Article

Ensemble-Based Out-of-Distribution Detection

Donghun Yang ^{1,2} , Kien Mai Ngoc ^{1,3} , Iksoo Shin ^{4,5}, Kyong-Ha Lee ^{1,2} and Myunggwon Hwang ^{1,2,*}

- ¹ Department of Data and HPC Science, University of Science and Technology, Daejeon 34113, Korea; yangdonghun3@kisti.re.kr (D.Y.); kienmn@kisti.re.kr (K.M.N.); kyongha@kisti.re.kr (K.-H.L.)
- ² Department of Intelligent Infrastructure Technology Research, Korea Institute of Science and Technology Information, Daejeon 34141, Korea
- ³ Research Data Sharing Center, Korea Institute of Science and Technology Information, Daejeon 34141, Korea
- ⁴ Department of ICT, University of Science and Technology, Daejeon 34113, Korea; sis@etri.re.kr
- ⁵ High Performance Embedded SW Research Section, Electronics and Telecommunications Research Institute, Daejeon 34129, Korea
- * Correspondence: mgh@kisti.re.kr; Tel.: +82-42-869-1631

Abstract: To design an efficient deep learning model that can be used in the real-world, it is important to detect out-of-distribution (OOD) data well. Various studies have been conducted to solve the OOD problem. The current state-of-the-art approach uses a confidence score based on the Mahalanobis distance in a feature space. Although it outperformed the previous approaches, the results were sensitive to the quality of the trained model and the dataset complexity. Herein, we propose a novel OOD detection method that can train more efficient feature space for OOD detection. The proposed method uses an ensemble of the features trained using the softmax-based classifier and the network based on distance metric learning (DML). Through the complementary interaction of these two networks, the trained feature space has a more clumped distribution and can fit well on the Gaussian distribution by class. Therefore, OOD data can be efficiently detected by setting a threshold in the trained feature space. To evaluate the proposed method, we applied our method to various combinations of image datasets. The results show that the overall performance of the proposed approach is superior to those of other methods, including the state-of-the-art approach, on any combination of datasets.

Keywords: out-of-distribution detection; confidence score; distance metric learning; Siamese network; triplet network; ensemble method; deep learning

Citation: Yang, D.; Ngoc, K.M.; Shin, I.; Lee, K.-H.; Hwang, M. Ensemble-Based Out-of-Distribution Detection. *Electronics* **2021**, *10*, 567. <https://doi.org/10.3390/electronics10050567>

Academic Editor: Jun Liu and Byung-Gyu Kim

Received: 4 February 2021

Accepted: 23 February 2021

Published: 28 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Deep learning has achieved state-of-the-art performance in various tasks, such as speech recognition [1,2], image classification [3,4], video prediction [5,6] and medical diagnosis [7,8]. Nevertheless, several problems with deep learning remain. This study is focused on two of them. The first is the closed-world assumption. Contemporary deep learning models are designed under the static and closed-world assumption that training and testing datasets have the same distribution [9]. However, in the real-world, data distributions may undergo complex and dynamic shifts over time, and even a novel dataset with an unseen distribution might be presented to the model during the test. These shifted and unseen data distributions may cause critical failures because the model attempts to predict the results under the closed-world assumption [10]. The second is the high confidence problem. It is generally known that modern deep learning models may yield improper predictions with high confidence even for unseen data distributions [11]. These problems, which are called out-of-distribution (OOD) problems [12], cause overfitting and complicate the calibration of deep learning models [13,14]. Therefore, to design an efficient deep learning model that can be used in the real-world, it is important to detect the OOD data well.

Various studies have been conducted to solve the OOD problem. A baseline model has been proposed to detect OOD data using a neural network's softmax value as a confidence score [12]. As an extension of the baseline method, the out-of-distribution detector for neural networks (ODIN) has been proposed to improve performance using temperature scaling and input preprocessing [15]. ODIN outperformed the baseline method; however, this method required hyperparameters to be tuned appropriately for each dataset. Additionally, approaches based on generative models and auxiliary datasets have been proposed [16–18]. The current state-of-the-art method uses confidence scores based on the Mahalanobis distance in a feature space [19]; however, its results are sensitive to the dataset complexity and the quality of the trained model. In that respect, we proposed the OOD detection method based on distance metric learning (DML) in our previous research [20]. This method can train a clumped feature space (in which data with the same label are located closely) by class using a DML-based network (instead of the softmax-based classifier) and can detect OOD samples efficiently in that feature space. Our previous method outperformed the state-of-the-art approach in 1-channel image datasets having relatively simple structures. However, it could not detect OOD well in the 3-channel image datasets having complex structures.

Herein, we propose a novel OOD detection method that uses not only DML-based networks but also softmax-based classifiers, as an extended version of our previous work. The proposed method can obtain more efficient feature space for OOD detection by an ensemble of the features trained using the softmax-based classifier and the DML-based networks, including Siamese and triplet networks [21,22]. The trained feature space has a more clumped distribution and can fit better on the Gaussian distribution by class, compared with using the state-of-the-art approach and our previous method. An example of the trained feature spaces is shown in Figure 1. Moreover, in the testing phase, the OOD data can be detected as follows: (1) measure the distance between the features of the input data and each class distribution as a confidence score and (2) set a threshold in that distance.

