CASE: Exploiting Intra-class Compactness and Inter-class Separability of Feature Embeddings for Out-of-Distribution Detection

Shuai Feng [*], Pengsheng Jin, Chongjun Wang

State Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
{shuaifeng, jps}@smail.nju.edu.cn, chjwang@nju.edu.cn

Abstract

Detecting out-of-distribution (OOD) inputs is critical for reliable machine learning, but deep neural networks often make overconfident predictions, even for OOD inputs that deviate from the distribution of training data. Prior methods relied on the widely used softmax cross-entropy (CE) loss that is adequate for classifying in-distribution (ID) samples but not optimally designed for OOD detection. To address this issue, we propose **CASE**, a simple and effective OOD detection method by explicitly improving intra-class **C**ompactness **A**nd inter-class **S**eparability of feature **E**mbeddings. To enhance the separation between ID and OOD samples, CASE uses a dual-loss framework, which includes a separability loss that maximizes the inter-class Euclidean distance to promote separability among different class centers, along with a compactness loss that minimizes the intra-class Euclidean distance to encourage samples to be close to their class centers. In particular, the class centers are defined as a free optimization parameter of the model and updated by gradient descent, which is simple and further enhances the OOD detection performance. Extensive experiments demonstrate the superiority of CASE, which reduces the average FPR95 by **37.11**% and improves the average AUROC by **15.89**% compared to the baseline method using a softmax confidence score on the more challenging CIFAR-100 model.

Introduction

In the open world, machine learning models may encounter out-of-distribution (OOD) inputs that deviate from the distribution of training data, also known as in-distribution (ID) (Drummond and Shearer 2006; He et al. 2015; Krizhevsky, Sutskever, and Hinton 2017). Such inputs can cause models to make predictions with high confidence at test time, leading to unexpected and potentially harmful consequences. For instance, in safety-critical applications like autonomous driving, the driving system must alert the driver when it detects unusual scenes or objects that it cannot safely handle (Yang et al. 2021). Similarly, a fraud detection system may fail to identify a new type of fraud that it has not encountered before (Phua et al. 2010), resulting in significant financial losses. Therefore, OOD detection is crucial in ensuring the reliability and safety of machine learning systems.

Numerous OOD detection methods (Hendrycks and Gimpel 2016; Liang, Li, and Srikant 2017; Lee et al. 2018; Liu et al. 2020; Huang, Geng, and Li 2021) have relied on employing off-the-shelf softmax cross-entropy (CE) loss to produce embeddings that are sufficient for classifying ID samples. However, this approach may not be optimally designed for OOD detection since it cannot guarantee high levels of intra-class compactness and inter-class separability (Liu et al. 2016, 2017; Wang et al. 2018). The quality of the learned features is generally correlated with these two factors, where intra-class compactness measures how close the features with the same label are to each other, and inter-class separability measures how far away the features with different labels are (Luo et al. 2019). When both intra-class compactness and inter-class separability are maximized simultaneously, the learned features become more discriminative (Liu et al. 2016) and promote strong ID-OOD separability. Hence, optimizing these two factors is crucial for developing effective OOD detection techniques.

In this paper, our high level idea is to establish the correlation between OOD detection performance and embedding properties in distance-based representation space. To achieve this goal, we propose an effective OOD detection method called **CASE**, which improves the intra-class **C**ompactness **A**nd inter-class **S**eparability of ID feature **E**mbeddings. To formalize our approach, we introduce two complementary losses. The first is the separability loss, which maximizes the inter-class Euclidean distance to promote clear separation among different class centers. The second is the compactness loss, which minimizes the intra-class Euclidean distance to encourage samples to cluster near their respective class centers. By working in tandem, these two losses help to produce more discriminative feature embeddings that exhibit higher intra-class compactness and higher inter-class separability in distance-based representation space. Overall, our proposed method provides a promising solution for improving OOD detection performance and ID classification accuracy.

In particular, we emphasize the importance of a simple and efficient class center definition method in further improving OOD detection performance. There are several methods available for defining class centers (Kirchheim, Filax, and Ortmeier 2022). Some methods (Miller et al. 2021; Feng et al. 2023) treat the class centers as constant parame-

[*] Corresponding author

ters, which can restrict the learned mapping since the centers cannot dynamically account for the class similarity in the input space. Other approaches (Hassen and Chan 2020; Ming et al. 2023) estimate the class centers from each batch of data individually, yet the estimated class centers might be inaccurate if the number N of points from one class in the batch is small. To overcome these issues, we treat the class centers as a free optimization parameter of the model and update them by gradient descent (Hsu et al. 2020). We demonstrate that this approach is more efficient and has the potential to further improve OOD detection performance.

In summary, our key results and contributions are as follows:

- We propose CASE, a simple and effective method for OOD detection, which jointly optimizes a compactness loss and a separability loss to explicitly improve intra-class compactness and inter-class separability of feature embeddings. We show that CASE can promote strong ID-OOD separability and boost different post-hoc OOD detection methods.
- We define the class centers as a free optimization parameter of the model and update them via gradient descent. This approach leads to further improvements in the OOD detection performance.
- We conduct extensive evaluation and ablation experiments to show that CASE can improve both OOD detection and ID classification accuracy. Compared with the CE loss, CASE reduces the average FPR95 by **50.47**%, improves the average AUROC by **9.55**%, and increases ID classification accuracy by **3.41**% on CIFAR-10 with the softmax confidence score.

