AUAAC: Area Under Accuracy-Accuracy Curve for Evaluating Out-of-Distribution Detection

Wonjik Kim[1][0000−0003−1067−8096], Masayuki Tanaka[1,2][0000−0002−5756−1904], and Masatoshi Okutomi[2][0000−0001−5787−0742]

[1] Artificial Intelligence Research Center
National Institute of Advanced Industrial Science and Technology
Umezono, Tsukuba 305-8560, Japan
kim-wonjik@aist.go.jp
[2] Department of Systems and Control Engineering, School of Engineering
Tokyo Institute of Technology
Meguro-ku, Tokyo 152-8550, Japan
{mtanaka,mxo}@sc.e.titech.ac.jp

Abstract. Determining whether input data are out-of-distribution (OOD) is important for real-world applications of machine learning. Various approaches to OOD detection have been proposed, and there is a growing interest in evaluating their performance. A commonly employed approach for OOD detection is training the network model using an in-distribution (IND) task and then applying a threshold to the probability estimated of unknown data. However, current evaluation metrics only assess the OOD detection performance while neglecting the IND task performance. To address this issue, we propose new evaluation metrics for OOD detection. Our novel metric, the area under the accuracy-accuracy curve (AUAAC), is designed to simultaneously evaluate both the IND task and OOD detection performances. Specifically, it calculates the area under the accuracy-accuracy curve after estimating the accuracy of the IND task and OOD detection for all thresholds. Flaws within the training dataset, such as contaminated labels or inaccurate annotations, disturb the network in properly performing the IND task. Nevertheless, the network may distinguish whether new input data are in IND because it was priorly exposed to IND data and trained by their features. The proposed AUAAC can asses such malfunction while existing evaluation metrics overlook the performance of the IND task and cannot identify such issues.

Keywords: Deep learning · Out-of-distribution detection · Evaluation metric.

1 Introduction

Learning-based deep neural network methods have demonstrated high performance with controlled data in various tasks. Those methods implicitly assume that training and test data distributions are the same. However, the test data distribution is often different from that of training data. That mismatch significantly

degrades the performance. Then, in practice, it is important to detect unexpected data before using it. This detection task is known as the out-of-distribution (OOD) detection task [5]. The main task is called as in-distribution (IND) task. In this paper, we focus on the classification task for the IND task.

Various OOD detection algorithms have been proposed after the OOD detection task was well formulated [5]. One of these approaches is to feed OOD data to the network during the IND task training phase, so that the network can learn differences between IND and OOD data [13, 2, 6, 25, 16, 8]. However, those approaches require IND and OOD data to train the network model. Then, OOD detection algorithms only require IND datasets have been proposed [14, 15, 19, 1, 3, 20]. The OOD detection performance of the network model is usually evaluated after training the network model for the IND task, implicitly assuming that the IND task performance of the network model is good enough. The OOD detection performance of the network model is usually evaluated after training the network model for the IND task, implicitly assuming that the IND task performance of the network model is good enough. However, we can easily improve the OOD detection performance if we allow the degradation of the IND task performance. For this reason, it is very important to evaluate simultaneously the IND task performance and the OOD detection performance.

The true negative rate (TNR) at 95% true positive rate (TPR) is a widely used metric to evaluate the OOD detection performance. The TPR is defined by TP/(TP+FN) and the TNR is defined by TN/(TN+FP), where TP, TN, FP, and FN are the numbers of true positive, true negative, false positive, and false negative, respectively. IND and OOD data are considered as positive and negative samples. The metric of the TNR at 95% TPR indicates the probability of correctly detecting OOD data when the network model can detect the IND data with 95% accuracy. Scherreik *et al.* [18] used maximum Youden's index [23] as a metric to evaluate the OOD detection performance. Youden's index is defined by the sum of TPR and TNR and is usually used to determine the proper parameter based on Receiver Operatorating Characteristic (ROC) curve. The area under the receiver operating characteristic curve (AUROC) has become a popular metric, as it summarizes TPR performance for all possible thresholds and is insensitive to threshold selection. The ROC curve for calculating AUROC depicts the relationship between true and false positive rates. An open-set F-score [10] is recently proposed for a metric to evaluate the OOD detection performance. Note that those three metrics can evaluate the OOD performance, while they commonly neglect the classification accuracy of the IND data.

