Suzuki et al., Int J Comput Softw Eng 2021, 6: 165
https://doi.org/10.15344/2456-4451/2021/165

International Journal of
Computer & Software Engineering

Research Article Open Access

Special Issue: Computational Analysis and Modeling

Improving Accuracy of Out-of-Distribution Detection and In-Distribution Classification by Incorporating JSD Consistency Loss

Kaiyu Suzuki and Tomofumi Matsuzawa*

Department of Information Sciences, Tokyo University of Science, Chiba, Japan

Abstract

Out-of-distribution (OOD) detection, the classification of samples not included in the training data, is essential to improve the reliability of deep learning. Recently, the accuracy of OOD detection through unsupervised representation learning is high; however, the accuracy of in-distribution classification (IN-D) is reduced. This is due to the cross entropy, which trains the network to predict shifting transformations (such as angles) for OOD detection. Cross entropy loss conflicts with the consistency in representation learning; that is, samples with different data augmentations applied to the same sample should share the same representation. To avoid this problem, we add the Jensen–Shannon divergence (JSD) consistency loss. To demonstrate its effectiveness for both OOD detection and IN-D classification, we apply it to contrasting shifted instances (CSI) based on the latest representation learning. Our experiments demonstrate that JSD consistency loss outperforms existing methods in both OOD detection and IN-D classification for unlabeled multi-class datasets.

Publication History:

Received: June 07, 2021
Accepted: June 23, 2021
Published: June 25, 2021

Keywords:

Out-of-distribution, Consistency loss, Data augmentation, SimSiam, SimCLRs

Introduction

Deep learning has garnered considerable attention owing to its ability to make accurate decisions based on images and other data and achieve state-of-the-art (SoTA) results in diverse tasks. In particular, for image classification tasks where sufficiently large datasets are available, AlexNet [1], VGG [2], ResNet [3], WideResNet [4], and EfficientNet [5] surpass all other non-deep learning methods. These deep learning classifiers achieved high in-distribution (IN-D) classification accuracy. However, these methods are weak in out-of-distribution (OOD), that is, samples of classes that do not exist in the training data. For example, suppose a classifier that solves the task of classifying ten types of dogs is given an image of a cat. In such cases, the classifier typically assigns a high probability to one of the classes, even though the image does not fit into any of the classes. Even if it does not assign a high probability to any class, it is difficult to determine whether it is an instance of OOD or an unconfident IN-D sample. Data that are not in the training data are often provided to classifiers, especially in real-world tasks. Thus, a method is needed to detect them.

OOD detection, also called novelty detection or anomaly detection, is a method for detecting out-ofdistribution samples that do not exist in the training data. This technique provides a measure to distinguish between the IN-D and OOD samples in the test data. Self-supervised learning is an approach for detecting OOD data in images. Self-supervised approaches are closely related to representation learning, which obtains features from the training samples. This method randomly transforms the training samples and obtains shared labels and feature representations. This learning produces feature representations that are useful for IN-D classification and OOD detection.

A notable previously reported self-supervised approach to detect OOD data is contrasting shifted instances (CSI) [6], based on SimCLR [7]. SimCLR applies two random data augmentations to each sample and learns to match the two obtained feature representations. CSI adds a constraint that images rotated at four different angles (0°, 90°, 180°, and 270°) are given as separate samples, resulting in different feature representations. CSI enhances the property aboveby adding a network–––a rot predictor–––that estimates the rotated angle from the feature representation. This method results in a high accuracy for OOD detection.

However, CSI does not consider the accuracy of classification. It is an issue that CSI reduces the accuracy of classification. The cause of this problem is the learning of the rot predictor, which is more specialized for angle prediction than the original representation learning. In contrast, new representation learning methods demonstrate higher classification accuracy. In particular, BYOL [8] and SimSiam [9] revealed that explicit contrastive learning loss in SimCLR is not necessary for representation learning. In other words, it is not necessary to specify a loss function that separates feature representations obtained from different samples, and the classification accuracy can be improved by simply learning to match two feature representations obtained from the same sample through different data augmentations. It is unclear whether this finding is compatible with OOD detection.

In this study, we developed a method that prevents the rot predictor from degrading the accuracy of classification in representation learning and improves the accuracy of classification by incorporating Jensen–Shannon divergence consistency loss. We also demonstrated how a method without explicit contrastive learning, such as SimSiam, affects OOD detection.