Figure 1. Examples of feature spaces trained using the softmax-based classifier and distance metric learning (DML)-based networks.

To evaluate the proposed OOD detection module, we applied our method to various combinations of 1-channel and 3-channel image datasets. Subsequently, we verified the performance of the OOD detection by comparing it with the previous approaches.

The remainder of this paper is organized as follows. Section 2 presents the related studies, Section 3 describes the proposed OOD detection method and Section 4 details our experiment. The paper is concluded in Section 5.

2. Related Work

In this section, we introduce several OOD detection methods and DML-based networks used in the proposed method and our previous approach.

2.1. OOD Detection Methods

This study mainly focuses on the OOD detection method based on the confidence score and threshold (among various OOD-related studies). Therefore, this section reviews some threshold-based OOD detection methods relevant for this study. The baseline method of OOD detection [12] was proposed based on the tendency of a well-trained neural network to assign a higher softmax score to in-distribution examples rather than OOD examples. In this approach, the OOD data can be detected using softmax as a confidence score and applying a threshold to it. The softmax function is shown in Equation (1); it comprises $f_i(x)$ for the logit of class i .

$$S_{baseline}(x) = \max_i \frac{\exp(f_i(x))}{\sum_{j=1}^C \exp(f_j(x))} \quad (1)$$

The ODIN method [15] was proposed to improve the performance of the baseline by using temperature scaling and adding small controlled perturbations to the input data. The temperature scaling T was applied to the baseline confidence scoring function, as shown in Equation (2). The ODIN method outperformed the baseline method; however, it required hyperparameters to be tuned appropriately for each dataset.

$$S_{ODIN}(x) = \max_i \frac{\exp(f_i(x)/T)}{\sum_{j=1}^C \exp(f_j(x)/T)} \quad (2)$$

Mahalanobis-based Approach [19], which demonstrates the state-of-the-art performance, uses the confidence score based on the Mahalanobis distance in a feature space. This approach was designed under the assumption that the well-trained output features of the softmax-based neural classifier can be fitted well to the class-conditional Gaussian distribution. The confidence score can be defined by calculating the Mahalanobis distance using the class mean and covariance of the feature map, thereby enabling the effective detection of OOD samples. Although this method outperformed the previous approaches, the results were sensitive to the dataset complexity and the quality of the trained model.

The DML-based approach [20] was proposed in our previous research. To train a more efficient feature space than the state-of-the-art approach, this method uses DML-based networks (as described in the next section) instead of the softmax-based classifier. The trained feature space has a more clumped distribution by class, and OOD samples can be detected by applying a threshold to this feature space. This method performed well in 1-channel image datasets with relatively simple structures. However, it could not detect OOD well in 3-channel image datasets with complex structures.

2.2. Networks Based on Distance Metric Learning (DML)

DML is a branch of machine learning algorithms that aims to learn similarities between data samples using a distance-based loss function [23]. As this method embeds similar data samples closer, DML-based networks can train more clumped feature spaces by class. In this section, we introduce two DML-based networks used in the proposed method and our previous research.

The Siamese network [21] comprises one cost function and two sub-networks that share parameters and have the same structure. When training the Siamese network, two inputs are passed through the sub-networks. One is an anchor input x_a , and the other can be a positive input x_p with the same label as the anchor or a negative input x_n with a different label. After the sub-networks, the distance between two output features is calculated using the cost function. The Siamese network uses contrastive loss as a cost function; hence, inputs with the same label are embedded closely in the feature space. In

the opposite case, they can be distant from one another when training the network. The contrastive loss function is shown in Equation (3), where M is a constant value

$$L_{contrastive} = \begin{cases} \|x_a - x_p\|_2^2 & \text{if Positive Pair} \\ \max(0, M - \|x_a - x_n\|_2^2) & \text{if Negative Pair} \end{cases} \quad (3)$$

In the testing phase, the test dataset is entered into one of the trained sub-networks; subsequently, the clumped vectors can be obtained by class. The Siamese network structure is shown in Figure 2a.

The triplet Network [22] is a network based on the Siamese network. It comprises a triplet loss function and three sub-networks. The triplet loss function is shown in Equation (4).

$$L_{triplet} = \max(0, \|x_a - x_p\|_2^2 - \|x_a - x_n\|_2^2 + M) \quad (4)$$

When training the triplet network, three inputs are provided: the anchor input x_a , positive input x_p and negative input x_n . Using these three inputs and the triplet loss function, the anchor input can be embedded farther from the negative input than the positive input. After training, the test can be performed as in the Siamese network. The structure of the triplet network is shown in Figure 2b.

Figure 2. Structures of DML-based networks.