However, in practice, OOD detection is not the final goal of the classification task. In this sense, we need to evaluate both IND classification accuracy and OOD detection performance. For example, if the network model has high IND classification accuracy with low OOD detection performance, that network model cannot work well in the practical situation because the input data includes both IND and OOD data. To address this issue, a concise pairwise formulation called OpenAUC [21] has been proposed for simultaneously evaluating both the IND task and OOD detection performance. OpenAUC is a concise pairwise

formulation where each pair consists of IND and OOD data. First, the metric checks whether the IND data are correctly classified into known IND classes, then confirms whether the OOD data are detected. This approach allows for a more comprehensive evaluation of OOD detection methods.

We introduce a new metric called the area under the accuracy-accuracy curve (AUAAC) which is designed to simultaneously evaluate both the IND task and OOD detection performance. We first estimate the accuracy of the IND task and OOD detection for all thresholds. Next, we draw the accuracy-accuracy curve (AAC), with the OOD detection accuracy on the horizontal axis and the IND task accuracy on the vertical axis. Finally, we calculate the area under the AAC (AUAAC) to comprehensively evaluate OOD detection performance.

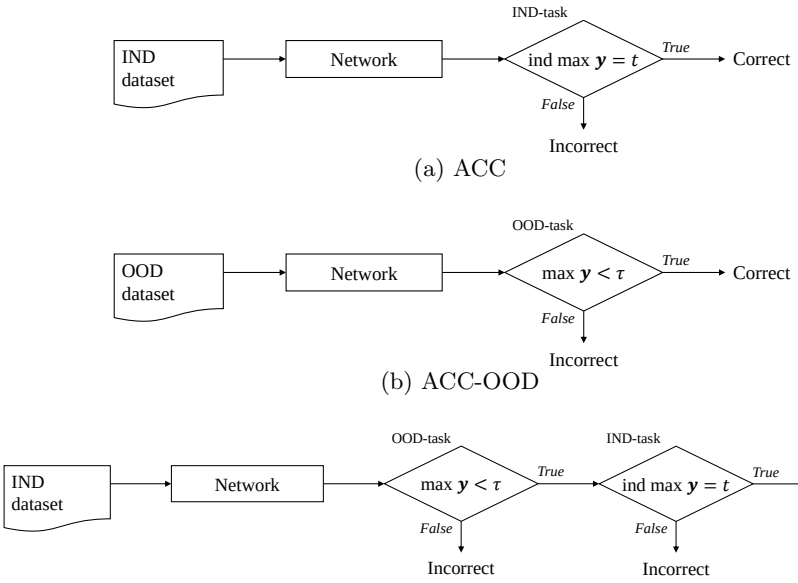
In order to validate the proposed metrics, we conducted experiments on two different network models using three IND datasets and five OOD datasets. We evaluated the OOD detection performance using three metrics: TNR at TPR 95%, AUROC, and AUAAC. Our analysis confirmed the importance of incorporating AUAAC to account for IND task performance. The dataset used for the IND task can potentially contain flaws, such as contaminated labels or incorrect annotations. When trained with such flawed data, the network may fail to perform correctly on the IND task. However, the network might still demonstrate an ability to detect OOD data. This is because the network previously encountered IND data and was trained by their features, enabling it to recognize characteristics that observe data as different from the IND. In other words, the network exposed to IND data allows it to distinguish whether the brand new data are in IND. Since existing evaluation metrics do not consider the performance of the IND task, they cannot catch out such malfunctioning networks. This highlights the importance of considering IND task performance even in OOD detection, and AUAAC can operate as an evaluation metric for this purpose. Additionally, we investigated the behavior of AUAAC during the training phase to gain further insights into its characteristics. Based on the experimental results, we concluded that the proposed metrics serve as valuable tools for evaluating network models in terms of both OOD detection and IND task performance.

2 Proposed Metric

Let $\mathbf{f} : \mathcal{X} \to \mathbb{R}^K$ be a network model which infers the class from an input image, where $\mathcal{X}$ is the input image space and K is the number of classes. The network model $\boldsymbol{f}$ has parameters to be trained with the training data.