Related Works

Representation learning

Representation learning is a method for obtaining valuable representations for classification and other purposes through self-supervised learning.

***Corresponding Author:** Prof. Tomofumi Matsuzawa, Department of Information Sciences, Tokyo University of Science, Chiba 278-8510, Japan; E-mail: t-matsu@is.noda.tus.ac.jp

Citation: Suzuki K, Matsuzawa T (2021) Improving Accuracy of Out-of-Distribution Detection and In-Distribution Classification by Incorporating JSD Consistency Loss. Int J Comput Softw Eng 6: 165. doi: https://doi.org/10.15344/2456-4451/2021/165

Copyright: © 2021 Suzuki et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

The major representation learning methods, such as SimCLR [7], BYOL [8], and SimSiam [9], share a common approach. First, x_1, x_2 are obtained by applying two data augmentations to a sample x. Second, z_1, z_2, that is, feature representations corresponding to x_1, x_2, are obtained from the neural network. Finally, the neural network learns to make the two obtained feature representations z_1, z_2 consistent. To achieve this policy, each method adopts a different method for the network and loss functions.

SimCLR (Figure 1a) is one of the earliest of these methods and is trained via the following algorithm.

$$x_1, x_2 = \mathrm{aug}(x), \mathrm{aug}(x)$$

$$z_1, z_2 = g_\theta(f_\theta(x_1)), g_\theta(f_\theta(x_2))$$

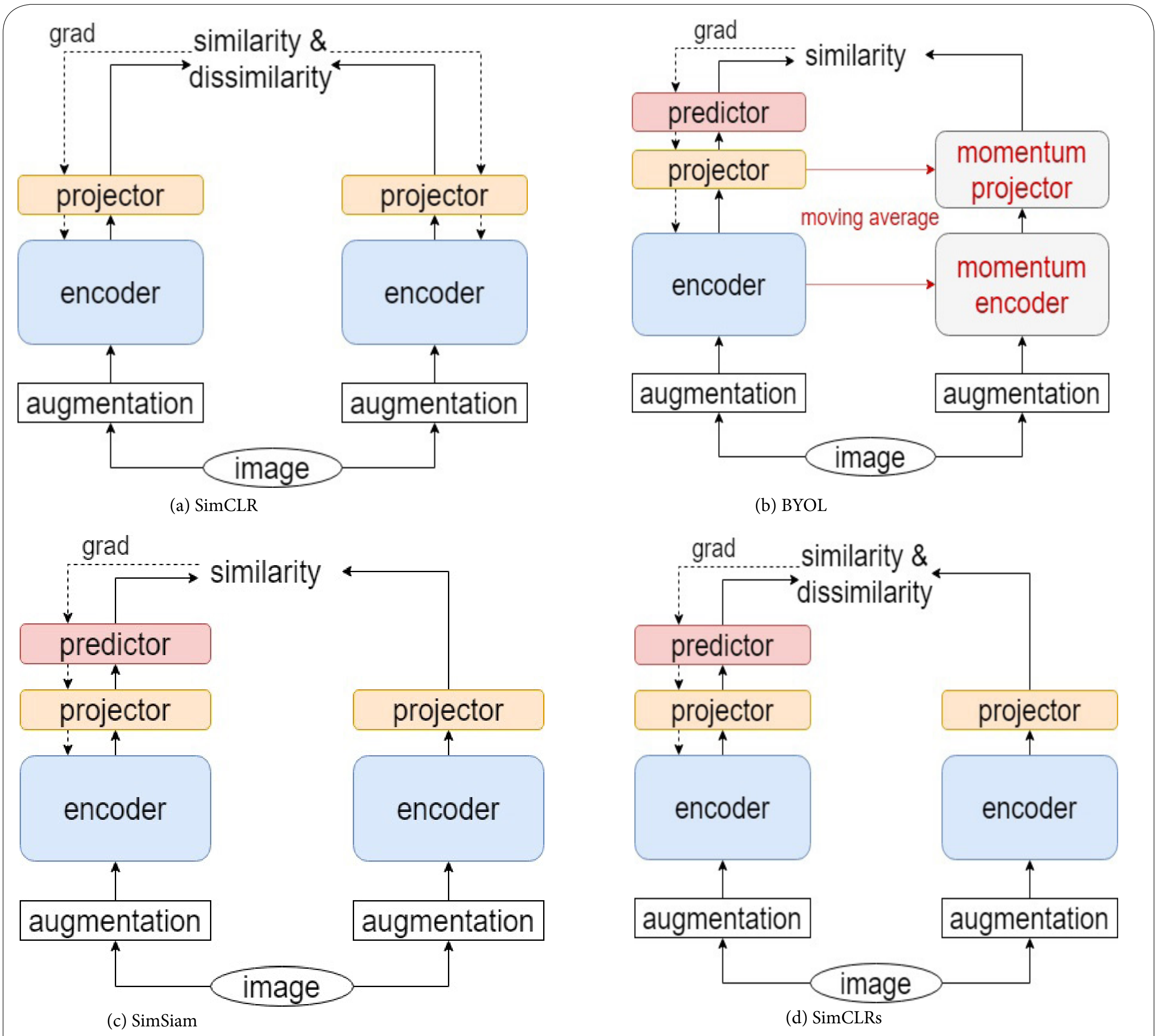
$$loss = D_{simclr}(z_1, z_2) + D_{simclr}(z_1 + z_2)$$