3. Methodology

In this section, we present our proposed method for detecting OOD samples. Our method is proposed to improve the Mahalanobis-based approach [19], which is the current state-of-the-art method, and the DML-based approach [20], which is our previous work. The state-of-the-art approach uses only a softmax-based classifier, and our previous approach uses only a DML-based network to train the feature space. Figure 1 shows well-trained feature spaces with a softmax-based classifier and DML-based networks. In these feature spaces, we may detect OOD well using these approaches. However, there is no guarantee that the networks are always trained well. If the networks are not trained well, we cannot detect the OOD well in that feature space. Therefore, in this study, we used both networks together—not only softmax-based classifier but also DML-based network—to train more efficient feature space for OOD detection. In the proposed method, the feature space is trained by an ensemble of the features trained using the softmax-based classifier and the DML-based network. With complementary interaction between the two networks by the ensemble, the trained feature space has a more clumped distribution, and it can better fit on the Gaussian distribution by class. Thus, OOD samples can be efficiently detected in that feature space. Figure 3 shows the overall structure of our proposed method.

Figure 3. Overall structure of the proposed out-of-distribution (OOD) detection method.

Except for the networks that train feature spaces, the state-of-the-art protocols (such as input preprocessing, Mahalanobis-based confidence score and feature ensemble) were also used to efficiently detect OOD samples in this study.

Input Preprocessing [15]. In the testing phase, to increase the confidence score based on the Mahalanobis distance, the input preprocessing technique is applied, in which a small controlled noise is added to the test samples. This technique results in a more separable in-distribution and OOD samples. The preprocessed test samples are obtained by Equation (5), where x represents the test sample, ϵ is the magnitude of noise and $M(x)$ is the confidence score based on the Mahalanobis distance

$$\hat{x} = x + \epsilon \text{sign}(\nabla_x M(x)) \quad (5)$$

Confidence Score based on Mahalanobis Distance [19]. The Mahalanobis distance between the test sample and the closest class distribution is used as a confidence score. The Mahalanobis distance of the l -th layer, $M_l(x)$, is calculated using Equation (6), where c is the class index and $f^l(x)$ is the feature of the test sample at the l -th layer; μ and Σ are the class mean and covariance matrix, respectively

$$M_l(x) = \max_c - (f^l(x) - \mu_c^l)^T \Sigma_l^{-1} (f^l(x) - \mu_c^l) \quad (6)$$

Feature Ensemble [19]. In the state-of-the-art approach, the feature ensemble technique was used to calculate the weighted sum of confidence scores from the feature set in some layers. Using this technique, we can ensemble the features trained using the softmax-based classifier and the DML-based network. Moreover, we can also measure and combine the confidence scores of the final feature and the other low-level features in the two networks. This means that effective layers can be assigned a higher weight, and ineffective layers can be assigned a lower weight. This is expressed as Equation (7), where M_{l_s} and α_{l_s} are the confidence score and its weight obtained from the feature set of the l -th layer in the softmax-based classifier, and M_{l_d} and α_{l_d} are the confidence score and its weight obtained from the feature set of the l -th layer in the DML-based network. In the experiments of this study, both weights were trained by logistic regression using a small validation dataset that consisted of 1000 images from each in- and out-of-distribution pair, similar to [19]. Here, $M(x)$ is the total confidence score based on the Mahalanobis distance.

$$M(x) = \sum_{l_s} \alpha_{l_s} M_{l_s}(x) + \sum_{l_d} \alpha_{l_d} M_{l_d}(x) \quad (7)$$

Figure 4 shows the overall process of the proposed method. In the training phase, features are extracted from the training samples using both the softmax-based classifier

and DML-based network trained on in-distribution dataset. Subsequently, the mean and covariance are calculated for each class from the extracted features.

Figure 4. Overall process of the proposed OOD detection method.

In the testing phase, features are extracted from the test samples consisting of the same ratio of in- and out-of-distribution datasets, with a small amount of controlled noise added. Thereafter, the Mahalanobis distance between the test samples and the closest class distribution is calculated using the class mean and covariance. The calculated Mahalanobis distances from the output features of several layers in the two networks are ensembled. Finally, OOD samples can be detected by applying a threshold to the ensembled Mahalanobis distance.

4. Experiments

In this section, the performance of the proposed OOD detection method is evaluated, analyzed and compared with the previous approaches on various combinations of datasets.

4.1. Experimental Setup

First of all, in order to fairly benchmark our method, we selected various standard datasets which have been most actively used for evaluating and comparing OOD detection methods until now [15–20]. For 1-channel image datasets, we chose Fashion-MNIST [24] and MNIST [25]. For 3-channel image datasets, we chose CIFAR-10 [4], CIFAR-100 [4], SVHN [26], Tiny ImageNet [27] and LSUN [28]. These datasets reflect real-world scenarios, such as handwriting, fashion, street signs and so on. Especially, the CIFAR and ImageNet datasets can be thought as collections containing most of the real-life images.

All proposed methods were implemented using Python 3.7 and PyTorch 1.5 on NVIDIA TITAN RTX 24 GB × 2. In the case of experiments for 1-channel image datasets, we used ResNet34 [3] for the softmax-based classifier and the ResNet34-based Siamese or triplet network (described in Section 2) for the DML-based network. These networks were trained with a learning rate of 0.001, a batch size of 32 and an Adam optimizer. For the experiments on 3-channel image datasets, we used the trained ResNet34 on each dataset, provided in the state-of-the-art approach (https://github.com/pokaxpoka/deep_Mahalanobis_detector (accessed on 27 February 2021)), as a softmax-based classifier. ResNet34-based Siamese or triplet network was also used for DML-based network. We trained these DML-based networks with a batch size of 256 and an Adam optimizer. The

learning rate was initialized at 0.001 and decreased to 0 by a cosine scheduler [29]. In addition, we used an early stopping method to prevent overfitting [30]. Subsequently, we applied our method to various combinations of standard datasets mentioned above.

