Computer Vision and Image Understanding 241 (2024) 103968

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Exploring using jigsaw puzzles for out-of-distribution detection

Yeonguk Yu, Sungho Shin, Minhwan Ko, Kyoobin Lee *

Gwangju Institute of Science and Technology (GIST), 123, Cheomdangwagi-ro, Buk-gu, Gwangju (61005), Republic of Korea

ARTICLE INFO

Communicated by Juergen Gall

MSC:
41A05
41A10
65D05
65D17

Keywords:
Neural networks
Image recognition
Out-of-distribution detection

ABSTRACT

Out-of-distribution (OOD) detection involves binary classification whether the given data is from outside the training data or not. Previous studies proposed outlier exposure (OE) that trains the model on an outlier dataset designed to represent potential future OOD data, thereby enhancing OOD detection performance. However, obtaining an outlier dataset representing all possible future OOD data can be challenging, and such dataset may be unavailable in some cases. This study proposes a novel approach to expose the model to jigsaw puzzles generated from training images as the outlier data. Specifically, the model is trained to have a low LogitNorm for given jigsaw puzzles. We argue that jigsaw puzzles can effectively represent future OOD data because they contain similar background information as the in-distribution data but with their semantic information destroyed. Our experimental results demonstrate that our approach outperforms previous competitive OOD detection methods and effectively detects semantically shifted OOD examples. Our code is available at https://github.com/gist-ailab/jigsaw-training-OOD.

1. Introduction

In machine learning, out-of-distribution (OOD) detection aims to recognize images outside trained data. This is important because models trained on a specific data distribution may not generalize well to data outside that distribution, leading to unreliable or incorrect predictions (Hendrycks and Gimpel, 2017; Liu et al., 2020). To illustrate this concept, consider a model trained to classify images of cats and dogs using a large dataset of related images. The model can accurately classify new images of these animals with high accuracy. However, if presented with an image of a car, the model performs poorly in classification because it has not encountered this type of data during training. Examples are illustrated in Fig. 1. In this case, it would be important for the model to have the ability to detect that the image is an OOD case and to alert the user that it is unusable for the model.

Various approaches have been employed to differentiate between in-distribution (ID) and out-of-distribution (OOD) data using neural network outputs. Output probability (Hendrycks and Gimpel, 2017; Liang et al., 2018) and output energy (Liu et al., 2020) are among the commonly used indicators. The output of a neural network is determined by a feature and a weight vector from the feature extractor and classification layer, respectively. In addition, training methods that produce deactivated features, resulting in low output confidence on data lacking in-distribution information, have been studied. For instance, OE (Hendrycks et al., 2019) uses an auxiliary dataset containing OOD objects to train the model, whereas virtual outlier synthesis (VOS Du et al., 2022) employs sampling strategy to generate OOD feature and train the model to produce low confidence on them. Compounded Corruptions (CnC Hebbalaguppe et al., 2023) uses hard augmentation-based corruption technique to generate OOD images and train the model to classify them as $K+1$ category. Exposing the model to proxy OOD, which are images used to represent OOD (2), has improved OOD detection performance. More recently, jigsaw puzzles have been used to identify the most appropriate block for OOD detection after completing the training (Yu et al., 2022). Therefore, we argue that jigsaw puzzles can serve as a proxy OOD to enhance OOD detection performance, as seen in previous pre-hoc methods.

This study investigates using jigsaw puzzles as proxy OOD during training stage to improve OOD detection performance. Specifically, we use jigsaw puzzles generated from the training images as proxy OOD so that the model is trained to produce relatively low L_2**-norm** of the **logit** (dubbed as **LogitNorm**). We argue that jigsaw puzzles have no spatial relationship, unlike ID images. Thus, model training with them can makes the model to produce high LogitNorm for given ID images while producing low LogitNorm for given OOD images during inference stage.

We provide analyses of our proposed training method with jigsaw puzzles. We conduct experiments on common OOD detection benchmarks and show that our simple method is effective. The key results and contributions of our study are summarized as follows:

- We introduce a novel training method to improve the OOD detection performance with jigsaw puzzles, where the model is trained to produce low LogitNorm for given jigsaw puzzles. Novelty of

* Corresponding author.
E-mail addresses: yeon_guk@gm.gist.ac.kr (Y. Yu), hogili89@gm.gist.ac.kr (S. Shin), mhko1998@gm.gist.ac.kr (M. Ko), kyoobinlee@gist.ac.kr (K. Lee).

https://doi.org/10.1016/j.cviu.2024.103968
Received 2 August 2023; Received in revised form 13 December 2023; Accepted 12 February 2024
Available online 15 February 2024
1077-3142/© 2024 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

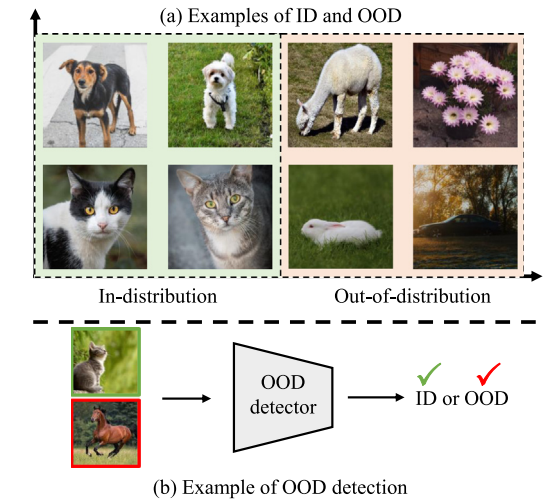


Fig. 1. Illustration of in-distribution (ID) and out-of-distribution (OOD) images (a) for a model trained to classify cats and dogs. The concept of the OOD detection is illustrated in (b). A reliable OOD detector should classify cat as ID and horse as OOD.