where $D_{simclr}(z_1, z_2)$ is defined in Equation 1.

$$-\sum_{i=1}^{N}\log\frac{\exp(\mathrm{sim}(z_{il},z_{i2}))}{\sum_{j\neq i}^{N}\sum_{k\in[1,2]}\exp(\mathrm{sim}(z_{i1},z_{ik}))}. \qquad (1)$$

Here, aug denotes random data augmentation, and sim denotes the cosine similarity. In addition, f_θ is a neural network such as ResNet [3] without the classification layer, and g_θ is a neural network consisting of one to three layers.

BYOL (Figure 1b) is an advanced version of SimCLR and MoCo [10]. There are two significant changes from SimCLR. The first change is ξ, which is the model parameter θ updated by a moving average, as proposed by MoCo. This implies that instead of predicting each z directly, prediction of z' output is learned using the model parameter ξ. The second change is to add a network h_θ to predict z and train the output p to predict z'. The algorithm is as follows:



(a) SimCLR

(b) BYOL

(c) SimSiam

(d) SimCLRs

Figure 1: Learning architecture of each Siamese Learning.

$$
\begin{aligned}
&x_1, x_2 = \mathrm{aug}(x), \mathrm{aug}(x)\\
&\mathrm{z}_1, \mathrm{z}_2 = g_\theta(f_\theta(x_1)), g_\theta(f_\theta(x_2))\\
&z_1', z_2' = g_\xi(f_\xi(x_1)), g_\xi(f_\xi(x_2))\\
&p_1, p_2 = h_\theta(z_1), h_\theta(z_2)\\
&loss = \mathrm{sim}(p_1, \mathrm{sg}(z_2') + \mathrm{sim}(p_2, \mathrm{sg}(z_1'),
\end{aligned}
$$

Here, sg denotes the stop gradient.

SimSiam (Figure 1c) is a simpler version of BYOL and can achieve the same or better accuracy. Instead of using the moving average model parameter ξ introduced in MoCo and BYOL, p learns to predict each other's z.

$$
\begin{aligned}
&x_1, x_2 = \mathrm{aug}(x), \mathrm{aug}(x)\\
&\mathrm{z}_1, \mathrm{z}_2 = g_\theta(f_\theta(x_1)), g_\theta(f_\theta(x_2))\\
&p_1, p_2 = h(z_1), h(z_2)\\
&loss = \mathrm{sim}(p_1, \mathrm{sg}(z_2) + \mathrm{sim}(p_2, \mathrm{sg}(z_1),
\end{aligned}
$$

This simple method learns representations that are useful for classification.

Surprisingly, BYOL and SimSiam, which remove the explicit contrastive loss, produce different representations for each sample. This result is because of implicit contrastive learning by batch normalization [11].

Out-of-Distribution by Distributionshifting Transformations

OOD detection aims to obtain a measure that distinguishes between samples included in the category of the training data and samples that are not. Rot [12] is an OOD detection method that uses self-supervised learning approaches. This method provides a sample of an image with four different rotations (0°, 90°, 180°, and 270°) and allows the neural network to predict the angle. This is because neural networks can predict the angle of rotation for images of the classes included in the training data, but not for other classes' images. To infer the angle information, it is necessary to capture not only the texture but also the shape.