For performance comparison, we considered the baseline model, ODIN, the Mahalanobis-based approach, and our previous method (as described in Section 2). Our previous method was trained using DML-based networks, and the other models were trained using ResNet34 in the same way as the proposed method. In the case of experiments for 3-channel image datasets, the trained ResNet34 (provided in the state-of-the-art approach (https://github.com/pokaxpoka/deep_Mahalanobis_detector (accessed on 27 February 2021))) for each dataset was used for those models, excluding our previous approach.

To evaluate the performance, the following metrics were used: the true negative rate (TNR) at a 95% true positive rate (TPR), the area under the receiver operating characteristic curve (AUROC), the detection accuracy (DTACC) and the area under the precision–recall curve (AUPR). Using these metrics, the performance of OOD detection methods can be evaluated without selecting a specific threshold [15]. Our source code is available on GitHub (https://github.com/yangdonghun3/Ensemble_based_OOD_Detection (accessed on 27 February 2021)).

We detail our experimental results on 1-channel and 3-channel image datasets in the next subsections.

4.2. Experimental Results on 1-Channel Image Datasets

Figure 5 presents the average performances of the proposed method and the other methods, in 10 experiments, from the 10th to the 50th epoch on 1-channel image datasets. In the case of (In) Fashion-MNIST/(Out) MNIST, our proposed method (ensemble-based) and previous method (DML-based) were superior to the others at all epoch points and metrics, as shown in Figure 5a. Furthermore, the ensemble-triplet (navy line) version of the proposed method showed the best performance among all methods. In addition to their better performance, the triplet-based approaches (purple and navy lines) consistently demonstrated robust states during the entire training phase. Meanwhile, the other models demonstrated unstable performances depending on the epoch. Consequently, the proposed method and our previous method can be considered as less sensitive when selecting hyperparameters for OOD detection in this experiment. Comparing ours, triplet-based approaches (purple and navy lines) performed better than the Siamese-based approaches (yellow and green lines). Table 1 presents the average performance at the epoch point, showing the best TNR in several experiments on 1-channel image datasets. Table 1a also shows that the proposed method and our previous method outperformed the other methods, including the state-of-the-art approach.

Table 1. Average performances at the epoch point, showing the best true negative rates (TNR) in 10 experiments on 1-channel image datasets (mean $\pm$ standard deviation).

(a) In: Fashion-MNIST/Out: MNIST							
Metric	Baseline	ODIN	Mahal.	OnlySiamese (Our Previous)	OnlyTriplet (Our Previous)	Ensemble-Siamese (Ours)	Ensemble-Triplet (Ours)
TNR at TPR 95%.	43.08 $\pm$ 8.45	76.84 $\pm$ 7.16	94.53 $\pm$ 2.84	97.22 $\pm$ 3.61	99.90 $\pm$ 0.10	99.37 $\pm$ 0.50	99.96 $\pm$ 0.04
AUROC.	86.75 $\pm$ 5.32	93.87 $\pm$ 1.85	98.18 $\pm$ 0.53	98.97 $\pm$ 0.63	99.76 $\pm$ 0.09	99.48 $\pm$ 0.21	99.82 $\pm$ 0.11
DTACC.	83.23 $\pm$ 3.45	87.71 $\pm$ 3.01	95.23 $\pm$ 1.00	96.57 $\pm$ 1.27	98.63 $\pm$ 0.28	97.85 $\pm$ 0.69	99.03 $\pm$ 0.34
AUPRin.	79.74 $\pm$ 9.23	92.12 $\pm$ 4.59	98.57 $\pm$ 0.47	99.22 $\pm$ 0.46	99.80 $\pm$ 0.08	99.58 $\pm$ 0.19	99.87 $\pm$ 0.08
AUPRout.	86.18 $\pm$ 4.03	94.64 $\pm$ 1.48	98.74 $\pm$ 1.01	98.04 $\pm$ 1.15	99.62 $\pm$ 0.13	98.92 $\pm$ 0.34	99.61 $\pm$ 0.35
(b) In: MNIST/Out: Fashion-MNIST							
Metric	Baseline	ODIN	Mahal.	OnlySiamese (Our Previous)	OnlyTriplet (Our Previous)	Ensemble-Siamese (Ours)	Ensemble-Triplet (Ours)
TNR at TPR 95%.	98.42 $\pm$ 1.07	99.51 $\pm$ 0.25	100.0 $\pm$ 0.00	100.0 $\pm$ 0.00	100.0 $\pm$ 0.00	99.99 $\pm$ 0.03	100.0 $\pm$ 0.00
AUROC.	99.04 $\pm$ 0.18	99.55 $\pm$ 0.14	100.0 $\pm$ 0.00	100.0 $\pm$ 0.01	100.0 $\pm$ 0.00	99.99 $\pm$ 0.03	99.99 $\pm$ 0.00
DTACC.	96.92 $\pm$ 0.64	97.86 $\pm$ 0.40	99.98 $\pm$ 0.01	99.94 $\pm$ 0.03	99.94 $\pm$ 0.03	99.95 $\pm$ 0.03	99.94 $\pm$ 0.02
AUPRin.	98.71 $\pm$ 0.45	99.64 $\pm$ 0.10	99.99 $\pm$ 0.00	99.99 $\pm$ 0.00	99.99 $\pm$ 0.00	99.91 $\pm$ 0.24	99.99 $\pm$ 0.00
AUPRout.	98.74 $\pm$ 0.26	99.42 $\pm$ 0.20	99.99 $\pm$ 0.04	99.95 $\pm$ 0.05	99.99 $\pm$ 0.00	99.96 $\pm$ 0.04	99.98 $\pm$ 0.03