We consider the problem of distinguishing IND and OOD data with a trained network model. For an input image $\boldsymbol{x}$, the network model estimates a probability $\boldsymbol{f}(\boldsymbol{x})$, and compare with the OOD threshold τ. An input image $\boldsymbol{x}$ is classified as IND if the maximum probability is greater than the threshold τ as follows:

$$\boldsymbol{x} \text{ is classified as } \begin{cases} \text{IND} & (\max \boldsymbol{f}(\boldsymbol{x}) \geq \tau) \\ \text{OOD} & (else) \end{cases} . \tag{1}$$



(a) ACC

(b) ACC-OOD

(c) ACC-IND

Fig. 1: Calculation flowchart of three metrics.

2.1 Three Accuracy Metrics

First, we review the performance evaluations of the IND classification and the OOD detection tasks. Here, we consider three accuracies related to the IND classification and the OOD detection tasks; ACC, ACC-IND, and ACC-OOD. ACC represents the classical accuracy focusing on the IND classification task only. ACC-IND is the accuracy of the IND classification task considering the OOD detection, which is newly introduced in this paper. ACC-OOD is the accuracy of the OOD detection task. The accuracy can be generally defined by

$$\text{Accuracy} = E_{(\boldsymbol{x},t)\sim\Omega}[g(\boldsymbol{f}(\boldsymbol{x}), t)]\,, \tag{2}$$

where Ω represents a dataset of the images and the labels, $\boldsymbol{x}$ represents the image, t is associated ground truth label index, and the function g is refer to the correctness function which specifies the property of the accuracy. Here, we consider ground truth label index has a positive integer value for the IND sample and a negative value for the OOD sample. We will show three different correctness functions with the dataset are associated to three accuracies of ACC, ACC-IND, and ACC-OOD.

For ACC which is the accuracy of the classical IND classification task, the correctness function g_{ACC} can be defined by

$$g_{\text{ACC}}(\boldsymbol{y}, t) = \begin{cases} 1 & (\text{ind}\max \boldsymbol{y} = t) \\ 0 & (else) \end{cases}\,, \tag{3}$$

where $\mathrm{ind}\max \boldsymbol{y}$ represents the index of maximum element of $\boldsymbol{y}$. The correctness function g_{ACC} returns one if the IND classification inference by the network model is correct. Then, ACC of the classical IND classification accuracy can be evaluated by Eq. 2 with the correctness function g_{ACC} in Eq. 3 and the IND dataset of Ω_{IND}. The flowchart for calculating ACC can be illustrated in Fig. 1a.

For ACC-OOD which is the OOD detection task, the function g_{OOD} can be defined by

$$g_{\mathrm{OOD}}(\boldsymbol{y}, t; \tau) = \begin{cases} 1 & (t \times (\max \boldsymbol{y} - \tau) > 0) \\ 0 & (else) \end{cases} . \tag{4}$$

If the maximum value of inference $\boldsymbol{y}$ is greater than the threshold τ, the associated sample is estimated as an IND sample. We set the positive integer value for the label of the IND sample and the negative value for the label of the OOD sample. Then, the function g_{OOD} in Eq. 4 returns one if the OOD detection inference is correct. ACC-OOD of the OOD detection accuracy can be evaluated by Eq. 2 with the function g_{OOD} in Eq. 4 and the OOD dataset of Ω_{OOD}, as shown in Fig. 1b. Based on the equations, it is obvious that the calculation of ACC-OOD solely relies on the OOD dataset and does not take into account the IND data.