CSI [6] is a method that combines contrastive learning and distribution-shifting transformations. Distribution-shifting transformations are augmentations that shift the data distribution in N ways. The model learns each feature from the N new samples applied to a single sample as another sample. CSI also has a neural network that can predict the type of shift applied in N. In CIFAR-10, as well as Rot, the most accurate distribution-shifting transformations are the four types of rotations (0°, 90°, 180,°and 270°). CSI constructs representation learning from SimCLR [7] as contrastive learning. In other words, CSI combines SimCLR and Rot. Thus, the model embeds the angle and sample information in detail.

CSI also defines the OOD detection score $S_{con}(x, \{x_m\})$ for sample x in Equation 2, where z is a feature representation, and $g_\theta(f_\theta(x))$ and $\{z_m\}$ are the feature representations of the training data set $\{x_m\}$.

$$S_{con}(x;\{x_m\}) = \max_m sim(z, z_m)\|z\| \tag{2}$$

The smaller the detection score, the higher is the probability that the sample is an OOD sample. This score means that in-distribution features have a higher similarity to the feature representations of the sample in the training data, and OOD features have a lower similarity. However, because the OOD sample is assigned to a random feature representation, it may be similar to the feature representation in the sample. To avoid this problem, the detection score is the similarity multiplied by the norm of z, because the norm of $f z$ is smaller when it is a random feature representation.

The detection score $S_{\text{con-SI}}$, which integrates $S_{\text{con-SI}}$ obtained from each rotated sample x^i, is expressed by the following equation 3.

$$S_{\text{con-SI}}(x;\{x_m\}) = \sum_{i=1}^{4} \lambda_i^{\text{con}} S_{\text{con}}(x^i;\{x_m\}) \tag{3}$$

Here, z^i is the feature representation extracted from the i-th rotation-applied to sample x^i, and the weights calculated from the training data are $\lambda_i^{\text{con}} = N / \sum_m^N \|z_m^i\|$.

Furthermore, the metric S_{cls-SI} is defined using the output $r_\theta^i(f_\theta(x^i))$ for the rotated i-th angle obtained from the rot predictor r_θ. This score is represented by the weights $\lambda_i^{\text{cls}} = N / \sum_m^N r_\theta^i(f_\theta(x^i))$ calculated from the training data, as in Equation 4

$$S_{\text{cls-SI}}(x) = \sum_{i=0}^{4} \lambda_i^{\text{cls}} r_\theta^i(f_\theta(x^i)) \tag{4}$$

The final OOD detection score S_{CSI} in CSI was calculated from S_{con-SI} and S_{cls-SI}, as in Equation 5.

$$S_{\text{cls}}(x,\{x_m\}) = S_{\text{con-SI}}(x;\{x_m\}) + S_{\text{cls-SI}}(x) \tag{5}$$

The final OOD detection score S_{CSI} in CSI was calculated from $S_{\text{con-SI}}$ and $S_{\text{cls-SI}}$, as in Equation 5.

$$S_{\text{CSI}}(x;\{x_m\}) = S_{\text{con-SI}}(x;\{x_m\}) + S_{\text{cls-SI}}(x) \tag{6}$$

A previous study compared these indices and showed that S_{CSI} is the best OOD detection score.

AugMix with Jensen–Shannon Divergence Consistency Loss

AugMix [13] is a method for data augmentation. This can improve the robustness and uncertainty of the model. As a feature, this method adds the Jensen–Shannon divergence (JSD) consistency loss to the loss function. This loss constrains the distance between each output probability distribution to be small when images to which the two types of data augmentation are applied, and the original image is input. The following loss function definesthis loss:

$$
\begin{aligned}
&x_1, x_2 = \mathrm{aug}(x), \mathrm{aug}(x)\\
&y, y_1, y_2 = p(x / l), P_\theta(x_1 / l), P_\theta(x_2 / l)\\
&M = \frac{1}{3}(y + y_1 + y_2)\\
&JSD = \frac{1}{3}(\mathrm{KL}(y / M) + \mathrm{KL}(y_1 / M) + \mathrm{KL}(y_2 / M))
\end{aligned}
$$

where KL is the Kullback–Leibler divergence.