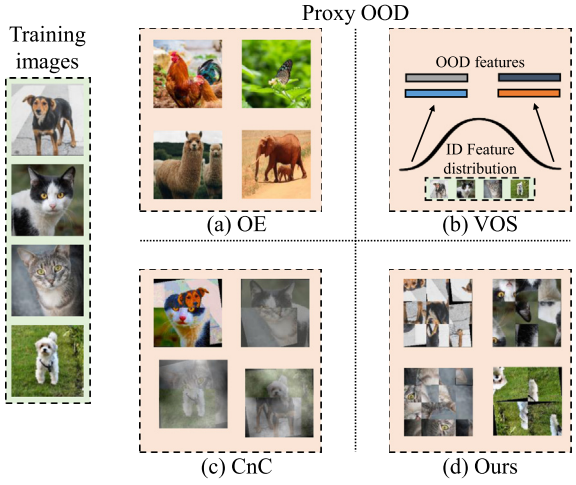


Fig. 2. Comparison of proxy OOD of various methods. For the given training images, outlier exposure (OE) exposes the model to auxiliary images (a), virtual outlier synthesis (VOS) exposes the model to OOD features, which are sampled from low-likelihood region of the class-conditional distribution estimated in the feature space (b), compound corruptions (CnC) exposes the model to corrupted images, generated by patch-based convex combination (e.g., Mixup Yun et al., 2019) and hard-augmentation (c), and proposed method exposes the model to jigsaw puzzles (d).

the paper lies in exposing the model to jigsaw puzzles as proxy OOD, which have not been previously investigated to the best of our knowledge.

- We conduct comprehensive experiments and show that our method consistently improves the OOD detection performance of the model and outperforms previous baselines.
- We conduct extensive analyses that demonstrates the effectiveness of our proposed method.

2. Related works

2.1. Out-of-distribution detection

Many areas in deep learning aim to detect unknown or novel input, which are different from the seen input in training distribution. For example, novelty detection aims to determine if a test data sample is normal or anomalous (known class vs. novel class) (Salehi et al., 2021; Almohsen et al., 2023; Lo et al., 2023). Similarly, OOD detection aims to determine if a test sample is from in-distribution or out-of-distribution. The main difference is that the novelty detection is based on the generative model for one-class training set, while OOD detection is based on discriminative model for multi-class training set. For instance, Salehi et al. proposed to use augmented auto-encoder with adversarial samples to determine input by degree of reconstruction, while Hendyricks et al. proposed to use output confidence. Although, novelty detection uses generative model mainly, insight from these approaches can also be used in OOD detection.

OOD detection with discriminative models can be categorized as post-hoc or pre-hoc methods. Post-hoc methods do not require modifications to the training method or architecture and can be applied to off-the-shelf models. Examples of post-hoc methods include maximum softmax probability (MSP Hendrycks and Gimpel, 2017) and its enhanced version (ODIN Liang et al., 2018), which uses input preprocessing and temperature scaling to distinguish the confidence of ID and OOD samples. In addition, the energy function (Liu et al., 2020) is also post-hoc method that utilize energy as an OOD indicator. Recently, Yu et al. (2022) proposed using the norm of the feature map for OOD detection.

In contrast, pre-hoc methods involve modifying the model to improve its OOD detection performance during the training stage. For example, OE (Hendrycks et al., 2019) trains the model to have low confidence in auxiliary data, resulting in enhanced OOD detection performance when combined with post-hoc methods. IsoMax loss (Macêdo et al., 2021), which replaces the softmax cross-entropy loss and follows the maximum entropy principle, is another pre-hoc method. Although OE-based approaches can significantly enhance OOD detection performance, they require access to an auxiliary dataset, which may not always be practical. Recently, Du et al. (2022) proposed to generate proxy OOD feature which are sampled from low-likelihood region of the class-conditional distribution. Also, Hebbalaguppe et al. (2023) proposed to generate proxy OOD using patch based convex combination and hard-augmentation. Our proposed method involves OE with augmented jigsaw puzzles and can be considered a pre-hoc method. Hence, we compare our method with pre-hoc methods.

2.2. Jigsaw puzzles in neural networks

The jigsaw puzzle technique was first introduced for use in computer vision tasks (Noroozi and Favaro, 2016) to predict image patch sequences. It demonstrated that solving jigsaw puzzles can help the model extract features for the original task. Jigsaw puzzles are actively used for pretext tasks learning in representation learning. For example, iterative reconstruction of jigsaw puzzles in high-dimensional space is used to initialize a network with transfer learning (Wei et al., 2019). In addition, Paumard et al. proposed Deepzzle that solves jigsaw puzzles with the network, which predict relative positions of two fragments and shortest path optimization on the graph (Paumard et al., 2020). Similarly, the jigsaw clustering method that clusters fragments from different images revealed enhanced performance (Chen et al., 2021). Also, Salehi et al. proposed novelty detection method using autoencoder, where the model is trained to solve anti-shortcut jigsaw puzzles for normal data and detect abnormal samples by choosing test samples fail to solve puzzles (Salehi et al., 2020). Unlike this work, we use jigsaw puzzles as proxy OOD for training a classification model and detect the OOD sample in the inference-stage by choosing test samples fail to produce meaningful logit. More recently, jigsaw puzzles are utilized as proxy OOD to select the more appropriate block for OOD detection (Yu et al., 2022). However, this is not pre-hoc method since it does not modify any training loss. In this work, we use jigsaw puzzles as proxy OOD for improving the OOD detection performance of neural networks during training.