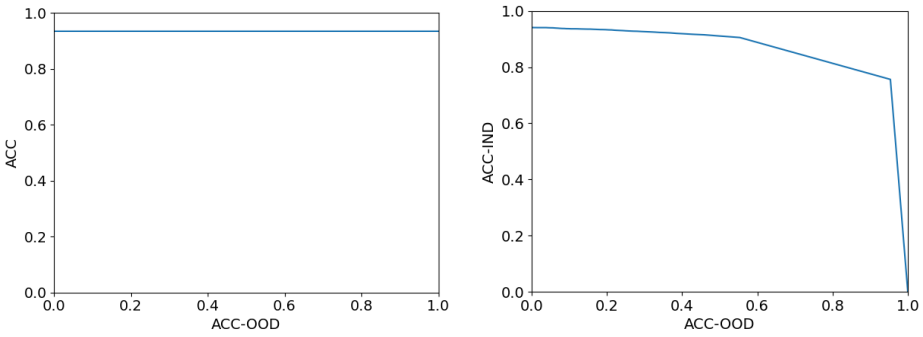
For the practical classification task, we need to evaluate the classification accuracy considering the OOD detection. Here, we introduce the new accuracy of ACC-IND (Fig. 1c), which can evaluate the IND classification accuracy after the OOD detection. If both the IND classification inference and the OOD detection inference are correct, we consider it to be correct for the ACC-IND. The associated correctness function of g_{IND} can be defined by

$$g_{\mathrm{IND}}(\boldsymbol{y}, t; \tau) = g_{\mathrm{ACC}}(\boldsymbol{y}, t) \times g_{\mathrm{OOD}}(\boldsymbol{y}, t; \tau) . \tag{5}$$

For the IND dataset, the function g_{OOD} should be one for the correct inference. In addition, the IND classification inference should also be correct, or the function g_{ACC} should be one for the correct inference. The function g_{IND} returns ones if the inference is correct. Then, ACC-IND of the IND classification accuracy considering the OOD detection can be evaluated by Eq. 2 with the function g_{IND} in Eq. 5 and the IND dataset of Ω_{IND}. Note that ACC-IND is evaluated based on the IND dataset Ω_{IND}.

2.2 AUAAC: Area Under Accuracy-Accuracy Curve

We propose a metric to evaluate both the IND classification accuracy and the OOD detection accuracy, which we call the area under the accuracy-accuracy curve (AUAAC). The AUAAC is defined based on ACC-IND and ACC-OOD, which were introduced in the previous section. As mentioned above, ACC-IND and ACC-OOD have the parameter of threshold τ. Then, we can draw the accuracy-accuracy curve whose horizontal and vertical axes are ACC-OOD and ACC-IND, changing the threshold parameter τ. Fig. 2b shows an example of the accuracy-accuracy curve. Following the idea of area under the ROC curve, we propose a metric of the AUAAC as the area under the accuracy-accuracy. The



(a) ACC / ACC-OOD graph (b) ACC-IND / ACC-OOD graph

Fig. 2: Example of the accuracy-accuracy curve.

maximum value of the AUAAC is one. It is worth noting that employing ACC instead of the proposed ACC-IND on the Y-axis, as depicted in Fig. 2a, restricts the examination of IND task performance through threshold value changes.

Recall that ACC-IND is evaluated only with the IND dataset and that ACC-OOD is evaluated only with the OOD dataset. Assuming the network model infers posterior probabilities, when the thresholding parameter τ is 0, ACC-OOD is zero because all data of the OOD dataset is wrongly classified as IND data. In addition, ACC-OOD is one for $\tau = 1$ because all data of the OOD dataset is correctly classified as OOD data. When the thresholding parameter τ increases, the ACC-OOD also increases. ACC-IND of $\tau = 0$ is identical to the classical ACC. In other words, ACC-IND at zero ACC-OOD equals the classical ACC. By definition of Eq. 5, when the thresholding parameter τ increases, the ACC-IND decreases. Then, the accuracy-accuracy curve is non-increase property.

Let's consider the real-world application. In the real-world application, input data includes the IND and the OOD data. Assuming we have a trained network model, we need to determine the suitable threshold parameter τ, because here is a trade-off relationship between ACC-IND and ACC-OOD as shown in Fig. 2b.

Recent deep neural networks often overfit the training data, which leads to improved performance on the IND task during continued training. However, the network model can become overconfident, resulting in high confidence estimates even for OOD data and a decrease in OOD detection accuracy. Deciding when to stop training becomes a complex multi-objective optimization problem when considering both IND task performance and OOD detection accuracy. The proposed AUAAC metric evaluates both performance metrics simultaneously and can be used as a simple criterion for stopping network training.