Turning to the experiment of (In) MNIST/(Out) Fashion-MNIST, except for the baseline method and ODIN, all other models showed nearly 100% OOD detection performance and stable states at all epoch points and all metrics, as shown in Figure 5b. This means that the three methods detected OOD samples perfectly during the entire training phase. Consequently, we considered that the proposed method, the previous method and the Mahalanobis-based approach could completely separate in-distribution samples and OOD samples because the structure of the MNIST dataset is simple. Table 1b also shows that the three methods detected the OOD samples perfectly.

(a) In: Fashion-MNIST/Out: MNIST

(b) In: MNIST/Out: Fashion-MNIST

Figure 5. Average performances in 10 experiments from 10 to 50 epochs on 1-channel image datasets.

4.3. Experimental Results on 3-Channel Image Datasets

To further verify the performance of the proposed method, additional experiments were performed on various combinations of 3-channel image datasets having more complex structures than 1-channel image datasets, and the results are reported in Tables 2–4. The tables show the average performance at the epoch point, showing the best TNR in 10 experiments on 3-channel image datasets. In the case of (In) SVHN, the proposed method, our previous approach and the Mahalanobis-based approach performed well on all combinations of datasets. In contrast, other models could not detect OOD well, as shown in Table 2. This table also shows that the overall performance of the proposed approaches (ensemble-Siamese, ensemble-triplet) was superior to the others, and the ensemble-triplet

method achieved the best TNR, approximately 0.16–1.40% higher than the state-of-the-art approach, among all methods. Moreover, except for OnlySiamese on (Out) CIFAR-10, our previous methods (OnlySiamese, OnlyTriplet) detected OOD well in most cases, compared with other combinations of 3-channel image datasets that have more complex structures.

Table 2. Average performances at the epoch point, showing the best TNR in 10 experiments on (In) SVHN/(Out) CIFAR-10, Tiny ImageNet and LSUN (mean $\pm$ standard deviation).

In-dist.	Out-dist.	Metric	Baseline	ODIN	Mahal.	OnlySiamese (Our Previous)	OnlyTriplet (Our Previous)	Ensemble-Siamese (Ours)	Ensemble-Triplet (Ours)
SVHN	CIFAR-10	TNR at TPR 95%.	78.26	79.83	97.63	82.46 $\pm$ 6.17	91.12 $\pm$ 3.72	98.45 $\pm$ 0.28	99.00 $\pm$ 0.45
		AUROC.	92.92	92.09	99.04	96.34 $\pm$ 1.31	98.08 $\pm$ 0.59	99.20 $\pm$ 0.08	99.46 $\pm$ 0.19
		DTACC.	90.03	89.44	96.37	91.10 $\pm$ 1.87	93.80 $\pm$ 1.20	96.85 $\pm$ 0.17	97.42 $\pm$ 0.66
		AUPRin.	95.06	93.96	99.58	98.38 $\pm$ 0.91	99.26 $\pm$ 0.24	99.65 $\pm$ 0.11	99.79 $\pm$ 0.10
		AUPRout.	85.66	86.83	96.15	89.04 $\pm$ 3.10	93.85 $\pm$ 1.75	96.55 $\pm$ 0.13	97.59 $\pm$ 0.78
SVHN	Tiny ImageNet	TNR at TPR 95%.	79.02	82.10	99.78	99.34 $\pm$ 0.34	99.76 $\pm$ 0.11	99.90 $\pm$ 0.04	99.94 $\pm$ 0.02
		AUROC.	93.51	91.99	99.82	99.73 $\pm$ 0.12	99.89 $\pm$ 0.04	99.87 $\pm$ 0.04	99.94 $\pm$ 0.03
		DTACC.	90.44	89.35	98.70	98.18 $\pm$ 0.42	99.00 $\pm$ 0.26	98.98 $\pm$ 0.13	99.21 $\pm$ 0.14
		AUPRin.	95.68	93.88	99.87	99.88 $\pm$ 0.07	99.95 $\pm$ 0.03	99.92 $\pm$ 0.03	99.97 $\pm$ 0.03
		AUPRout.	86.18	88.12	99.06	98.99 $\pm$ 0.46	99.31 $\pm$ 0.33	99.23 $\pm$ 0.17	99.48 $\pm$ 0.31
SVHN	LSUN	TNR at TPR 95%.	74.29	77.34	99.77	99.75 $\pm$ 0.18	99.93 $\pm$ 0.04	99.94 $\pm$ 0.06	99.98 $\pm$ 0.01
		AUROC.	91.58	89.43	99.75	99.81 $\pm$ 0.16	99.93 $\pm$ 0.03	99.87 $\pm$ 0.05	99.95 $\pm$ 0.02
		DTACC.	88.96	87.19	99.28	98.89 $\pm$ 0.39	99.41 $\pm$ 0.26	99.23 $\pm$ 0.19	99.58 $\pm$ 0.13
		AUPRin.	94.19	92.12	99.64	99.84 $\pm$ 0.25	99.97 $\pm$ 0.02	99.92 $\pm$ 0.07	99.98 $\pm$ 0.01
		AUPRout.	83.95	85.47	99.05	99.27 $\pm$ 0.33	99.22 $\pm$ 0.33	98.93 $\pm$ 0.43	99.38 $\pm$ 0.37