Related Work

Post-hoc OOD Detection Methods

The advantage of post-hoc methods is that they can be easily applied without any modifications to the training procedure or objective (Yang et al. 2021). OOD detection research originated from a simple baseline (Hendrycks and Gimpel 2016), where a test sample with high maximum softmax probability (MSP) score is classified as an ID example rather than an OOD example. However, DNNs frequently produce overconfident predictions, even for inputs that are significantly different from the training data, which raises doubts about using softmax confidence directly for OOD detection. This approach was further extended in ODIN (Liang, Li, and Srikant 2017) by using temperature scaling strategy and input preprocessing strategy. Mahalanobis distance-based scores (Lee et al. 2018) were used for OOD detection by modeling the class-conditional distributions of softmax neural classifiers using multivariate Gaussian distributions. Energy-OOD (Liu et al. 2020) proposes using an energy score for OOD detection, which is theoretically more aligned with the probability density of the inputs and is less likely to result in overconfident predictions. ReAct (Sun, Guo, and Li 2021) rectifies feature vectors by thresholding their elements with a certain magnitude. These methods utilize the traditional CE loss to train the model and focus on using the outputs or features of the model to estimate OOD uncertainty. The significant challenge is to develop a scoring function $S(\boldsymbol{x})$ that can effectively capture OOD uncertainty.

OOD Detection with Auxiliary OOD Data

While the exact OOD test distribution cannot be predicted, several prior methods have explored the use of synthetic data generated by GANs (Lee et al. 2017) or unlabeled data (Hendrycks, Mazeika, and Dietterich 2018) as auxiliary OOD training data. The inclusion of auxiliary OOD sets during training has been shown to improve OOD detection performance by explicitly regularizing the model through fine-tuning, resulting in lower confidence on anomalous examples (Liu et al. 2020; Hendrycks, Mazeika, and Dietterich 2018; Geifman and El-Yaniv 2019; Mohseni et al. 2020) Selectively training on auxiliary OOD data that induces uncertain OOD scores can improve OOD detection performance on both clean and perturbed adversarial OOD inputs, as proposed by the informative outlier mining approach (Chen et al. 2021). He et al. (He et al. 2022) propose a novel OOD detection framework called RONF, which improved the performance of both OOD and ID classification. Their approach included a virtual OOD data synthesis strategy called BFM that generated virtual OOD data outside of the ideal ID boundary. Ming et al. (Ming, Fan, and Li 2022) propose a novel posterior sampling-based outlier mining framework termed POEM, which allows for the efficient use of outlier data and promotes learning a compact decision boundary between ID and OOD data for improved detection. However, in our work, we do not use any auxiliary OOD sets during training.

OOD Detection with Model Re-training

Recent studies have investigated alternative loss functions for retraining models, as opposed to training models directly with traditional softmax CE loss. For instance, Generalized ODIN (G-ODIN) (Hsu et al. 2020) builds upon ODIN (Liang, Li, and Srikant 2017) by incorporating a specialized training objective known as DeConf-C, and selecting hyperparameters such as the perturbation magnitude on ID data to improve model performance. Another approach, Logit Normalization (Logit-Norm) (Wei et al. 2022), offers a simple fix to the overconfidence problem in NNs by enforcing a constant vector norm on the logits during training. IsoMax loss (Macêdo et al. 2021) proposes a seamless OOD detection approach based on logit isotropy and the maximum entropy principle. This method serves as a drop-in replacement for CE loss. An improved version, IsoMax+ (Macêdo and Ludermir 2021), performs isometrization of the distances used in IsoMax loss and replaces the entropic score with the minimum distance score. Recently, Feng et al. (Feng et al. 2023) propose CESED, an improved CE loss applied to the scalable Squared Euclidean Distance vector, which exploits hyperspherical evenly-distributed class centroids for OOD detection. Although our proposed method in this work shares a similar spirit with CESED, there are significant differences in the class center definition method and training objective.

Preliminaries

OOD Detection

We are focusing on a supervised multi-class classification problem. The input space is represented by $\mathcal{X}$, and the ID labels are denoted by $\mathcal{Y} = \{1, 2, ..., N\}$. The training set, denoted by $\mathcal{D}_{\text{in}}^{\text{tr}} = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^{n}$, is i.i.d. sampled from the joint distribution $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$. The marginal distribution over $\mathcal{X}$ is called the ID and is represented by $\mathcal{P}_{\mathcal{X}}$.

OOD detection is typically formulated as a binary classification problem. During test time, the aim is to determine whether a sample $\boldsymbol{x} \in \mathcal{X}$ is from $\mathcal{P}_{\mathcal{X}}$ (ID) or not (OOD). Ideally, the OOD samples should follow a distribution whose label set has no intersection with $\mathcal{Y}$. The decision is usually made using a thresholding mechanism as follows:

$$G_\lambda(x) = \begin{cases} \text{ID}\ , & \text{if } S(x) \geq \lambda \\ \text{OOD}, & \text{if } S(x) < \lambda \end{cases} \tag{1}$$

Where $S(x)$ is a scoring function that determines whether a sample $\boldsymbol{x}$ belongs to the ID or OOD. λ is the threshold for the decision and is commonly chosen to ensure that a significant proportion (e.g., 95%) of the ID data is correctly classified (Hendrycks and Gimpel 2016).