3 Experiments

First, we evaluated the OOD detection performance of two network models, ResNet-34 and DenseNet-VC-100, trained with three different IND datasets. We

employed three evaluation metrics: TNR at TPR 95%, AUROC, and proposed AUAAC, to evaluate the OOD detection performance. Subsequently, we examined the behavior of ACC, ACC-IND, ACC-OOD, and proposed AUAAC of ResNet-34 during the training phase to gain insights into the characteristics of the proposed AUAAC.

3.1 Experimental setup

We conducted comparative experiments under the following conditions to confirm the behavior of the proposed AUAAC metric compared to other evaluation metrics for OOD detection. For the IND datasets, we used the standard split of CIFAR-100 [11], CIFAR-10 [11], and SVHN [17]. These datasets contain RGB images with 32×32 pixels. For the OOD test dataset, we used iSUN [22], LSUN [24], and TinyImageNet[12]. As same as in [14], we used two variants of TinyImageNet, and LSUN sets: '(C)' stands for using a 32×32 image crop, and '(R)' stands for using resized images to 32×32 pixels. We also used CIFAR-100, CIFAR-10, and SVHN as OOD if a model was not trained with them. We employed ResNet-34 [4] and DenseNet-BC-100 [7] to train the image classification models. The networks were trained by categorical cross-entropy loss function using Adam [9] with a 0.001 learning rate in 400 epochs, and the batch size is 128.

Additionally, we investigated the behavior of ACC, ACC-IND, ACC-OOD, and AUAAC of ResNet-34 during the training phase. In calculating ACC-IND and ACC-OOD, a threshold was set so that 95% of IND data were correctly determined as IND. For a dataset that represents known classes, we used the standard split of CIFAR-10 [11]. For the OOD performance evaluation, we use TinyImageNet[12] with 32×32 image crop. The network was trained by categorical cross-entropy loss function using Adam [9] with a 0.001 learning rate in 1000 epochs. The batch size is 128.

3.2 Results

Table 1 presents the results of OOD detection using DenseNet-BC-100. TNR at TPR 95% shows a similar tendency as AUROC since both metrics only evaluate OOD detection accuracy. since AUAAC does not directly evaluate OOD detection performance, it may exhibit less sensitivity toward OOD detection compared to other metrics. This trend is evident in Tables 1 and 2, where the values of AUAAC show a similar trend to AUROC but exhibit smaller changes compared to TNR at TPR 95% and AUROC.

By incorporating AUAAC alongside other metrics, we can better understand the network's performance. For instance, AUROC helps determine the network's ability to effectively distinguish between IND and OOD data. On the other hand, AUAAC allows us to assess not only OOD detection performance but also the network's performance in the IND task. While a high AUROC suggests good OOD detection capability, it does not guarantee strong performance in the IND task. Conversely, a high AUAAC indicates potential competence in both the IND task and OOD detection. Table 3 presents the OOD detection

Table 1: OOD detection performance of DenseNet-BC-100 in three metrics. Results reported in percentage.

IND	OOD	Metrics		
		TNR at TPR 95%	AUROC	AUAAC
CIFAR-100	iSUN	3.3	37.2	37.8
	LSUN (C)	13.8	70.7	54.8
	LSUN (R)	2.8	34.9	37.8
	ImageNet (C)	6.9	50.6	44.4
	ImageNet (R)	4.0	39.5	39.6
	SVHN	11.2	70.0	54.1
CIFAR-10	iSUN	31.0	87.4	81.9
	LSUN (C)	42.2	91.3	84.7
	LSUN (R)	32.5	88.5	83.0
	ImageNet (C)	30.1	87.7	82.2
	ImageNet (R)	28.6	86.7	81.7
	SVHN	46.2	92.8	85.0
SVHN	iSUN	60.4	93.7	89.2
	LSUN (C)	55.4	91.3	86.1
	LSUN (R)	57.5	92.9	88.5
	ImageNet (C)	60.6	93.6	88.8
	ImageNet (R)	61.7	93.6	89.3
	CIFAR-10	53.6	91.6	86.1
	CIFAR-100	52.8	91.4	86.1

results for ResNet34, which was trained on Cifar10 as an IND dataset but had a Top-1 accuracy of 6.2% due to incorrect class labels. The results indicate that even when trained with the wrong labels, the metrics TNR at TPR 95% and AUROC, which only evaluate OOD detection performance, exhibit acceptable values. However, it is meaningful that the AUAAC, which also assesses the performance of the IND task, demonstrates significantly lower values. In other words, the existing evaluation metrics cannot identify if the network does not perform adequately in IND tasks for some reason. It highlights the advantage of utilizing the proposed AUAAC metric, which can simultaneously assess the IND task and OOD detection performance.