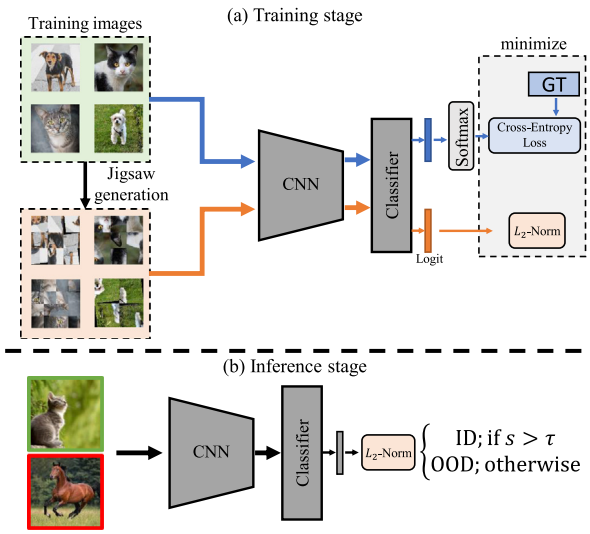


Fig. 3. Proposed framework where the model is exposed on jigsaw puzzles and trained to minimize the norm of logits for given jigsaw puzzles (a). During the inference stage, the test image is considered as ID when the logit is higher than the threshold (b); otherwise, the image is considered as OOD.

3. Proposed method

This study proposes the framework using jigsaw puzzles during training stage to improve OOD detection performance. This framework is based on the concept that the network that has learned about spatial relationships of ID objects will produce low confidence for the given OOD objects with destroyed spatial relationship of ID objects. The model learns to classify ID images with correct spatial relationships while also training to output low LogitNorm for jigsaw puzzle images where spatial relationships have collapsed. Therefore, we train the model to have low LogitNorm for given jigsaw puzzles. Fig. 3 illustrates the proposed framework.

We first describe the general setting of image classification and jigsaw puzzle generation to ease understanding (Section 3.1). Next, we describe the training procedure of our framework (Section 3.2). Lastly, we present the OOD detection method with the network (Section 3.3).

3.1. Preliminaries

We first provide an overview of the supervised learning problem in the image classification network. Specifically, the network is trained using cross-entropy loss on a training dataset denoted as $D_{in} = (x_i, y_i)_{n=1}^{N}$, where $x_i \in \mathbb{R}^{3\times W\times H}$ represents the RGB image input, and $y_i \in 1, 2, \dots, K$ represents the label with K class categories. The OOD detection method qualifies as a post-hoc method if it does not modify during the training phase and as a pre-hoc method otherwise.

We generate jigsaw puzzles from the training images to incorporate them into model training. We create 3 × 3 jigsaw puzzles, a format commonly used in previous research studies (Esteva et al., 2021; Chen et al., 2021). Our primary objective is to leverage jigsaw puzzles as a proxy OOD to improve OOD detection performance, as they lack spatial relationship, rather than to enhance the jigsaw puzzle itself.

3.2. Training loss: exposure on jigsaw puzzles

To train the model with training images and jigsaw puzzles, which are generated from training images, we utilize two loss: (1) cross-entropy loss and (2) L_2-norm loss as shown in Fig. 3-a. We use the cross-entropy loss for training the classification network and the L_2-norm loss to improve the OOD detection performance of the classification network. Our design of the L_2-norm loss is based on the previous research that the norm of the logit affects the confidence of the model (Xu et al., 2020; Wei et al., 2022). Also, we do not utilize ReLU function in the training stage unlike the inference stage as it helps us to produce more calibrated model. Therefore, the model will produce confidence according to the spatial relationship by training the model to have low confidence on the jigsaw puzzles and high confidence on the original training images.

In particular, the training loss of the model, $\mathcal{L}$, is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{norm}, \tag{1}$$

where the L_{ce} is the cross-entropy loss between the probability distribution produced for training images and ground truth. Moreover, $\mathcal{L}_{norm}$ and λ refer to the loss for jigsaw puzzles and their corresponding weight, respectively. $\mathcal{L}_{ce}$ and $\mathcal{L}_{norm}$ are formulated as follows:

$$\mathcal{L}_{ce} = \mathbb{E}_{(x,y)\sim D_{in}}[\log(p_y)] \tag{2}$$

$$\mathcal{L}_{norm} = \mathbb{E}_{(x)\sim D_{jigsaw}} \|v\|_2^2, \tag{3}$$

where p_y, v, and $\|\cdot\|_2^2$ refer to probability for given ground-truth class y, calculated logit for jigsaw puzzles, and L_2-norm for the given vector, respectively. Thus, the model is trained to have zero norm value for jigsaw puzzles to minimize $\mathcal{L}_{norm}$. Moreover, the model is trained to classify the original image as the ground-truth class to minimize $\mathcal{L}_{ce}$. For the weight parameter λ, we use 1.0 for all experiments.