Softmax Cross-Entropy(CE) Loss

The simplicity and probabilistic interpretation of the softmax function have made it a popular choice among many NNs. When combined with the CE loss, they form a widely used component in DNNs. The formulation for this combination is as follows:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{b} \log \frac{\exp(\boldsymbol{z_{y_i}})}{\sum_{j=1}^{N} \exp(\boldsymbol{z_j})} \tag{2}$$

where b is the batch size, $\boldsymbol{z}_{y_i}$ represents the output value of the last fully connected layer of the correct class label y_i, and $\boldsymbol{z}_j$ is the output value of the last fully connected layer of the j-th class.

Although the softmax CE loss is widely used in deep learning models, it is not optimally designed for OOD detection considering that it does not explicitly encourage intra-class compactness and inter-class separability (Liu et al. 2016, 2017; Wang et al. 2018). To mitigate this limitation, we propose an effective OOD detection method by improving intra-class compactness and inter-class separability of feature Embeddings.

Methods

In this section, we begin by describing the architecture of proposed CASE-based classifier. We then provide a detailed description of the proposed training objective and the inference process.

Overview Architecture

The general architecture, depicted in Figure 1, comprises three main components:

(1) A deep neural network feature exactor, $f_e : \mathcal{X} \mapsto \mathbb{R}^e$ that projects an input image $\boldsymbol{x}$ to a high dimensional feature embedding $\boldsymbol{z} = f_e(\boldsymbol{x})$ (often referred to as the penultimate layer).

(2) A Class Centers Module, that obtains a trainable parameter $\boldsymbol{W}$, representing a set of class centers ($\boldsymbol{w_1}$,$\boldsymbol{w_2}$,...,$\boldsymbol{w_N}$), which is a free optimization parameter of the model and is obtained through a new FC1 layer. We use the *kaiming initialization* to initialize the parameter $\boldsymbol{W}$.

(3) A Scaling Module, that calculates a scaling parameter s for each sample. The parameter s is obtained by passing the extracted feature embedding $\boldsymbol{z}$ through an additional linear layer FC2, followed by a sigmoid activation function σ.

Finally, the output of our distance-based classifier is a vector of scalable Euclidean distance:

$$\begin{aligned} \boldsymbol{d} &= \frac{1}{s} \cdot e\left(\boldsymbol{z}, \boldsymbol{W}\right) \\ &= \frac{1}{s} \cdot \left(\|\hat{\boldsymbol{z}} - \hat{\boldsymbol{w}}_{\boldsymbol{1}}\|_2^2, \|\hat{\boldsymbol{z}} - \hat{\boldsymbol{w}}_{\boldsymbol{2}}\|_2^2, \ldots, \|\hat{\boldsymbol{z}} - \hat{\boldsymbol{w}}_{\boldsymbol{N}}\|_2^2\right)^T \end{aligned} \tag{3}$$

where $\|\cdot\|_2^2$ denotes the squared Euclidean norm, $\hat{\boldsymbol{z}} = \boldsymbol{z}/\left\|\boldsymbol{z}\right\|_2$, $\hat{\boldsymbol{w}}_j = \boldsymbol{w}_j/\left\|\boldsymbol{w}_j\right\|_2$, $(j = 1, 2, \ldots, N)$.

CASE: A Compactness Loss and a Separability Loss

Training Objective To improve the quality of feature embeddings for both OOD detection and ID classification, we propose a novel training objective termed **CASE**, which aims to enhance the intra-class **C**ompactness **A**nd inter-class **S**eparability of feature **E**mbeddings. Specifically, the term is defined as follows:

$$\mathcal{L}_{\text{CASE}} = \mathcal{L}_{\text{sep}} + \lambda_c \cdot \mathcal{L}_{\text{comp}} \tag{4}$$

where the coefficient λ_c is used to adjust the relative importance of the two components in the proposed training objective, **CASE**, which consists of a separability loss that promotes separability among different class centers and a compactness loss that encourages samples to be close to their class centers. By jointly optimizing these two losses, our training objective aims to produce more discriminative feature embeddings in distance-based representation space. The training scheme of our CASE loss is shown in Algorithm 1 in Appendix A.

A Compactness Loss To promote each sample to be aligned with its class center, we introduce a compactness loss, which is an improved softmax CE loss applied to the scalable squared Euclidean distance vector. The use of the scalable squared Euclidean distance vector ensures that the compactness loss is effective across a wide range of distance values, as it can be appropriately scaled. Specifically, the term is defined as follows:

$$\mathcal{L}_{comp} = -\sum_{i=1}^{b} \log \frac{\exp(-\frac{1}{s_i} \cdot \boldsymbol{e_{y_i}})}{\sum_{j=1}^{N} \exp(-\frac{1}{s_i} \cdot \boldsymbol{e_j})} \tag{5}$$

where $\boldsymbol{e}_j \, (j = 1, 2, \ldots, N)$ is the squared Euclidean norm between a latent feature embedding $f_e\left(\boldsymbol{x}\right)$ and its class center $\boldsymbol{w}_{y_i}$. Remembering that $\boldsymbol{e} = (\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_N)$ is defined in equation (3). Besides, each input logit is scaled by a scaling parameter s, which is given by:

$$s = \sigma\left(BN\left(w_s f_e(\boldsymbol{x}) + b_s\right)\right) \tag{6}$$

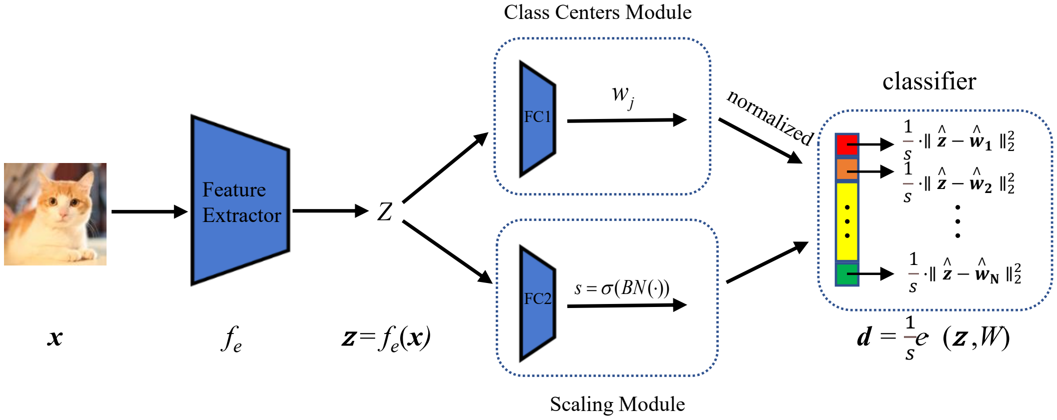

Class Centers Module
classifier
FC1
w_j
normalized
Feature Extractor
Z
FC2
$s=\sigma(BN(\cdot))$
$\frac{1}{s}\cdot\| \hat{z}-\hat{w}_1 \|_2^2$
$\frac{1}{s}\cdot\| \hat{z}-\hat{w}_2 \|_2^2$
$\frac{1}{s}\cdot\| \hat{z}-\hat{w}_N \|_2^2$
$\boldsymbol{x}$
f_e
$\boldsymbol{z}=f_e(\boldsymbol{x})$
$\boldsymbol{d}=\frac{1}{s}e\ (\boldsymbol{z},W)$
Scaling Module

Figure 1: An overview of our proposed CASE-based classifier.

where s is obtained by passing the feature $f_e(\boldsymbol{x})$ from the penultimate layer of the NN through an additional linear layer, followed by optional batch normalization BN for faster convergence, and a sigmoid activation function σ. The learnable weight and bias parameters of this added branch are denoted by w_s and b_s, respectively. The G-ODIN method (Hsu et al. 2020) also employs a scaling parameter to construct a dividend/divisor structure for a classifier, which encourages NNs to exhibit behavior similar to the decomposed confidence effect. However, in our proposed method, the primary objective of the scaling parameter is to adjust the squared Euclidean distance. This enables us to control the distance metric and improve the discriminative power of the feature embeddings, thereby enhancing the performance of our distance-based classifier.

A Separability Loss To encourage greater separation between different classes, we propose a separability loss that is designed to maximize the inter-class Euclidean distance. By maximizing this distance, the separability loss aims to increase the margin between classes in distance-based feature space. The separability loss is defined as follows:

$$L_{\text{sep}}=\frac{1}{N}\sum_{j=1}^{N}\frac{1}{N-1}\sum_{k=1}^{N}\mathbf{1}_{[j\neq k]}\exp(\|\hat{\boldsymbol{w}}_j-\hat{\boldsymbol{w}}_k\|_2^2) \quad (7)$$

where $\hat{\boldsymbol{w}}_k=\boldsymbol{w}_k/\|\boldsymbol{w}_k\|_2$, $(k=1,2,\dots,N)$ is the normalized class center, $\mathbf{1}_{[j\neq k]}\in\{0,1\}$ is an indicator function evaluating to 1 if $j\neq k$ and 0 otherwise.

Inference

Temperature scaling has been demonstrated to effectively distinguish between softmax scores for ID and OOD images, thereby enabling reliable OOD detection. In our proposed method, we adopt the temperature scaling approach that was introduced in ODIN (Liang, Li, and Srikant 2017). Specifically,

$$S_k(\boldsymbol{x};T)=\frac{\exp\left(f_k(\boldsymbol{x})/T\right)}{\sum_{j=1}^{N}\exp\left(f_j(\boldsymbol{x})/T\right)} \quad (8)$$

where $T\in R^{+}$ is the temperature scaling parameter, $f_j(\boldsymbol{x})$ is the output value of the last fully connected layer of the j-th class. For our method, $f_j=\|\hat{\boldsymbol{z}}-\hat{\boldsymbol{w}}_{\boldsymbol{j}}\|_2^2$, $(j=1,2,\dots,N)$ without the scaling parameter s.

Furthermore, we incorporate a small perturbation ϵ to the input, as suggested in ODIN. This perturbation is added to the input data prior to feeding it into the neural network, with the aim of increasing the sensitivity of the classifier to subtle differences between ID and OOD images:

$$\tilde{\boldsymbol{x}}=\boldsymbol{x}-\varepsilon\operatorname{sign}\left(-\nabla_{\boldsymbol{x}}\log S_{\hat{y}}(\boldsymbol{x};T)\right) \quad (9)$$

where $S_{\hat{y}}(\boldsymbol{x};T)=\max_k S_k(\boldsymbol{x};T)$, $(k=1,2,\dots,N)$. It is worth noting that the perturbation can be straightforwardly computed by back-propagating the gradient of the loss w.r.t the input.