Adding the above loss function to cross entropy loss, the model learns to have probability distributions close to the original image, even when applying data augmentation. This loss function is also robust to noise that is not included in the data augmentation during training.

Method

The purpose of this study was not only to improve the anomaly detection performance of CSI [6] but also to extract better image feature representations. Better image feature representations are general-purpose representations that are easily applied to classification and object detection tasks.

CSI has the following problems regarding the performance of representation learning.

- base representation learning, SimCLR is low performance; and
- rot predictor reduces the performance of representation learning.

The first problem indicates that although the base of representation learning in CSI is SimCLR [7], the accuracy decreases by approximately 10% compared to the new method. The second problem implies that training with a rot predictor reduces the performance of representation learning. To solve these problems, we propose a method (Figure 2) that introduces the two algorithms detailed in the following subsections.

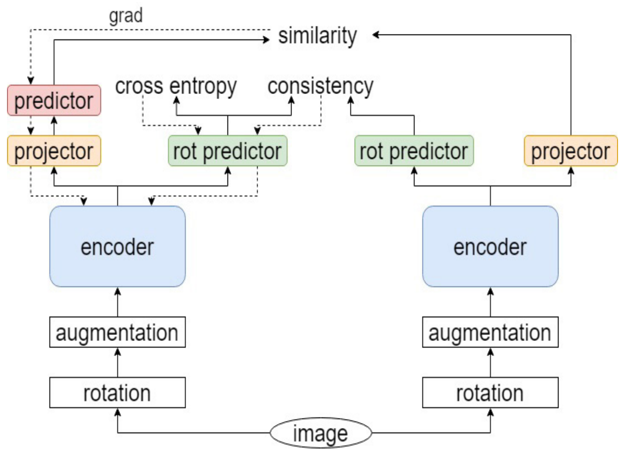


Figure 2: proposed method.

Contrastive learning like SimSiam

SimSiam [9] is a representation learning method that differs from SimCLR, mainly in the following three aspects:

1. projector composed of multiple layers;
2. adding a network to predict each other's feature representation; and
3. implicit contrastive learning.

Because of these differences, SimSiam can perform representation learning with a higher accuracy than Sim- CLR, as described in related works.

SimSiam projects different samples to different representations without explicit contrastive learning because of batch normalization[11] in the network. In addition, implicit contrastive learning means that there is no need to constrain the different samples represented differently by a loss function since the final classification falls into some categories. The classification accuracy is high, even if images of the same class have highly similar feature representations.

In contrast, in good detection, matching all representations to the central representation of each class may lead to overlearning. This problem arises because learning features representing in-distribution well are quite different from learning features representing outof-distribution. Therefore, it is necessary to test whether explicit contrastive learning or implicit contrastive learning is better in such situations.

JSD consistency loss

Training with a rot predictor reduces representation learning performance, because cross entropy loss ignores the consistency of the feature representation until the probability distribution is completely sharp. However, the ideal probability distribution output is not one-hot, and the probability distributions between similar samples need to be closer than those between different samples. The usual cross entropy cannot learn such a consistency. To solve this problem, we introduce the loss function such that the expressions of the probability distributions y_1, y_2 that predict the rotation angles of x_1, x_2 are consistent. In this study, we adopted AugMix [13] and introduced the following JSD: suppose the one-hot distribution of the correct value of the rotation angle yt, the distribution of the predicted probability of the rotation angle y_1, y_2, and the mean of the three distributions M.

The JSD consistency loss is represented by Equation 6.

$$JSD = \frac{1}{2}(\mathrm{KL}(y_1|M), \mathrm{KL}(y_2|M)) \qquad (6)$$

As in AugMix, this loss function reduces overoptimization for cross entropy, which interferes with representation learning and encourages feature representations to match between x_1, x_2 in the process of learning consistency. Whereas AugMix is constrained using the distribution y obtained from the original image, our As in AugMix, this loss function reduces overoptimization for cross entropy, which interferes with representation learning and encourages feature representations to match between x_1, x_2 in the process of learning consistency. Whereas AugMix is constrained using the distribution y obtained from the original image, our method calculates the loss function using the correct answer value. This loss is multiplied by 12, similar to AugMix, and added to the cross entropy.