3.3. OOD detection: using LogitNorm

For the OOD detection task, we use the norm of the elements larger than the threshold α in the logits for the given test image as an indicator value showing closeness of the test image to the ID images (i.e., ID-ness). We consider the larger elements in logits, as the small elements can be in logits for given OOD images (e.g., negative value or confused small positive value in a logit also increases the norm). Thus, to improve the OOD detection performance, we filter the unnecessary value below the confident threshold α (i.e., the small values that do not contribute to the decision of final prediction). The α is calculated by simply averaging the second-largest element in the logit for all training images as follows:

$$\alpha = \frac{1}{N}\sum_{i=1}^{N} \hat{v}_i, \tag{4}$$

where the N and $\hat{v}_i$ refer to the number of training images and the second-largest value in ith logit respectively. Note that the second-largest value represents the model's confusing score for the incorrect class, while the largest value represents the model's confident score for the correct class. Thus, α can be interpreted as the confident threshold in the logit.

Subsequently, if LogitNorm is sufficient, we consider the test image as an ID image as shown in Fig. 3-b. The OOD detection using our framework is formulated as follows:

$$G(x; f) = \begin{cases} \text{ID} & \text{if } \|\text{ReLU}(v - \alpha)\|_2^2 > \tau \\ \text{OOD} & otherwise, \end{cases} \tag{5}$$

where $\|v\|_2^2$ is the norm of the logit for the given test image x. τ is the chosen threshold so that 95% of the ID data is correctly classified (i.e., true positive rate of 95%) using the trained network f. Thus, our OOD detector classifies the image as ID and OOD when the norm of the logit is sufficiently high and low, respectively.

Table 1
Performance of OOD detection on CIFAR benchmarks. The methods have no access to OOD data during training and validation. The best result is indicated in bold. All values are percentages averaged over five runs.

In-distribution	Method	OOD												Average	
		SVHN		Textures		LSUN(c)		LSUN(r)		iSUN		Places365			
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CIFAR10	Baseline	26.30	96.15	9.54	97.98	25.24	96.30	30.27	95.51	28.69	95.72	25.90	95.13	24.32	96.13
	VOS	8.09	98.31	1.59	99.48	8.45	98.19	13.10	97.59	11.82	97.83	13.02	97.06	9.34	98.08
	CnC	**0.30**	**99.93**	**0.02**	**99.99**	0.95	99.77	6.51	98.46	8.21	98.07	**0.00**	**100.00**	2.67	99.37
	Ours	2.60	99.40	0.03	**99.99**	**0.36**	**99.93**	**1.29**	**99.75**	**1.58**	**99.67**	0.04	99.99	**0.98**	**99.79**
CIFAR100	Baseline	68.05	83.91	55.41	88.08	63.09	86.32	61.11	87.43	65.00	85.60	65.83	84.39	63.08	85.95
	VOS	34.07	92.13	17.46	96.65	38.75	92.27	38.40	92.40	43.27	90.96	40.93	91.35	35.48	92.63
	CnC	**2.65**	**99.33**	**0.28**	**99.93**	16.91	96.22	28.04	93.14	35.51	90.13	0.04	**99.99**	13.91	96.46
	Ours	20.95	94.18	0.52	99.49	**6.66**	**98.44**	**8.76**	**98.00**	**17.65**	**95.77**	**0.03**	99.57	**9.10**	**97.57**

4. Experiments

In this section, we compare our proposed framework with other pre-hoc methods. Following previous works (Du et al., 2022; Liu et al., 2020; Sun et al., 2021; Yu et al., 2022), we use CIFAR (Krizhevsky et al., 2009) benchmarks with CIFAR10 and CIFAR100 as the ID datasets, and SVHN (Netzer et al., 2011), Textures (Cimpoi et al., 2014), LSUN (Yu et al., 2015), iSUN (Xu et al., 2015), and Places365 (Zhou et al., 2018) as OOD datasets. Detailed experimental settings are as follows.

4.1. Training settings

We use a vision transformer (ViT; Dosovitskiy et al., 2021) with a patch size of 16, input image size of 224, depth of 12, and embedding size of 192 (a.k.a. 'Tiny' variant or ViT-T/16) for experiments. ViT was pretrained with ImageNet (Deng et al., 2009) using the open-released checkpoint by timm repo (Wightman, 2019) and finetuned on ID dataset. Since the ViT is pretrained on ImageNet, we train the ViT with an initial learning rate of 0.003 using SGD optimizer with momentum 0.9. In addition, the model is finetuned using a batch size of 128 for 15 epochs with a weight decay of 0.0005. All input image were resized to 224×224. For the weight parameter of $\mathcal{L}_{norm}$ λ, we use 1.0.

4.2. Evaluation metrics

We adopted the two commonly used evaluation metrics: FPR95 and AUROC.

- **FPR95** refers to the false positive rate (FPR) at a true positive rate (TPR) of 95% (i.e., FPR@TPR95); that is, the OOD detection threshold is set to obtain TPR 95%. Thus, a smaller FPR95 indicates better OOD detection performance.
- **AUROC** stands for the area under the receiver operating characteristic curve. It can be interpreted as the probability of correctly classifying an OOD sample as OOD for the given randomly selected sample. A higher AUROC indicates better performance in detecting OOD samples.