Experiments and Results

Common Setup

Datasets and Training Details Consistent with the common benchmarks in the literature, we use CIFAR-10 and CIFAR-100 (Krizhevsky, Hinton et al. 2009) as ID datasets. For OOD test datasets, nine different natural image datasets are employed, including CIFAR datasets, TinyImageNet (cropped and resized)(Deng et al. 2009a), LSUN (cropped and resized)(Yu et al. 2015), iSUN (Xu et al. 2015), SVHN (Netzer et al. 2011), Textures (Cimpoi et al. 2014), Places365 (Zhou et al. 2017). In our main experiments, the ResNet-18 is employed as the backbone for CIFAR-10, while the ResNet-34 is used for CIFAR-100. All experiments are repeated five times, and we report the average performance. Further details regarding the training process can be found in Appendix B.1.

Hyperparameter Tuning In the proposed method, two hyperparameters play a crucial role: the temperature scaling parameter, denoted by T, and the input perturbation parameter, denoted by ϵ. More details of Hyperparameter tuning are given in Appendix B.2.

Evaluation Metrics We report the following metrics: (1) the false positive rate (**FPR95**) of OOD samples when the true positive rate of ID samples is at 95%, (2) the area under the receiver operating characteristic curve (**AUROC**), and (3) the area under the precision-recall (**AUPR**), where we treat success/normal classes as positive, (4) In-distribution test error (**IDTE**).

Main Results and Discussion

Comparison with Post-hoc OOD Detection Methods Table 1 presents a comprehensive comparison of competitive OOD detection methods. For pre-trained model-based approaches such as MSP (Hendrycks and Gimpel 2016), ODIN (Liang, Li, and Srikant 2017), Mahalanobis (Lee et al. 2018), Energy (Liu et al. 2020), and ReAct (Sun, Guo, and Li 2021), the standard practice is to train the model using the softmax CE loss. As demonstrated in Table 1, the proposed CASE method yields significant improvements in both OOD detection performance and ID classification accuracy. Due to space constraints, the average performance across nine OOD datasets is reported. The expanded version is presented in Table 5 in Appendix C. The results indicate that CASE reduces the average FPR95 by **50.47**%, enhances the average AUROC by **9.55**%, and improves the average AUPR by **7.83**% compared to the baseline MSP (Hendrycks and Gimpel 2016) on the CIFAR-10 model. Moreover, due to the improved embedding quality, CASE also reduces the ID test error (IDTE) by **3.41**%. Similarly, on CIFAR-100, CASE also achieves a considerable performance improvement.

CASE Outperforms Different Training Loss Functions In Table 2, we present a comparative analysis of different training loss functions. To isolate the impact of the loss function during training, we maintain the same OOD scoring function, namely ODIN score (Liang, Li, and Srikant 2017), for test-time evaluation. Our results show that for the more challenging CIFAR-100 is employed as the ID dataset, training with CASE reduces the average FPR95 by **29.84**% compared to the CE baseline. Furthermore, it improves the average AUROC by **10.56**% and the average AUPR by **10.61**%. Similarly, CASE significantly outperforms all the compared training loss functions with a considerable margin on the CIFAR-10 model. A more comprehensive version of Table 2 is available in Table 6 in Appendix C. These observations highlight the necessity and effectiveness of improving the intra-class compactness and inter-class separability of feature embeddings through the use of CASE loss.

In particular, different from the class center definition method used in recent work CESED (Feng et al. 2023), which exploits hyperspherical predefined evenly-distributed class centroids for OOD detection, we define class centers as a free optimization parameter of the model and updated by gradient descent in our proposed method. The results presented in Table 2 indicate that CASE reduces the average FPR95 by **5.13**% and improves the average AUROC by **1.18**% compared to CESED on the CIFAR-100 model. The difference in OOD detection performance between CASE and CESED underlines the effectiveness of the class center definition method adopted in our approach, which further promotes OOD detection performance.

CASE can Improve Existing Scoring Functions Table 3 highlights that our proposed CASE loss not only outperforms but also boosts competitive OOD scoring functions. All the OOD scoring functions considered in the comparison are originally developed based on models trained using the softmax CE loss, making them natural comparison candidates. Specifically, the following methods are examined: 1) MSP (Hendrycks and Gimpel 2016), which employs softmax confidence score for OOD detection. 2) ODIN (Liang, Li, and Srikant 2017), which leverages temperature scaling and input perturbation to improve OOD detection. 3) Energy score (Liu et al. 2020), which utilizes the information in logits for OOD detection, where logits represent the negative log of the denominator in the softmax function.

The results in Table 3 provide compelling evidence that the proposed CASE loss can significantly improve the performance of various downstream OOD scoring functions compared with CE loss and CESED loss. For instance, on the CIFAR-100 model, combined with the ODIN score, CASE loss reduces the average FPR95 by **29.82**% and **12.33**%, respectively. Additionally, the results also demonstrate that the CASE loss facilitates the softmax score and energy score to achieve favorable OOD detection performance. The OOD detection performance on each OOD test dataset can be found in Table 7 in Appendix C.