Experiment

In this study, we conducted experiments to demonstrate that the new contrastive learning and JSD consistency loss improve OOD detection accuracy and produces better feature representations. In the proposed method, one of the scores defined by CSI [6] was set as the score for OOD detection. We also considered the usefulness of feature representation for classification. We fixed the network after the contrastive learning training, and trained only the added linear layer supervised to show the accuracy of the classification.

Dataset

We set CIFAR-10 as the training dataset, and the following seven datasets as the OOD: SVHN, LSUN, ImageNet, CIFAR-100 and Interp. OOD datasets are equivalent to the experimental condition of Unlabelled CIFAR-10 in CSI [6]; however, this study regards LSUN(FIX) and ImageNet(FIX) in the previous paper as LSUN and ImageNet.

Contrastive learning

SimSiam [9] and SimCLRs were used as the experimental targets for the contrastive learning method. The hyperparameters of SimSiam, such as architecture, data augmentation, and learning rate, follow those shown in previous studies. In addition, SimCLRs shares all the hyperparameters with SimSiam, except for the loss function, as shown in Figure 1d. However, the loss function is represented by D_{simclr}. SimCLRs follows the following algorithm:

$$x_1, x_2 = \text{aug}(x), \text{aug}(x)$$

$$\mathrm{z}_1, \mathrm{z}_2 = g_\theta(f_\theta(x_1)), g_\theta(f_\theta(x_2))$$

$$z'_1, z'_2 = g_\xi(f_\xi(x_1)), g_\xi(f_\xi(x_2))$$

$$p_1, p_2 = h_\theta(z_1), h_\theta(z_2)$$

$$loss = D_{simclr}(p_1, \text{sg}(z'_2) + D_{simclr}(p_2, \text{sg}(z'_1),$$

Contrastive learning

We showed that the addition of a rotation angle prediction network, the rot predictor, degrades the performance of representation learning in the previous method. We also demonstrated that the incorporation of JSD consistency loss in training the rot predictor improves the performance of representation learning. To highlight these results, we compared the following four types of methods.

- w/o rot
- w/ rot, w/o rot predictor
- w/ rot, w/ rot predictor, w/o JSD consistency loss
- w/ rot, w/ rot predictor, w/o JSD consistency loss

The first is the baseline, which is used when OOD detection is not required. The second method trains images with four different rotations. The four rotations quadruple the batch size; therefore, the batch size of the original image should be 1/4. The third method is the same as CSI, which is trained with the rot predictor. In this case, the total loss is the sum of the similarity loss for representation learning and the cross entropy loss for learning the rot predictor. Finally, we added the proposed JSD loss to the third one during training.

OOD detection score with JSD

In this experiment, we adopted $S_{\text{con-SI}}$, represented by the expression 3 as the OOD detection score in the proposed method, that is, CSI with JSD consistency loss. The OOD detection scores of the comparison targets correspond to those of the original CSI, as shown in Table 1.

ROT	PRED	JSD	ODD
-	-	-	S_{con}
✓	-	-	$S_{\text{con-SI}}$
✓	✓	✓	S_{CSI}
✓	✓	✓	$S_{\text{con-SI}}$

Table 1: Correspondence between presence/absence of each method and OOD detection score.

ROT denotes the presence of rotation augmentation, PRED denotes the presence of a rotation predictor, and JSD denotes the presence of a JSD consistency loss.

Parameters

ResNet18 [3] was used as the neural network f_θ to extract the features. In addition, in both SimCLRs and Sim-Siam, according to [9], the projector g_θ is constructedas shown in Table 2, and the predictor h_θ, as shown in Table 3.

Layer name	Input size	Output size
Linear	512	2048
BN	2048	2048
ReLU	2048	2048
Linear	2048	2048
BN	2048	2048

Table 2: projector

Layer name	Input size	Output size
Linear	2048	512
BN	512	512
ReLU	512	512
Linear	512	2048

Table 3: predictor

The original SimCLR does not have a predictor network, and the projector network consists of a single layer with an input size of 512 and an output size of 2048.

In this study, the rot predictor was constructed as shown in Table 4.

Layer name	Input size	Output size
Linear	2048	2048
BN	2048	2048
ReLU	2048	2048
Linear	2048	512

Table 4: rot predictor.