In the case of (In) CIFAR-10, the proposed method and the Mahalanobis-based approach detected OOD samples well. Meanwhile, the other models did not perform well, as shown in Table 3. Furthermore, our proposed methods (ensemble-Siamese, ensemble-triplet) showed the best performances for all combinations of datasets. Specifically, their TNR values were approximately 0.01–0.61% higher than the state-of-the-art approach's. In the case of experiments on (In) CIFAR-100, similarly, the proposed method and the Mahalanobis-based approach outperformed other methods, as shown in Table 4. This table also presents that the proposed methods (ensemble-Siamese, ensemble-triplet) were superior to the others, and showed the best TNR, approximately 1.80–7.17% higher than the state-of-the-art approach. However, our previous methods (OnlySiamese, OnlyTriplet) showed similar or worse performance to the ODIN method in both cases of (In) CIFAR-10 and (In) CIFAR-100. It was considered that these poor results came about because CIFAR-10 and CIFAR-100 have more complex structures than SVHN.

Table 3. Average performances at the epoch point, showing the best TNR in 10 experiments on (In) CIFAR-10/(Out) SVHN, Tiny ImageNet and LSUN (mean $\pm$ standard deviation).

In-Dist.	Out-Dist.	Metric	Baseline	ODIN	Mahal.	OnlySiamese (Our Previous)	OnlyTriplet (Our Previous)	Ensemble-Siamese (Ours)	Ensemble-Triplet (Ours)
CIFAR-10	SVHN	TNR at TPR 95%.	32.47	86.60	96.93	50.69 $\pm$ 5.69	51.77 $\pm$ 7.84	96.94 $\pm$ 0.18	97.01 $\pm$ 0.26
		AUROC.	89.88	96.65	99.23	90.75 $\pm$ 1.29	90.39 $\pm$ 2.01	99.21 $\pm$ 0.05	99.22 $\pm$ 0.04
		DTACC.	85.06	91.09	95.99	84.41 $\pm$ 1.58	84.24 $\pm$ 2.30	96.04 $\pm$ 0.11	96.06 $\pm$ 0.17
		AUPRin.	85.40	92.53	98.44	81.68 $\pm$ 3.18	79.35 $\pm$ 6.11	98.23 $\pm$ 0.15	98.29 $\pm$ 0.24
		AUPRout.	93.96	98.52	99.65	95.01 $\pm$ 0.54	95.01 $\pm$ 0.95	99.65 $\pm$ 0.04	99.65 $\pm$ 0.03
CIFAR-10	Tiny ImageNet	TNR at TPR 95%.	44.72	72.51	97.10	72.86 $\pm$ 3.98	78.92 $\pm$ 4.18	97.69 $\pm$ 0.11	97.51 $\pm$ 0.10
		AUROC.	91.02	94.04	99.47	91.82 $\pm$ 1.50	93.90 $\pm$ 1.22	99.51 $\pm$ 0.04	99.48 $\pm$ 0.02
		DTACC.	85.05	86.48	96.32	86.08 $\pm$ 1.51	88.28 $\pm$ 1.48	96.60 $\pm$ 0.09	96.52 $\pm$ 0.09
		AUPRin.	92.49	94.21	99.48	87.99 $\pm$ 2.37	90.71 $\pm$ 1.78	99.47 $\pm$ 0.07	99.41 $\pm$ 0.12
		AUPRout.	88.40	94.09	99.48	92.82 $\pm$ 1.32	94.67 $\pm$ 1.13	99.51 $\pm$ 0.04	99.44 $\pm$ 0.08
CIFAR-10	LSUN	TNR at TPR 95%.	45.44	73.83	98.57	80.09 $\pm$ 4.99	82.73 $\pm$ 3.93	99.01 $\pm$ 0.07	98.94 $\pm$ 0.07
		AUROC.	91.04	94.14	99.70	94.32 $\pm$ 1.42	95.40 $\pm$ 0.91	99.65 $\pm$ 0.02	99.67 $\pm$ 0.03
		DTACC.	85.26	86.69	97.41	88.95 $\pm$ 1.93	90.08 $\pm$ 1.32	97.89 $\pm$ 0.10	97.78 $\pm$ 0.11
		AUPRin.	92.45	94.21	99.70	91.97 $\pm$ 1.75	93.52 $\pm$ 1.25	99.18 $\pm$ 0.13	99.33 $\pm$ 0.18
		AUPRout.	88.55	94.34	99.70	94.84 $\pm$ 1.33	95.74 $\pm$ 0.95	99.69 $\pm$ 0.04	99.71 $\pm$ 0.02

Table 4. Average performances at the epoch point, showing the best TNR in 10 experiments on (In) CIFAR-100/(Out) SVHN, Tiny ImageNet and LSUN (mean $\pm$ standard deviation).