CASE Improves ID-OOD Separability Table 1, Table 2 and Table 3 present a noteworthy finding that our proposed method significantly improves the OOD detection performance. For example, compared with CESED, when evaluated against SVHN as OOD data, our approach reduces the FPR95 from **38.02**% to **30.99**%, corresponding to a direct improvement of **7.03**% on the more challenging CIFAR-100 model. To further illustrate the difference in OOD detection performance among CE, CESED and CASE, we present a visualization comparison of the distribution of ODIN scores for both ID and OOD data derived from networks trained with the three losses. Figure 2(a) and Figure 2(b) reveal that the ODIN scores for both ID and OOD data tend to cluster around high values when trained with CE loss and CESED loss. In contrast, the network trained with CASE loss results in an overall smoother distribution of ODIN scores for ID data, as evidenced by Figure 2(c). This implies that the use of CASE loss makes the ODIN scores more distinguishable between ID and OOD data, allowing for more effective OOD detection.

Ablation Studies

This section is dedicated to presenting the ablation results of the various factors that influence the performance of CASE. Specifically, we provide the ablation results on loss components, loss weights, network capacity and architecture based on the more challenging CIFAR-100 task, while acknowledging that similar trends apply to the less challenging CIFAR-10 task. Meanwhile, we delve deeper into CASE's performance on more challenging large-scale benchmarks. It is worth noting that each ablation study isolates the impact of a single factor by maintaining the other hyperparameters at their default settings (refer to Section).

ID	Method	**FPR95↓**	**AUROC↑**	**AUPR↑**	**IDTE↓**
CIFAR-10 (ResNet-18)	MSP(Hendrycks and Gimpel 2016)	70.65	86.57	88.39	8.34
	ODIN (Liang, Li, and Srikant 2017)	52.53	88.41	88.32	8.34
	Mahalanobis(Lee et al. 2018)	47.22	86.77	86.38	8.34
	Energy(Liu et al. 2020)	59.21	87.74	88.41	8.34
	ReAct(Sun, Guo, and Li 2021)	52.81	89.76	90.30	8.40
	CASE(ours)	**20.18**$^{\pm 1.3}$	**96.12**$^{\pm 0.18}$	**96.22**$^{\pm 0.13}$	**4.93**$^{\pm 0.09}$
CIFAR-100 (ResNet-34)	MSP(Hendrycks and Gimpel 2016)	84.45	72.81	74.74	30.45
	ODIN (Liang, Li, and Srikant 2017)	77.18	78.14	78.70	30.45
	Mahalanobis (Lee et al. 2018)	60.38	81.56	81.32	30.45
	Energy (Liu et al. 2020)	81.00	77.33	78.50	30.45
	ReAct(Sun, Guo, and Li 2021)	75.52	80.62	82.05	31.32
	CASE(ours)	**47.34**$^{\pm 0.93}$	**88.70**$^{\pm 0.10}$	**89.31**$^{\pm 0.25}$	**25.44**$^{\pm 0.46}$

Table 1: Comparison with competitive post-hoc OOD detection methods. ↑ indicates larger values are better and ↓ indicates smaller values are better. The reported values are percentages averaged across the nine OOD test datasets described in Section , and the standard deviation is presented in the upper right corner. Bold numbers are superior results.

ID	training loss	**FPR95↓**	**AUROC↑**	**AUPR↑**
CIFAR-10 (ResNet-18)	CE	52.53	88.41	88.32
	G-ODIN(Hsu et al. 2020)	**17.10**	**96.75**	**97.01**
	IsoMax+ (Macêdo and Ludermir 2021)	25.59	94.31	94.27
	LogitNorm(Wei et al. 2022)	50.75	88.18	86.80
	CESED (Feng et al. 2023)	25.29	95.33	96.63
	CSAE(ours)	20.18	96.12	96.22
CIFAR-100 (ResNet-34)	CE	77.18	78.14	78.70
	G-ODIN (Hsu et al. 2020)	71.45	81.91	83.49
	IsoMax+ (Macêdo and Ludermir 2021)	53.42	86.23	86.74
	LogitNorm (Wei et al. 2022)	81.83	74.57	75.32
	CESED (Feng et al. 2023)	52.47	87.52	88.25
	CASE(ours)	**47.34**	**88.70**	**89.31**

Table 2: OOD detection performance comparison with different training loss functions. All values are percentages averaged over nine OOD test datasets. Bold numbers are superior results.

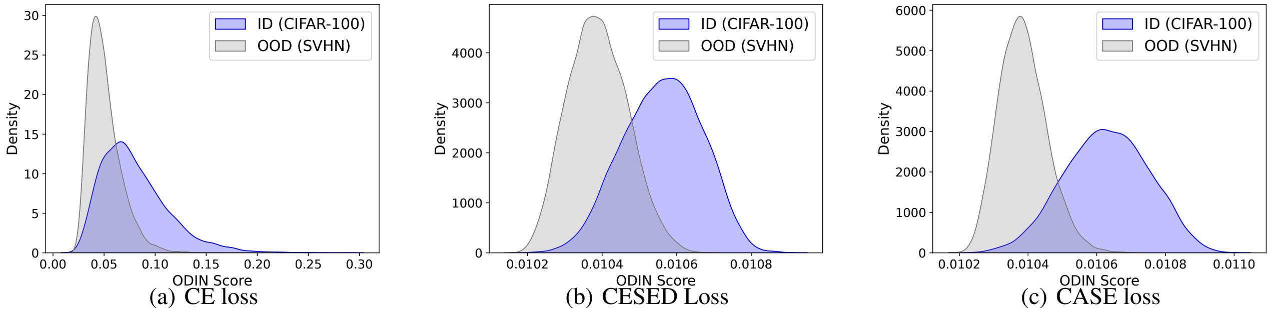
Ablation on Different Loss Components We investigate the effects of loss components on OOD detection. The results in Table 4 suggest that $\mathcal{L}_{comp}$ and $\mathcal{L}_{sep}$ work together to improve intra-class compactness and inter-class separability of feature embeddings, which are desirable for both ID classification and OOD detection. Specifically, for ID classification, training with both $\mathcal{L}_{comp}$ and $\mathcal{L}_{sep}$ reduces the average ID test error from **26.22**% to **25.44**% compared to training with $\mathcal{L}_{comp}$ alone, indicating that promoting inter-class separability and a moderate level of intra-class compactness can improve the ability to discriminate ID classes. Regarding OOD detection, higher inter-class separability proves to be beneficial, as explicitly encouraged through the separability loss $\mathcal{L}_{sep}$. Therefore, incorporating $\mathcal{L}_{sep}$ leads to a reduction in the average FPR95 by **5.24**%, an improvement in the average AUROC by **1.15**%, and an enhancement in the average AUPR by **1.01**%. Nevertheless, improving inter-class separability via $\mathcal{L}_{sep}$ alone without $\mathcal{L}_{comp}$ is inadequate for both ID classification and OOD detection.