The rot predictor in the original CSI is composed of a single layer.

Results

The experimental results of this study were compared with those of a previous study, CSI [6], in Table 5.

SimSiam and SimCLRs both exhibit the best performance in representation learning when the proposed method (i.e., the incorporation of JSD) was introduced. The area under the ROC curve (AUROC) of OOD detection were higher than that of the existing CSI and without JSD for all OOD datasets except for SVHN and Interp. However, the accuracy of AUROC for SVHN was lower than that of previous studies. Furthermore, in applying SimSiam to Interp, the highest AUROC was obtained when only ROT was applied.

Discussion

Consistency loss for rotation predictor

The experimental results clearly indicate that training the rot predictor with a cross entropy without the JSD consistency loss reduces the classification accuracy. In particular, the classification

Method	ROT	PRED	JSD	Accuracy IN-D	OOD AUROC between CIFAR-10 and				
					SVHN	LSUN	ImageNet	CIFAR-100	nterp
SimCLR(CSI[6])	✓	✓	-	-	**99.8**	90.3	93.3	89.2	79.3
SimCLRs	-	-	-	89.9	86.3	89.7	88.0	85.5	81.0
SimCLRs	✓	-	-	92.2	97.4	91.0	91.5	90.1	82.1
SimCLRs	✓	✓	-	90.9	98.9	89.7	92.8	89.5	79.3
SimCLRs(ours)	✓	✓	✓	**92.3**	98.0	**92.3**	**93.4**	**90.8**	82.8
SimSiam	-	-	-	91.4	91.4	91.7	90.5	87.7	81.9
SimSiam	✓	-	-	91.9	97.4	91.0	91.4	89.8	82.9
SimSiam	✓	✓	-	89.9	99.1	89.1	92.3	88.9	78.7
SimSiam(ours)	✓	✓	✓	**92.3**	99.2	**92.3**	**93.4**	90.1	81.5

Table 5: Classification accuracy and OOD detection AUROC. ROT is the rotation augmentation; PRED means rot predictor; and IN-D means in-distribution (CIFAR-10 in this experiment).

accuracy was lower than that without the rot predictor. Thus, OOD detection by conventional CSI has a trade-off relationship with the classification.

Adding the JSD consistency loss addresses this tradeoff and improves the accuracy of OOD detection and the classification accuracy for the original contrastive learning. This result was not just because they suppressed the cross entropy. It is possible to learn better intermediate representations (i.e., feature representations) by introducing a loss function such that the probability distributions of image samples from the same image at the exact angle match.

Detection score with JSD consistency loss

In our experiments, S_{con-SI}, which is represented by Equation 3, is adopted as the OOD detection score in the proposed method. The reason for using S_{con-SI} instead of SCSI is that the introduction of this loss improves the accuracy of OOD AUROC, compared to using S_{cls-SI} in combination. This is because the accuracy of representation learning is improved. However, this loss reduces the angle prediction accuracy of the rot predictor. Therefore, the accuracy of S_{cls-SI} and S_{CSI} decreased compared to the accuracy without the JSD consistency loss.

Contrastive learning method for accuracy

As shown in the experimental results, regular SimSiam is more accurate than SimCLRs for unsupervised class classification using representation learning. On the other hand, when rotation augmentation is applied, SimSiam accuracy is lower. This is probably because the rotated images should be treated as separate samples, but they have very close feature values. In SimCLRs, the feature representations of the images in a batch are explicitly specified to be different; therefore, even if they are initially from the same samples, they are learned to have different representations for each rotated sample. In contrast, SimSiam implicitly performs contrastive learning through normalization. Therefore, if the image representations are genuinely close, they can be the same even after normalization. Because the rotated images are close to each other, it is difficult to determine whether they should be closer or further away from other data augmentation methods. It is assumed that this conflicts with cross entropy learning for separation and reduces the accuracy.

In contrast, there is little difference in the absolute accuracy when adding the JSD consistency loss. Cross entropy prevented the learning of feature representations from conflicting with the learning of feature representations, thereby moderating the intense conflicts that occurred, especially in SimSiam. Moreover, the model learns the feature representation through the predictive distribution of the rotation angle.

It is demonstrated that the final accuracy is affected by providing data whose distribution within a batch differs significantly from that of regular representation learning. The proposed method, with the JSD consistency loss, reduces this effect.