Figure 3 shows the change in ACC, ACC-IND, AUAAC, and ACC-OOD for a thousand epochs of network training. Figure 3a shows that the classification accuracy stably improves as the number of epochs increases. Figure 3b shows that the ACC-OOD achieves a peak at 572 epochs and then gradually falls off. The moving average graph more explicitly confirms the trend of accuracy decay. The graphs of AUAAC in Figure 3a and the accuracy of OOD detection in Figure 3b show similar tendencies in peaks and valleys. This is because the ACC-OOD is

Table 2: OOD detection performance of ResNet-34 in three metrics. Results reported in percentage.

IND	OOD	Metrics		
		TNR at TPR 95%	AUROC	AUAAC
CIFAR-100	iSUN	14.0	75.4	58.6
	LSUN (C)	15.5	75.6	58.5
	LSUN (R)	15.7	77.6	59.9
	ImageNet (C)	17.4	78.5	60.3
	ImageNet (R)	14.3	76.4	59.2
	SVHN	12.2	72.9	57.2
CIFAR-10	iSUN	39.2	89.4	84.9
	LSUN (C)	46.7	91.9	86.9
	LSUN (R)	40.2	89.7	85.1
	ImageNet (C)	39.5	88.2	83.8
	ImageNet (R)	34.4	86.5	82.4
	SVHN	38.8	91.4	86.3
SVHN	iSUN	67.8	94.2	90.4
	LSUN (C)	67.4	94.1	90.2
	LSUN (R)	67.0	93.9	90.1
	ImageNet (C)	70.4	94.9	91.0
	ImageNet (R)	70.0	94.6	90.7
	CIFAR-10	62.2	92.6	88.9
	CIFAR-100	61.5	92.3	88.6

also taken into account when calculating AUAAC. However, we can see that the range of variation in AUAAC is smaller than that of OOD detection accuracy.

Table 4 summarizes four metrics of ACC, ACC-IND, ACC-OOD, and AUAAC at several epochs. For example, early stopping is an important technique to avoid overfitting. But, it is not an easy task to determine when we should stop the training. If we want to maximize the classical ACC, Table 1 shows that we need to continue the training until 990 epochs. On the other hand, if we focus on ACC-OOD, which represents the OOD detection accuracy, it might be good to stop at 572 epochs. However, ACC-OOD does not reflect any information about ACC and/or ACC-IND. So, there is no guarantee that the network model has sufficient good accuracy in terms of the IND classification accuracy. The proposed AUAAC can evaluate both the IND classification accuracy and the OOD detection performance. Therefore, we can use AUAAC for the stopping criteria. If we train the network to maximize the AUAAC, the network model is expected to have reasonably good IND classification accuracy and OOD detection accuracy.

We also show the Accuracy-Accuracy Curves (AACs) and the proposed metric of AUAAC at several epochs in Figure 4. As shown in Figure 4, the AAC varies as

Table 3: OOD detection performance of ResNet-34 trained with wrong label (Top-1 accuracy of 6.2%) in three metrics. Results reported in percentage.

IND	OOD	Metrics		
		TNR at TPR 95%	AUROC	AUAAC
CIFAR-10	iSUN	51.2	92.5	4.3
	LSUN (C)	46.9	92.4	4.2
	LSUN (R)	53.7	93.5	4.4
	ImageNet (C)	47.7	92.3	4.1
	ImageNet (R)	48.5	92.4	4.2
	SVHN	40.9	90.7	3.8