Ablation on the Loss Weights Figure 3(a) and Figure 3(b) shown in Appendix D illustrate the influence of different loss weights λ_c on the OOD detection performance of CASE. The results reveal that the performance of CASE remains relatively stable for moderate adjustments of λ_c (e.g., from 0.5 to 2), with the best FPR95 of **47.34**% and the best AUROC of **88.70**% achieved at approximately λ_c = 1.0. As shown in Table 1, we have demonstrated the effectiveness of CASE, where the loss weight λ_c is set to 1.0 by default to balance the initial scale between the $\mathcal{L}_{com}$ and $\mathcal{L}_{sep}$. This implies that CASE provides a simple and effective solution for boosting OOD detection performance, without requiring excessive hyperparameter tuning of the loss scale.

Ablation on Network Capacity and Architecture To evaluate the effectiveness and generality of CASE, we conducted experiments using two different network architectures: ResNet-50 and DenseNet-101. The results of the ResNet-50 experiments are presented in Table 8 in Appendix D, where we observe that training with CASE improves the quality of feature embeddings, leading to a reduction in the average FPR95 by **19.13**%, an improvement in the average AUROC by **3.94**%, and an enhancement in the average AUPR by **3.48**% compared to the CESED method (Feng et al. 2023). We then verify the generality of CASE by repeating the experiments using the DenseNet-101 architecture. The results shown in Table 9 in Appendix D demonstrate that CASE also outperforms all compared methods in terms of the OOD detection performance.

ID	CIFAR-10			CIFAR-100		
Score	**FPR95↓**	**AUROC↑**	**AUPR↑**	**FPR95↓**	**AUROC↑**	**AUPR↑**
	CE loss / CESED loss / **CASE loss(ours)**					
Softmax	70.65/26.37/**23.33**	86.57/95.13/**95.72**	88.39/95.51/**96.03**	84.45/79.08/**65.83**	72.81/80.52/**85.73**	74.74/82.98/**87.18**
Energy	59.21/23.50/**20.42**	87.74/95.35/**95.83**	88.41/95.53/**95.98**	81.00/80.54/**49.48**	77.33/73.01/**84.50**	78.50/73.58/**84.16**
ODIN	52.53/25.07/**20.18**	88.41/95.35/**96.12**	88.32/95.68/**96.22**	77.18/59.67/**47.34**	78.14/85.67/**88.70**	78.70/86.46/**89.31**

Table 3: OOD detection performance comparison with various scoring functions using CE loss, CESED loss and CASE loss. All values are percentages averaged over nine OOD test datasets. Bold numbers are superior results.



(a) CE loss (b) CESED Loss (c) CASE loss

Figure 2: Distribution of ODIN scores from ResNet-34 trained on CIFAR-100 with (a) Cross-Entropy (CE) loss, (b) CESED loss and (c) CASE loss(ours).

Loss Components		**FPR95↓**	**AUROC↑**	**AUPR↑**	**IDTE↓**
$\mathcal{L}_{comp}$	$\mathcal{L}_{sep}$				
	✓	94.35	48.51	49.05	99.06
✓		52.58	87.65	88.30	26.22
✓	✓	**47.34**	**88.70**	**89.31**	**25.44**

Table 4: Ablation on CASE's different components. All values are percentages averaged over nine OOD test datasets. Bold numbers are superior results.

CASE is Competitive on Large-scale Datasets To further examine the performance of CASE on real-world tasks, we analyzed its performance on more challenging and realistic large-scale datasets. Specifically, we utilize ImageNet-10 as the ID dataset, which is a subset of ImageNet (Deng et al. 2009b) comprising 10 randomly selected classes. For OOD test datasets, we adopt the same one used in (Huang and Li 2021), which include subsets of iNaturalist (Van Horn et al. 2018), SUN (Xiao et al. 2010), Places365 (Zhou et al. 2017) , and Textures (Cimpoi et al. 2014). Importantly, each OOD dataset's categories do not overlap with those in the ID dataset. Figure 4(a) and Figure 4(b) shown in Appendix D illustrate the results in terms of FPR95 and AUROC, respectively. Our findings indicate that CASE consistently outperforms CE and CESED on all OOD test sets, demonstrating the advantages of improving the intra-class compactness and inter-class separability of feature embeddings with a trainable class center parameter.