Contrastive learning method for OOD-detection

For regular classification applications, SimSiam [9] outperformed the original SimCLR [7] and SimCLRs. In contrast, in OOD-Detection, different results were obtained for different tasks. SVHN has a very high AUROC of 99.9% in the original CSI. However, even Sim-Siam + CSI + JSD, which is the best of the proposed methods, achieves 99.2%, which is lower than the original CSI. In addition, the accuracy is not consistent between SimCLRs and SimSiam. Furthermore, in Interp, which is a dataset of two blended images from CIFAR-10, the accuracy is inconsistent between SimCLRs and SimSiam. These datasets are considered to be more affected by the contrastive learning method than the others.

Future Works

This study aimed to balance the performance of feature representation and OOD Detection. However, it is necessary to verify how the performance of OOD detection is affected by training classifications on pre-trained networks. We will conduct experiments to verify how the accuracy changes when training on a more extensive network and larger image size.

Although many datasets have higher OOD detection accuracy than CSI in existing studies, the difference between datasets with inferior results should be clarified. In particular, clarifying the difference in accuracy by contrastive learning leads to further improvements in OOD detection accuracy.

Conclusions

The purpose of this study was to detect OOD data and extract feature representations useful for classifying classes. As a result, the proposed method improves the accuracy of both processes.

We propose adding the JSD consistency loss, which constrains the distribution of the predicted probability of the rotation angle to be consistent among the same samples. To improve the basic accuracy, we also experimented with SimSiam [9] instead of SimCLR [7], which was the basis of the previously reported CSI [6]. In addition, SimCLRs, in which SimCLR replaced only the loss function of SimSiam, was set up, and a comparison experiment was conducted.

The experimental results show that the proposed method, which incorporates the JSD consistency loss, outperforms the existing methods for both IN-D classification accuracy and most OOD datasets for both Sim-Siam and SimCLRs. These results show that the proposed method achieves both classification and OOD detection with high accuracy.

However, the loss function of SimSiam and SimCLRs does not make a significant difference in the final result. The incompatibility of implicit contrastive learning causes rotational augmentation, which yields similar images. Although the JSD consistency loss moderates this conflict to the same level as explicit contrastive learning, we will study contrastive learning for more enhanced OOD detection in the future.

Competing Interests

The authors declare that they have no competing interests.

References

1. Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Communications of the ACM 60: 84-90.
2. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recog-nition. Published as a conference paper at ICLR.
3. He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778.
4. Zagoruyko S, Komodakis N (2016) Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, Proceedings of the British Machine Vision Conference (BMVC), pp. 87.1-87.12.
5. Tan M, Le QV (2019) EfficientNet: Rethinking Model Scaling for Convolutional Neural Net-works.
6. Tack J, Mo S, Jeong J, Shin J (2020) Csi: Novelty detection via contrastive learning on distribu-tionally shifted instances. 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.
7. Tian Y, Yu L, Chen X, Ganguli S (2020) Understanding self-supervised learning with dual deep networks. 34th Conference on Neural Information Processing Systems (NeurIPS 2020).
8. Grill JB, Strub F, Altché F, Tallec C, Richemond P, et al. (2020) Bootstrap your own latent - a new approach to self-supervised learning. 33: 21271-21284.
9. Chen X, He K (2021) Exploring simple siamese representation learning. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
10. He K, Fan H, Wu Y, Xie S, Girshick R (2020) Momentum contrast for un-supervised visual rep-resentation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pat-tern Recognition (CVPR).
11. Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on International Con-ference on Machine Learning 37: 448-456.
12. Hendrycks D, Mazeika M, Kadavath S, Song D (2019) Using self-supervised learning can im-prove model robustnessand uncertainty. 33rd Conference on Neural Information Processing Sys-tems (NeurIPS 2019), Vancouver, Canada.
13. Hendrycks D, Mu N, Cubuk ED, Zoph B, Gilmer J, et al. (2020) AugMix: A simple data processing method to improve robustness and uncertainty. Proceedings of the International Conference on Learning Representations (ICLR).

This article was originally published in a special issue:

Computational Analysis and Modeling

Handled by Editor:

Prof. Shigeo Akashi
Department of Information Sciences
Tokyo University of Science
Japan