4.3. Baseline methods

We compare our method against a baseline method and two pre-hoc methods that use proxy OOD during the training stage. The methods used are as follows:

- For Baseline (Hendrycks and Gimpel, 2017), we use the cross-entropy-trained ViT. The OOD detection performance of MSP indicates the low limit of OOD detection performance using ViT.
- For VOS (Du et al., 2022), we expose the ViT with OOD features sampled from the low likelihood region of the class-conditional distribution estimated in the feature space using the last block. For other hyperparameters, we followed the setting for CIFAR10 in Du et al. (2022). Moreover, we use the same training setting to train the ViT with generated proxy OOD features and use energy as the OOD score after training.
- For CnC (Hebbalaguppe et al., 2023), we follow their settings to generate proxy OOD for CIFAR10 benchmark. Subsequently, we train the ViT to classify generated proxy OOD as $K+1$ category with our training setting and use the probability of $K+1$ category as the OOD score after training.

5. Results

We provide our experimental results on CIFAR benchmarks in Table 1. We report the performance of OOD detection for the ViT architecture using previous pre-hoc OOD detection methods. The performance is calculated using FPR95 and AUROC on six OOD datasets: (1) SVHN (Netzer et al., 2011), (2) Textures (Cimpoi et al., 2014), (3) LSUN(c) (Yu et al., 2015), LSUN(r) (Yu et al., 2015), iSUN (Xu et al., 2015), and Places365 (Zhou et al., 2018). LSUN(c) and LSUN(r) refer to the dataset containing center-cropped and resized image from the original dataset. Our proposed method achieved the best average performance on CIFAR10 and CIFAR100. Our method reduces the average FPR95 by **62.92%** and **25.52%** compared to the second best results on CIFAR10 and CIFAR100, respectively.

As shown in Table 1, the performance of CnC on SVHN outperforms our method, where SVHN consists of blurred images. We argue that the hard augmentation of CnC contributes towards improvement, where the model is exposed to blurred images generated by hard augmentation. In contrast, our method outperforms other methods on LSUN(c), LSUN(r), and iSUN which does not contain blurred images.

6. Discussion

This section provides extensive experiments and analysis to explain the proposed framework. First, we give the classification accuracy on ID test set (Section 6.1). Then, we provide the OOD detection performance using various post-hoc methods instead of LogitNorm on the network trained by our framework (Section 6.2). Subsequently, we provide the performance when detecting semantically shifted OOD using our model (Section 6.4). Lastly, we demonstrate the location of jigsaw puzzles in the feature space (Section 6.4).

6.1. Classification accuracy evaluation

Maintaining high in-distribution classification accuracy is also a significant challenge for pre-hoc based OOD detection methods in real-world scenarios. Our approach trains the model with L_2-norm loss for given jigsaw puzzles during the training stage. It may reduce classification accuracy since it forces the network to produce low confidence

Table 2
Classification accuracy evaluation of various pre-hoc methods. The best result is indicated in bold. All values are percentages averaged over five runs.

Training method	ID training set	
	CIFAR10	CIFAR100
Baseline	97.37	87.35
VOS	96.55	**87.45**
CnC	97.09	87.20
Ours	**97.45**	87.19

Table 3
Average OOD detection performance using various post-hoc methods on our model. The best result is indicated in bold. All values are percentages averaged over five runs.

Training method	Detection method	Average performance	
		FPR95↓	AUROC↑
Baseline	MSP	24.31	96.13
L_2-Norm loss (Ours)	MSP	3.19	99.42
	ODIN	1.69	99.66
	Energy	1.40	99.73
	LogitNorm (Ours)	**0.98**	**99.79**

Table 4
Performance of detecting OOD in different in-distribution dataset. The models are trained on ImageNet and MNIST and evaluated on detecting OOD. The best result is indicated in bold and the second-best result is indicated in underline.

Method	In-distribution dataset			
	ImageNet		MNIST	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Baseline	68.59	80.49	3.54	<u>98.87</u>
VOS	<u>51.03</u>	<u>85.68</u>	<u>2.42</u>	**99.47**
CnC	73.40	76.26	6.37	98.50
Ours	**49.91**	**86.47**	**1.90**	**99.47**

for given similar images with different spatial relationship. To evaluate our approach in this perspective, we present a classification accuracy for ID test set in Table 2. Our method results in a small improvement in classification accuracy for CIFAR10 compared to the baseline trained by cross-entropy loss. However, our model has a small reduction for CIFAR100. We argue that this small reduction in classification accuracy is acceptable for improving the OOD detection performance. Moreover, the model trained with our framework has similar accuracy as models trained by other pre-hoc based OOD detection methods.

6.2. Comparison with other post-hoc methods

As other post-hoc OOD detection methods can be applied to our model after the training stage, we evaluate the performance of these methods using our model. The results are provided in Table 3. Our model performs better than other post-hoc methods, but using the LogitNorm to detect OOD outperforms them. For example, OOD detection performance of our model significantly outperformed the OOD detection performance of baseline models. Moreover, OOD detection performance using LogitNorm outperformed OOD detection performance of MSP, ODIN, and energy methods.