In-Dist.	Out-Dist.	Metric	Baseline	ODIN	Mahal.	OnlySiamese (Our Previous)	OnlyTriplet (Our Previous)	Ensemble-Siamese (Ours)	Ensemble-Triplet (Ours)
CIFAR-100	SVHN	TNR at TPR 95%.	20.25	62.76	91.94	45.91 $\pm$ 5.15	48.11 $\pm$ 7.16	93.60 $\pm$ 0.94	93.80 $\pm$ 0.54
		AUROC.	79.45	93.94	98.36	88.43 $\pm$ 1.48	88.91 $\pm$ 2.13	98.47 $\pm$ 0.10	98.46 $\pm$ 0.10
		DTACC.	73.20	88.04	93.66	82.12 $\pm$ 1.60	82.22 $\pm$ 2.12	94.44 $\pm$ 0.40	94.54 $\pm$ 0.20
		AUPRin.	64.83	88.97	96.41	74.74 $\pm$ 3.90	76.63 $\pm$ 6.01	96.57 $\pm$ 0.38	96.63 $\pm$ 0.38
		AUPRout.	89.02	96.91	99.34	94.19 $\pm$ 0.79	94.61 $\pm$ 0.89	99.24 $\pm$ 0.09	99.21 $\pm$ 0.06
CIFAR-100	Tiny ImageNet	TNR at TPR 95%.	20.40	49.19	90.12	63.45 $\pm$ 8.59	65.20 $\pm$ 6.86	92.24 $\pm$ 0.30	92.22 $\pm$ 0.35
		AUROC.	77.17	87.62	98.06	88.32 $\pm$ 2.62	88.98 $\pm$ 2.12	98.43 $\pm$ 0.08	98.39 $\pm$ 0.10
		DTACC.	70.82	80.11	93.02	82.86 $\pm$ 2.60	82.99 $\pm$ 2.16	93.79 $\pm$ 0.17	93.77 $\pm$ 0.14
		AUPRin.	79.74	87.06	98.11	82.67 $\pm$ 3.14	84.41 $\pm$ 2.80	98.39 $\pm$ 0.15	98.21 $\pm$ 0.29
		AUPRout.	73.30	87.39	98.02	89.89 $\pm$ 2.63	90.43 $\pm$ 2.17	98.42 $\pm$ 0.07	98.42 $\pm$ 0.08
CIFAR-100	LSUN	TNR at TPR 95%.	18.78	45.59	90.71	75.23 $\pm$ 8.88	72.73 $\pm$ 7.13	97.21 $\pm$ 0.39	97.08 $\pm$ 0.28
		AUROC.	75.75	85.64	98.13	92.36 $\pm$ 2.53	92.10 $\pm$ 2.16	99.20 $\pm$ 0.09	99.17 $\pm$ 0.06
		DTACC.	69.89	78.26	93.47	87.06 $\pm$ 3.03	86.40 $\pm$ 2.55	96.26 $\pm$ 0.23	96.24 $\pm$ 0.24
		AUPRin.	77.62	84.49	98.34	89.58 $\pm$ 2.90	89.49 $\pm$ 2.46	98.89 $\pm$ 0.11	98.81 $\pm$ 0.12
		AUPRout.	71.98	85.70	97.74	93.13 $\pm$ 2.61	92.64 $\pm$ 2.24	99.22 $\pm$ 0.10	99.20 $\pm$ 0.10

In summary, our previous method trained using only a DML-based network outperformed other models, including the state-of-the-art approaches trained using only a softmax-based classifier, on simple datasets (such as 1-channel image datasets). However, OOD could not be detected well in complex image datasets (such as 3-channel datasets). In contrast, the proposed method trained by the ensemble of the two networks outperformed all other methods for the combinations of simple image datasets and combinations of complex datasets, showing up to 7% higher TNR than the state-of-the-art approach.

5. Conclusions

This study proposed a novel OOD detection method that can train more efficient feature space. The proposed method uses an ensemble of the features trained using the softmax-based classifier and the DML-based network. With a complementary interaction between these two networks, the trained feature space has a more clumped distribution, and it can be better fitted to the Gaussian distribution by class. Thus, OOD samples can be efficiently detected by setting a threshold in this feature space. To verify the proposed method, we applied our OOD detection approach to various combinations of standard datasets which have been most actively used for evaluating and comparing OOD detection methods. After that, we compared its performance with previous approaches. The results showed that the overall performance of the proposed approach made it superior to other methods, including the state-of-the-art approach trained using only a softmax-based classifier and our previous method trained using only a DML-based network. We believe that the proposed approach has the potential to be applied to designing various machine learning models that can be efficiently used in the real-world, where data distributions undergo complex changes.