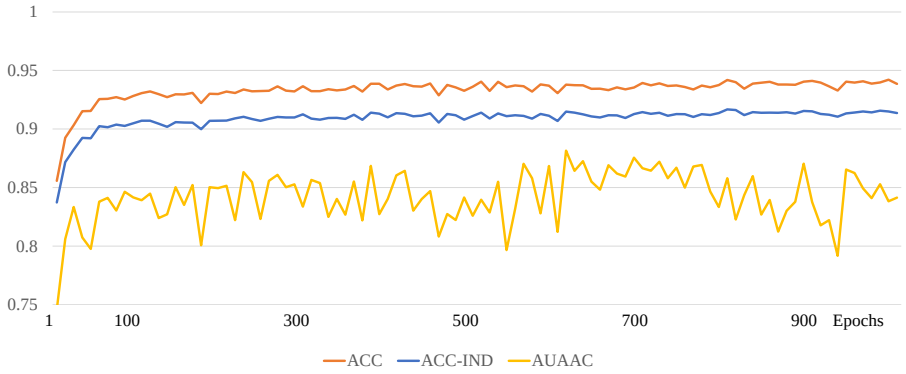
Table 4: ACC, ACC-IND, ACC-OOD, and AUAAC of different training epochs.

Epoch	ACC	ACC-IND	ACC-OOD	AUAAC
572	0.935	0.910	**0.520**	0.879
612	0.939	0.914	0.486	**0.882**
796	0.943	**0.918**	0.450	0.872
884	0.**944**	0.915	0.418	0.851

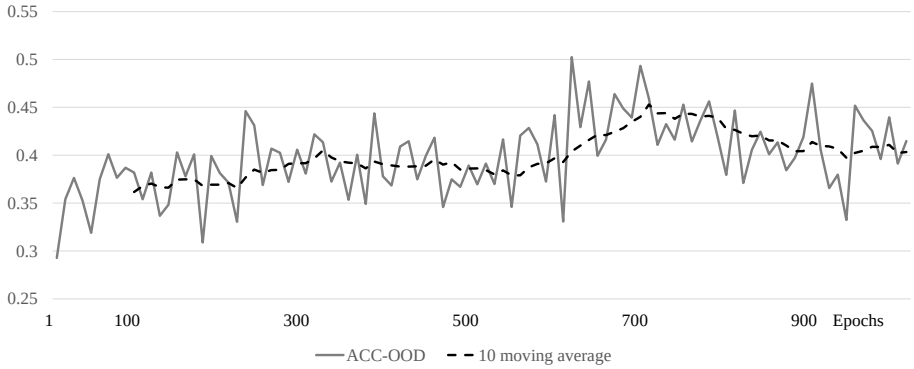
the training progresses and the area changes. Therefore, it is also possible to check at what epoch and at what threshold the desired IND classification accuracy and OOD detection accuracy are achieved. For instance, when comparing the AACs at 700 epochs and 1000 epochs, the AUAAC of the 700 epoch is larger, suggesting superior performance. Nevertheless, if an OOD detection rate of 0.8 is sufficient, then training for 1000 epochs is the preferred choice since it yields greater accuracy for IND classification at the OOD detection rate of 0.8. On the other hand, if a desired OOD detection rate of 95% is targeted, training for 700 epochs would be more suitable. Thus, the proposed method can facilitate the selection of networks and thresholds based on more intricate decision criteria.

4 Conclusions

We have proposed a new metric named the area under the accuracy-accuracy curve (AUAAC) to simultaneously evaluate IND task and OOD detection accuracy. Initially, we evaluated the IND classification accuracy of the IND task and OOD detection performance changing the threshold. Subsequently, we constructed an accuracy-accuracy curve (AAC), plotting the OOD detection accuracy (ACC-OOD) on the X-axis and the IND task accuracy (ACC-IND) on the Y-axis. Finally, we have determined that the proposed AUAAC is good criteria for early stopping. When both IND task performance and OOD detection accuracy are



(a) ACC, ACC-IND and AUAAC graphs.



(b) ACC-OOD and 10 moving average graphs.

Fig. 3: Training graphs of four-different metrics. The threshold τ is selected that 95% of IND data are determined as IND data.