We argue that our model works well with the LogitNorm because we train the model to produce high LogitNorm for ID and low LogitNorm for jigsaw puzzles, which do not have spatial relationship. Nevertheless, since producing low LogitNorm can be interpreted as generating low confidence, other methods that use confidence or logit, may benefit from our training framework.

6.3. Results in other datasets

To validate our method in other datasets, we use ImageNet (Deng et al., 2009) and MNIST (Deng, 2012). For the OOD dataset, iNaturalist (Van Horn et al., 2018), SUN (Sun et al., 2016), Places (Zhou et al., 2018), and Textures (Cimpoi et al., 2014) are used for ImageNet benchmark following other research (Li et al., 2023; Sun et al., 2022). Also, Fashion-MNIST (Xiao et al., 2017) and Kuzushiji-MNIST (Clanuwat et al., 2018) are used as OOD dataset for MNIST. Specifically, we use the same setting for the CIFAR benchmark for both dataset except ViT-small variant and zero weight decay are used for ImageNet and initial learning rate of 0.0003 and weight decay of 1e-04 are used for MNIST. We report the average FPR95 and AUROC for these benchmarks in Table 4. We find that our method shows competitive OOD detection performance in other datasets as well. This demonstrates the generalizability of our method across various datasets.

Table 5
Performance of detecting semantically-shifted OOD. The models are trained on CIFAR10 and evaluated on detecting CIFAR100 as OOD. The best result is indicated in bold. All values are percentages averaged over five runs.

Method	CIFAR100	
	FPR95↓	AUROC↑
Baseline	36.24	93.86
VOS	25.77	94.91
CnC	45.62	87.46
Ours	**19.39**	**95.70**

Table 6
Averaged performance (FPR95↓) of OOD detection using various number of jigsaw puzzles. The best result and the second-best result are indicated in bold and underline, respectively. All values are percentages averaged over three runs.

In-distribution	Number of jigsaw puzzles (n^2)					
	2	3	4	5	7	9
CIFAR10	1.76	<u>0.98</u>	**0.97**	1.12	1.23	2.50
CIFAR100	17.93	**9.10**	9.76	<u>9.73</u>	18.24	21.67

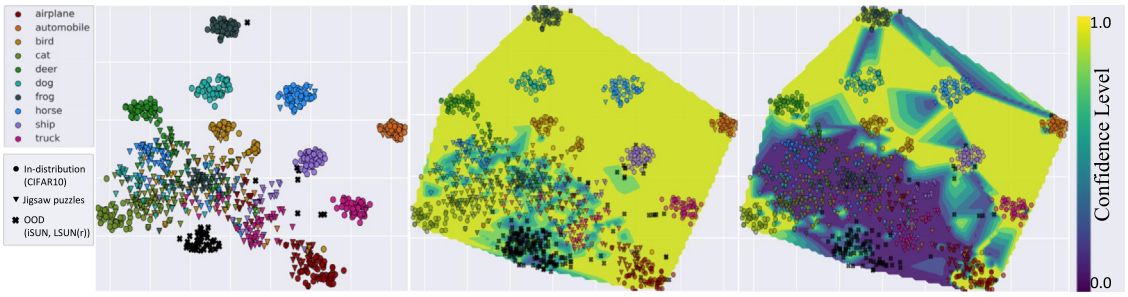
6.4. Detecting semantically-shifted OOD

According to previous research, semantically shifted OOD is the most difficult OOD to detect (Hsu et al., 2020). However, it is more common than non-semantically shifted OOD. Typically, object information, such as a novel object category, is considered the semantic meaning of an image, while non-object information, such as textures or colors, is considered non-semantic. Hence, detecting semantically shifted OOD is challenging but crucial.

To evaluate our framework's ability to detect semantically shifted OOD, we tested its OOD detection performance on CIFAR100 using a CIFAR10 trained network. The results are presented in Table 5. Our framework perform better in detecting semantically and non-semantically shifted OOD. Our results reduced the FPR95 by 25.34% compared to the second-best result (VOS). Since semantically shifted OOD have different spatial relationships compared to ID data, we argue that our training framework can effectively improve detection performance in this area. However, other methods, such as VOS, do not consider spatial relationships and only focus on whether features are outside of the training distribution, whereas CnC method considers images that are corrupted by hard augmentation. Therefore, these methods do not effectively detect semantically shifted OOD.

6.5. Effect of the number of jigsaw puzzles

To demonstrate the effect of jigsaw puzzles numbers, we conducted an experiment to evaluate the OOD detection performance for trained models with various numbers of jigsaw puzzles. In Table 6, we provide the OOD detection performance with different puzzle numbers. We observed that the number of jigsaw puzzle affects the OOD detection performance of the model. Specifically, when the number of jigsaw puzzles is two (i.e., total number is four), we find that the model has low in-distribution accuracy (e.g., 87.19 → 86.45) because the resulting



(a) Feature space (b) Confidence map of baseline (c) Confidence map of ours

Fig. 4. Visualization of the feature distribution (a), confidence map of baseline model (b), and confidence map of our model (c). In feature space (a), Class-wise features are indicated by color, with ID, jigsaw puzzles, and OOD represented by a circle, inverted triangle, and cross shape, respectively. In the confidence map, yellow and dark blue areas represent high and low confidence, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 7
Computational burden for various training methods. The time consumption refers to the training-time for an epoch.