Conclusion

In this paper, we propose CASE, a simple and effective method for OOD detection by improving intra-class compactness and inter-class separability of feature embeddings. CASE simultaneously optimizes a compactness loss and a separability loss to promote the separation between ID and OOD samples. The separability loss maximizes the inter-class Euclidean distance to encourage separability among different class centers, while the compactness loss minimizes the intra-class Euclidean distance to encourage samples to be close to their respective class centers. In particular, the class centers are defined as a free optimization parameter of the model and updated by gradient descent. Under this new training method, a distance-based representation space can be achieved to benefit OOD detection. Extensive evaluations and ablation studies demonstrate that CASE can improve both OOD detection and ID classification accuracy. We hope that our insights inspire further research on distance-based representation learning and loss function design for enhancing OOD detection.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China (Grant No. 62192783, 62376117), the Collaborative Innovation Center of Novel Software Technology and Industrialization at Nanjing University.

References

Chen, J.; Li, Y.; Wu, X.; Liang, Y.; and Jha, S. 2021. Atom: Robustifying out-of-distribution detection using outlier mining. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21*, 430–445. Springer.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3606–3613.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009a. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009b. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Drummond, N.; and Shearer, R. 2006. The open world assumption. In *eSI Workshop: The Closed World of Databases meets the Open World of the Semantic Web*, volume 15, 1.

Feng, S.; Jiang, W.; Chen, M.; Du, Y.; Cheng, H.; Ge, Y.; and Wang, C. 2023. CESED: Exploiting Hyperspherical Predefined Evenly-Distributed Class Centroids for OOD Detection. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, 262–270. SIAM.

Geifman, Y.; and El-Yaniv, R. 2019. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, 2151–2159. PMLR.

Hassen, M.; and Chan, P. K. 2020. Learning a neural-network-based representation for open set recognition. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, 154–162. SIAM.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

He, R.; Han, Z.; Lu, X.; and Yin, Y. 2022. RONF: reliable outlier synthesis under noisy feature space for out-of-distribution detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4242–4251.

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

Hendrycks, D.; Mazeika, M.; and Dietterich, T. 2018. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*.

Hsu, Y.-C.; Shen, Y.; Jin, H.; and Kira, Z. 2020. 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*, 10951–10960.

Huang, R.; Geng, A.; and Li, Y. 2021. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34: 677–689.

Huang, R.; and Li, Y. 2021. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8710–8719.

Kirchheim, K.; Filax, M.; and Ortmeier, F. 2022. Multi-Class Hypersphere Anomaly Detection. In *2022 26th International Conference on Pattern Recognition (ICPR)*, 2636–2642. IEEE.

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

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2017. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6): 84–90.

Lee, K.; Lee, H.; Lee, K.; and Shin, J. 2017. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*.

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

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

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33: 21464–21475.

Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; and Song, L. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 212–220.

Liu, W.; Wen, Y.; Yu, Z.; and Yang, M. 2016. Large-margin softmax loss for convolutional neural networks. *arXiv preprint arXiv:1612.02295*.

Luo, Y.; Wong, Y.; Kankanhalli, M.; and Zhao, Q. 2019. $\mathcal{G}$-softmax: improving intraclass compactness and interclass separability of features. *IEEE transactions on neural networks and learning systems*, 31(2): 685–699.

Macêdo, D.; and Ludermir, T. 2021. Enhanced isotropy maximization loss: Seamless and high-performance out-of-distribution detection simply replacing the softmax loss. *arXiv preprint arXiv:2105.14399*.

Macêdo, D.; Ren, T. I.; Zanchettin, C.; Oliveira, A. L.; and Ludermir, T. 2021. Entropic out-of-distribution detection. In *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Miller, D.; Sunderhauf, N.; Milford, M.; and Dayoub, F. 2021. Class anchor clustering: A loss for distance-based open set recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3570–3578.

Ming, Y.; Fan, Y.; and Li, Y. 2022. Poem: Out-of-distribution detection with posterior sampling. In *International Conference on Machine Learning*, 15650–15665. PMLR.

Ming, Y.; Sun, Y.; Dia, O.; and Li, Y. 2023. How to Exploit Hyperspherical Embeddings for Out-of-Distribution Detection? In *The Eleventh International Conference on Learning Representations*.

Mohseni, S.; Pitale, M.; Yadawa, J.; and Wang, Z. 2020. Self-supervised learning for generalizable out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5216–5223.

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

Phua, C.; Lee, V.; Smith, K.; and Gayler, R. 2010. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*.

Sun, Y.; Guo, C.; and Li, Y. 2021. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34: 144–157.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8769–8778.

Wang, F.; Cheng, J.; Liu, W.; and Liu, H. 2018. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7): 926–930.

Wei, H.; Xie, R.; Cheng, H.; Feng, L.; An, B.; and Li, Y. 2022. Mitigating neural network overconfidence with logit normalization. In *International Conference on Machine Learning*, 23631–23644. PMLR.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, 3485–3492. IEEE.

Xu, P.; Ehinger, K. A.; Zhang, Y.; Finkelstein, A.; Kulkarni, S. R.; and Xiao, J. 2015. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*.

Yang, J.; Zhou, K.; Li, Y.; and Liu, Z. 2021. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*.

Yu, F.; Seff, A.; Zhang, Y.; Song, S.; Funkhouser, T.; and Xiao, J. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6): 1452–1464.