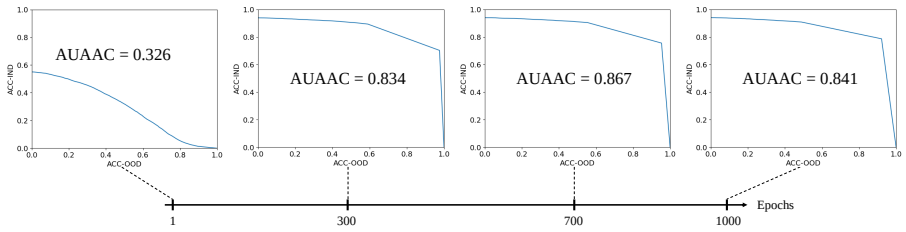


Fig. 4: AUAACs of different training epochs.

considered, determining the optimal point to stop training becomes a complex multi-objective optimization problem. In this context, the proposed AUAAC metric, which evaluates both performance metrics simultaneously, serves as a

straightforward criterion to determine the optimal stopping point for network training. Furthermore, the AAC can provide guidance for selecting a threshold to distinguish between IND and OOD data. By choosing the desired accuracy of IND task or OOD detection, we can readily confirm the accuracy of the other metric.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP23K19985.

References

1. Bibas, K., Feder, M., Hassner, T.: Single layer predictive normalized maximum likelihood for out-of-distribution detection. Advances in Neural Information Processing Systems **34**, 1179–1191 (2021)
2. Dhamija, A.R., Günther, M., Boult, T.: Reducing network agnostophobia. Advances in Neural Information Processing Systems **31** (2018)
3. Dong, X., Guo, J., Li, A., Ting, W.T., Liu, C., Kung, H.: Neural mean discrepancy for efficient out-of-distribution detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19217–19227 (2022)
4. 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. pp. 770–778 (2016)
5. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. Proceedings of International Conference on Learning Representations (2017)
6. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. Proceedings of the International Conference on Learning Representations (2019)
7. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
8. Katz-Samuels, J., Nakhleh, J.B., Nowak, R., Li, Y.: Training ood detectors in their natural habitats. In: International Conference on Machine Learning. pp. 10848–10865. PMLR (2022)
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
10. Kong, S., Ramanan, D.: Opengan: Open-set recognition via open data generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 813–822 (2021)
11. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
12. Le, Y., Yang, X.: Tiny imagenet visual recognition challenge. CS 231N **7**(7), 3 (2015)
13. Lee, K., Lee, H., Lee, K., Shin, J.: Training confidence-calibrated classifiers for detecting out-of-distribution samples. arXiv preprint arXiv:1711.09325 (2017)
14. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint arXiv:1706.02690 (2017)

15. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. Advances in Neural Information Processing Systems **33**, 21464–21475 (2020)
16. Moller, F., Botache, D., Huseljic, D., Heidecker, F., Bieshaar, M., Sick, B.: Out-of-distribution detection and generation using soft brownian offset sampling and autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 46–55 (2021)
17. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011 (2011)
18. Scherreik, M.D., Rigling, B.D.: Open set recognition for automatic target classification with rejection. IEEE Transactions on Aerospace and Electronic Systems **52**(2), 632–642 (2016)
19. Sun, Y., Guo, C., Li, Y.: React: Out-of-distribution detection with rectified activations. Advances in Neural Information Processing Systems **34**, 144–157 (2021)
20. Sun, Y., Li, Y.: Dice: Leveraging sparsification for out-of-distribution detection. In: European Conference on Computer Vision. pp. 691–708. Springer (2022)
21. Wang, Z., Xu, Q., Yang, Z., He, Y., Cao, X., Huang, Q.: Openauc: Towards auc-oriented open-set recognition. Advances in Neural Information Processing Systems **35**, 25033–25045 (2022)
22. Xu, P., Ehinger, K.A., Zhang, Y., Finkelstein, A., Kulkarni, S.R., Xiao, J.: Turkergaze: Crowdsourcing saliency with webcam based eye tracking. arXiv preprint arXiv:1504.06755 (2015)
23. Youden, W.J.: Index for rating diagnostic tests. Cancer **3**(1), 32–35 (1950)
24. 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 preprint arXiv:1506.03365 (2015)
25. Yu, Q., Aizawa, K.: Unsupervised out-of-distribution detection by maximum classifier discrepancy. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9518–9526 (2019)