Author Contributions: Conceptualization, D.Y.; methodology, D.Y., I.S. and K.-H.L.; software, D.Y., K.M.N. and I.S.; validation, D.Y., K.M.N., K.-H.L. and M.H.; investigation, D.Y.; resources, M.H.; data curation, D.Y.; writing—original draft preparation, D.Y.; writing—review and editing, D.Y., I.S., K.M.N., K.-H.L. and M.H.; visualization, D.Y. and I.S.; supervision, M.H.; project administration, M.H.; funding acquisition, M.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Korea Institute of Science and Technology Information (KISTI).

Acknowledgments: This research was supported by Korea Institute of Science and Technology Information (KISTI).

Conflicts of Interest: The authors declare no conflict of interest.

References

- Hannun, A.; Case, C.; Casper, J.; Catanzaro, B.; Diamos, G.; Elsen, E.; Prenger, R.; Satheesh, S.; Sengupta, S.; Coates, A.; et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv* **2014**, arXiv:1412.5567.
- Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
- He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
- Krizhevsky, A.; Hinton, G. Learning Multiple Layers of Features from Tiny Images. Master's Thesis, University of Toronto, Toronto, ON, Canada, 2009.
- Villegas, R.; Yang, J.; Zou, Y.; Sohn, S.; Lin, X.; Lee, H. Learning to generate long-term future via hierarchical prediction. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 3560–3569.
- Tulyakov, S.; Liu, M.Y.; Yang, X.; Kautz, J. Mocogan: Decomposing motion and content for video generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1526–1535.
- Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015; pp. 1721–1730.
- Wu, N.; Phang, J.; Park, J.; Shen, Y.; Huang, Z.; Zorin, M.; Jastrzebski, S.; Févry, T.; Katsnelson, J.; Kim, E.; et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Trans. Med. Imaging* **2019**, *39*, 1184–1194. [[CrossRef](#)] [[PubMed](#)]
- Hsu, Y.C.; Shen, Y.; Jin, H.; 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, Seattle, WA, USA, 13–19 June 2020; pp. 10951–10960.
- Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete problems in AI safety. *arXiv* **2016**, arXiv:1606.06565.
- Nguyen, A.; Yosinski, J.; 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, Boston, MA, USA, 7–12 June 2015; pp. 427–436.
- Hendrycks, D.; Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv* **2016**, arXiv:1610.02136.
- Davis, J.; Goadrich, M. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 233–240.
- Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
- Liang, S.; Li, Y.; Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv* **2017**, arXiv:1706.02690.
- Lee, K.; Lee, H.; Lee, K.; Shin, J. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv* **2017**, arXiv:1711.09325.
- Papadopoulos, A.A.; Rajati, M.R.; Shaikh, N.; Wang, J. Outlier exposure with confidence control for out-of-distribution detection. *arXiv* **2019**, arXiv:1906.03509.
- Hendrycks, D.; Mazeika, M.; Dietterich, T. Deep anomaly detection with outlier exposure. *arXiv* **2018**, arXiv:1812.04606.
- Lee, K.; Lee, K.; Lee, H.; Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 7167–7177.
- Yang, D.; Shin, I.; Ngoc, K.M.; Kim, H.; Yu, C.; Hwang, M. Out-of-Distribution Detection Based on Distance Metric Learning. In Proceedings of the 9th International Conference on Smart Media and Applications (SMA 2020), Jeju Island, Korea, 17–19 September 2020.
- Koch, G.; Zemel, R.; Salakhutdinov, R. Siamese Neural Networks for One-Shot Image Recognition. Master's Thesis, University of Toronto, Toronto, ON, Canada, 2015. Available online: <http://www.cs.toronto.edu/~gkoch/files/msc-thesis.pdf> (accessed on 27 February 2021).
- Hoffer, E.; Ailon, N. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 84–92.
- Suárez, J.L.; García, S.; Herrera, F. A Tutorial on Distance Metric Learning: Mathematical Foundations, Algorithms, Experimental Analysis, Prospects and Challenges. *Neurocomputing* **2021**, *425*, 300–322. [[CrossRef](#)]
- Xiao, H.; Rasul, K.; Vollgraf, R. Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms. *arXiv* **2017**, arXiv:1708.07747.
- LeCun, Y. The MNIST Database of Handwritten Digits. 1998. Available online: <http://yann.lecun.com/exdb/mnist/> (accessed on 27 February 2021).
- Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; Ng, A. Reading Digits in Natural Images with Unsupervised Feature Learning. *NIPS* **2011**. Available online: <http://ufldl.stanford.edu/housenumbers/> (accessed on 27 February 2021).

-
27. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.-F. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
 28. Yu, F.; Seff, A.; Zhang, Y.; Song, S.; Funkhouser, T.; Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv* **2015**, arXiv:1506.03365.
 29. Loshchilov, I.; Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv* **2016**, arXiv:1608.03983.
 30. Prechelt, L. Early stopping-but when? In *Neural Networks: Tricks of the Trade*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 55–69.