Method	Time consumption	Memory consumption
Baseline	28.90 s	6710 MB
VOS	31.12 s	6750 MB
CnC	260.39 s	11,366 MB
Ours	57.40 s	11,366 MB

jigsaw puzzles are close to original samples and fail to divide the object in image sometimes. Also, large number of jigsaw puzzles drops the OOD detection performance. We argue that this is because the jigsaw puzzles with too much destroyed semantic information do not work well for proxy OOD. For example, the model can only focus on the dividing edge of the jigsaw puzzles rather than the spatial relationship. Thus, we believe that 3, 4 or 5 for jigsaw puzzles number is more appropriate for generating proxy OOD.

6.6. Jigsaw puzzles in feature space

To illustrate the working mechanism of our framework, we provide t-SNE (van der Maaten and Hinton, 2008) visualization of the feature distribution of ID, jigsaw puzzles, and OOD in Fig. 4. Specifically, we visualize the following: (i) the feature produced from the last attention block of ViT for the given ID, jigsaw puzzles, and OOD images (Fig. 4-a); (ii) we visualize the confidences produced from baseline model which trained by cross-entropy loss (Fig. 4-b); (iii) our model trained by the proposed framework (Fig. 4-c). In these visualizations, confidence refers to the MSP for the given images.

Fig. 4-a shows that the ID features are clustered around the same class features, while OOD features are outside the class-wise decision boundary. Jigsaw puzzles are located between ID and OOD features. Subsequently, as shown in Fig. 4-(b,c), baseline and proposed models produce high confidence on ID images since it is trained to classify them. However, our model produces low confidence on images with different spatial relationships compared to the one of ID (i.e., jigsaw puzzles and OOD images), while the baseline model produces relatively high confidence. Our model generates low confidence for jigsaw puzzles and OOD images is because it is trained to produce a low LogitNorm for jigsaw puzzles used as proxy OOD. Using jigsaw puzzles aids the model in more precisely learning the decision boundary for the ID class, resulting in the model generating a low logitnorm for OOD images.

6.7. Limitations

We provide the computational burden for various training methods in Table 7. The experiments are conducted using Pytorch with Nvidia RTX4090. The time consumption and memory consumption are calculated for one Epoch. Time consumption refers to the elapsed time to train the ViT-tiny model for 50,000 CIFAR100 images using batch size of 128. Also, Memory consumption represents the allocated GPU memory for a given batch (size of 128). We find that our method require twice as much time for training and GPU memory since it process the jigsaw puzzle as well as original training images. However, it is much faster than CnC since it requires much more time for processing hard augmentations to create a proxy OOD.

Also, another observation is that our method does not achieve state-of-the-art performance in the ResNet18 architecture. Specifically, we use an ImageNet-pretrained ResNet18 for the CIFAR100 benchmark, using the same settings reported in the experiment section. We observed that our method improves OOD detection performance (FPR95 is reduced from 38.35 to 14.47 and AUROC is 92.94). However, when we utilize the CnC method with ResNet, the FPR95 is 20.66 and AUROC is 95.85. This demonstrates that the CnC method and our method have similar performances for the ResNet architecture despite that our method outperforms CnC when using ViT. We argue that this is because the CNN lacks the ability to connect spatial relationships between image patches, unlike ViT.

7. Conclusion

In this study, we introduce a novel training framework, a simple yet effective approach for improving OOD detection performance by teaching the network the spatial relationship of ID objects. Specifically, we propose L_2-norm loss that forces the network to have a low norm of the logit for given jigsaw puzzles generated from training images, while training to classify the category of training images using cross-entropy loss. We compare our framework with other pre-hoc OOD detection methods and demonstrate that it outperforms other methods on CIFAR benchmarks. Specifically, our method achieves the best average performance on CIFAR10 and CIFAR100. It reduces the average FPR95 by **62.92%** and **25.52%** compared to the second best results on CIFAR10 and CIFAR100, respectively. Moreover, extensive experiments demonstrated that our framework does not decrease the ID accuracy and benefits of detecting semantically shifted OOD. Finally, we provided the visualization of ID, jigsaw puzzles, and OOD features to explain the working mechanism of our framework. We hope our study will benefit other OOD detection methods considering spatial relationships for OOD detection.

CRediT authorship contribution statement

Yeonguk Yu: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Sungho Shin:** Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Minhwan Ko:** Investigation, Validation, Writing – original draft. **Kyoobin Lee:** Funding acquisition, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Kyoobin Lee reports financial support was provided by Korea Ministry of Science and ICT.

Data availability

I have shared the link to code in the paper.

Acknowledgment

This work was partially supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00951, Development of Uncertainty-Aware Agents Learning by Asking Questions) and by ICT R&D program of MSIT/IITP[2020-0-00857, Development of Cloud Robot Intelligence Augmentation, Sharing and Framework Technology to Integrate and Enhance the Intelligence of Multiple Robots].

References

Almohsen, R., Patel, S., Adjeroh, D.A., Doretto, G., 2023. A robust likelihood model for novelty detection. arXiv preprint arXiv:2306.03331.

Chen, P., Liu, S., Jia, J., 2021. Jigsaw clustering for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11526–11535.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A., 2014. Describing textures in the wild. In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., Ha, D., 2018. Deep learning for classical japanese literature. arXiv preprint arXiv:1812.01718.

Deng, L., 2012. The mnist database of handwritten digit images for machine learning research. IEEE Signal Process. Mag. 29 (6), 141–142.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: CVPR09.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR.

