\boldsymbol{\mu}_k$ for OOD detection

$$g(\mathbf{x}) = \begin{cases} \text{ID}, & \text{if } \max_k \cos(\boldsymbol{\mu}_k, \mathbf{z}) \geq \lambda \\ \text{OOD}, & \text{otherwise} \end{cases},$$

where λ is the threshold.

- The score function $S(\mathbf{x}) = \max_k \cos(\boldsymbol{\mu}_k, \mathbf{z})$ measures the similarity between the test input's feature and within-class mean features.
- CTM is extremely simple and has low computational complexity.

Relation to Influence measures

$$K_g(\mathbf{z}, \mathbf{z}') = \frac{\langle \nabla_W g_W(\mathbf{z}), \nabla_W g_W(\mathbf{z}') \rangle}{\|\nabla_W g_W(\mathbf{z})\| \|\nabla_W g_W(\mathbf{z}')\|}.$$

$$K_g(\boldsymbol{\mu}_k, \mathbf{z}) = \frac{p_k - 1/C}{(1 - 1/C)(\|\mathbf{p}\|^2 - 1/C)} \cdot \cos(\boldsymbol{\mu}_k, \mathbf{z})$$

- $K_g(\boldsymbol{\mu}_k, \mathbf{z})$ and $\cos(\boldsymbol{\mu}_k, \mathbf{z})$ are positively correlated
- Smaller $\cos(\boldsymbol{\mu}_k, \mathbf{z})$ indicates less influence between the typical ID feature $\boldsymbol{\mu}_k$ and the test input's feature \mathbf{z} . This can be signal of a OOD input.

Experiments

Experiment Settings

Datasets and models

ID Dataset	OOD Datasets	Model architectures
CIFAR-10	SVHN, LSUN C, LSUN R, iSUN, Places365, Textures	DenseNet-101
CIFAR-100	SVHN, LSUN C, LSUN R, iSUN, Places365, Textures	DenseNet-101
ImageNet	iNaturalist, SUN, Places365, Textures	ResNet50

Evaluation metrics: FPR95, AUROC, AUPR

Results

Table 2. OOD detection results on CIFAR benchmarks.

Method	CIFAR-10		CIFAR-100	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
Softmax score	48.73	92.46	80.13	74.36
MaxLogit	26.44	94.47	69.98	80.31
Energy score	26.55	94.57	68.45	81.19
ODIN	24.57	93.71	58.14	84.49
Mahalanobis	31.42	89.15	55.37	82.73
KNN	16.61	96.71	42.34	87.56
CTM (Ours)	18.23	96.40	41.76	89.11

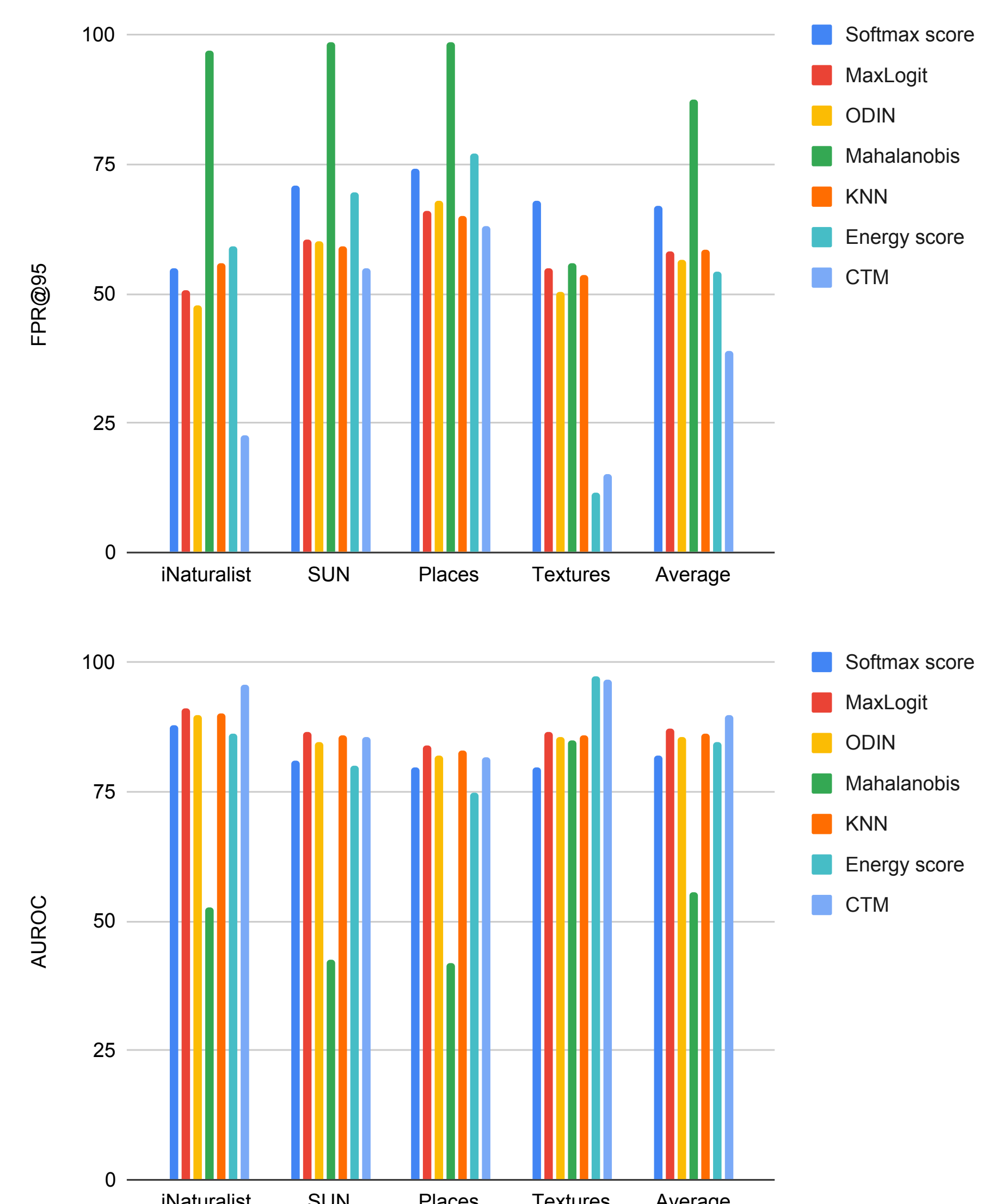


Figure 3. OOD Detection results on ImageNet benchmark