Du, X., Wang, Z., Cai, M., Li, Y., 2022. VOS: Learning what you don't know by virtual outlier synthesis. In: Proceedings of the International Conference on Learning Representations.

Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., Socher, R., 2021. Deep learning-enabled medical computer vision. NPJ Digit. Med. 4 (1), 1–9.

Hebbalaguppe, R., Ghosal, S.S., Prakash, J., Khadilkar, H., Arora, C., 2023. A novel data augmentation technique for out-of-distribution sample detection using compounded corruptions. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part III. Springer, pp. 529–545.

Hendrycks, D., Gimpel, K., 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: Proceedings of International Conference on Learning Representations.

Hendrycks, D., Mazeika, M., Dietterich, T., 2019. Deep anomaly detection with outlier exposure. In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=HyxCxhRcY7.

Hsu, Y.-C., Shen, Y., Jin, H., Kira, Z., 2020. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10951–10960.

Krizhevsky, A., Hinton, G., et al., 2009. Learning Multiple Layers of Features from Tiny Images. Citeseer.

Li, J., Chen, P., He, Z., Yu, S., Liu, S., Jia, J., 2023. Rethinking out-of-distribution (OOD) detection: Masked image modeling is all you need. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11578–11589.

Liang, S., Li, Y., Srikant, R., 2018. Enhancing the reliability of out-of-distribution image detection in neural networks. In: 6th International Conference on Learning Representations, ICLR 2018.

Liu, W., Wang, X., Owens, J., Li, Y., 2020. Energy-based out-of-distribution detection. Adv. Neural Inf. Process. Syst..

Lo, S.-Y., Oza, P., Patel, V.M., 2023. Adversarially robust one-class novelty detection. IEEE Trans. Pattern Anal. Mach. Intell. 45 (4), 4167–4179. http://dx.doi.org/10.1109/TPAMI.2022.3189638.

van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (86), 2579–2605, URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

Macêdo, D., Ren, T.I., Zanchettin, C., Oliveira, A.L.I., Ludermir, T., 2021. Entropic out-of-distribution detection: Seamless detection of unknown examples. IEEE Trans. Neural Netw. Learn. Syst. 1–15. http://dx.doi.org/10.1109/TNNLS.2021.3112897.

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

Noroozi, M., Favaro, P., 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In: European Conference on Computer Vision. Springer, pp. 69–84.

Paumard, M.-M., Picard, D., Tabia, H., 2020. Deepzzle: Solving visual jigsaw puzzles with deep learning and shortest path optimization. IEEE Trans. Image Process. 29, 3569–3581.

Salehi, M., Arya, A., Pajoum, B., Otoofi, M., Shaeiri, A., Rohban, M.H., Rabiee, H.R., 2021. Arae: Adversarially robust training of autoencoders improves novelty detection. Neural Netw. 144, 726–736.

Salehi, M., Eftekhar, A., Sadjadi, N., Rohban, M.H., Rabiee, H.R., 2020. Puzzle-AE: Novelty detection in images through solving puzzles. arXiv:2008.12959.

Sun, Y., Guo, C., Li, Y., 2021. ReAct: Out-of-distribution detection with rectified activations. In: Advances in Neural Information Processing Systems.

Sun, Y., Ming, Y., Zhu, X., Li, Y., 2022. Out-of-distribution detection with deep nearest neighbors. In: International Conference on Machine Learning. PMLR, pp. 20827–20840.

Sun, X., Yang, J., Sun, M., Wang, K., 2016. A benchmark for automatic visual classification of clinical skin disease images. In: European Conference on Computer Vision. pp. 206–222.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S., 2018. The iNaturalist species classification and detection dataset. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8769–8778. http://dx.doi.org/10.1109/CVPR.2018.00914.

Wei, H., Xie, R., Cheng, H., Feng, L., An, B., Li, Y., 2022. Mitigating neural network overconfidence with logit normalization. In: International Conference on Machine Learning. ICML.

Wei, C., Xie, L., Ren, X., Xia, Y., Su, C., Liu, J., Tian, Q., Yuille, A.L., 2019. Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1910–1919.

Wightman, R., 2019. PyTorch image models. http://dx.doi.org/10.5281/zenodo.4414861, https://github.com/rwightman/pytorch-image-models.

Xiao, H., Rasul, K., Vollgraf, R., 2017. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv:cs.LG/1708.07747.

Xu, P., Ehinger, K.A., Zhang, Y., Finkelstein, A., Kulkarni, S.R., Xiao, J., 2015. TurkerGaze: Crowdsourcing saliency with webcam based eye tracking. http://dx.doi.org/10.48550/ARXIV.1504.06755, URL: https://arxiv.org/abs/1504.06755.

Xu, K., Rui, L., Li, Y., Gu, L., 2020. Feature normalized knowledge distillation for image classification. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (Eds.), Computer Vision – ECCV 2020. Springer International Publishing, Cham, pp. 664–680.

Yu, Y., Shin, S., Lee, S., Jun, C., Lee, K., 2022. Block selection method for using feature norm in out-of-distribution detection. http://dx.doi.org/10.48550/ARXIV.2212.02295, URL: https://arxiv.org/abs/2212.02295.

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

Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y., 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6023–6032.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A., 2018. Places: A 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. 40 (6), 1452–1464. http://dx.doi.org/10.1109/TPAMI.2017.2723009.