Block Selection Method for Using Feature Norm in Out-of-Distribution Detection

Yeonguk Yu, Sungho Shin, Seongju Lee, Changhyun Jun, and Kyoobin Lee
Gwangju Institute of Science and Technology
{yeon_guk, hogili89, lsj2121, junch9634}@gm.gist.ac.kr, kyoobinlee@gist.ac.kr

Abstract

Detecting out-of-distribution (OOD) inputs during the inference stage is crucial for deploying neural networks in the real world. Previous methods typically relied on the highly activated feature map outputted by the network. In this study, we revealed that the norm of the feature map obtained from a block other than the last block can serve as a better indicator for OOD detection. To leverage this insight, we propose a simple framework that comprises two metrics: FeatureNorm, which computes the norm of the feature map, and NormRatio, which calculates the ratio of FeatureNorm for ID and OOD samples to evaluate the OOD detection performance of each block. To identify the block that provides the largest difference between FeatureNorm of ID and FeatureNorm of OOD, we create jigsaw puzzles as pseudo OOD from ID training samples and compute NormRatio, selecting the block with the highest value. After identifying the suitable block, OOD detection using FeatureNorm outperforms other methods by reducing FPR95 by up to 52.77% on CIFAR10 benchmark and up to 48.53% on ImageNet benchmark. We demonstrate that our framework can generalize to various architectures and highlight the significance of block selection, which can also improve previous OOD detection methods. Our code is available at https://github.com/gistailab/block-selection-for-OOD-detection.

1. Introduction

Neural networks have widely been utilized in the real world, such as in autonomous cars [9, 21] and medical diagnoses [7,38]. In the real world, neural networks often encounter previously unseen input that are different from the training data. If the system fails to recognize those input as unknown input, there can be a dangerous consequence. For example, a medical diagnosis system may recognize an unseen disease image as one of the known diseases. This gives rise to the importance of the out-of-distribution (OOD) detection, which makes users operate a neural network system more safely in the real world.

In practice, various outputs of the network can be used as an indicator to separate the in-distribution (ID) and out-of-distribution (OOD) data. For instance, output probability [15], calibrated output probability [28], and output energy [30] are used as an indicator. The output of a neural network is commonly calculated using a feature vector of the feature extractor and a weight vector of the classification layer. It is known that the norm of the feature vector can be an indicator of input image quality [23, 37, 40] or level of awareness [46]. Thus, we ask the following question: Can we use the norm of the feature as an indicator to separate ID and OOD?

In this paper, we first reveal the key observation concerning the last block of neural networks sometimes deteriorating owing to the overconfidence issue [10,11]. Empirically, we show that OOD images highly activate filters of the last block (i.e., large norm; see Figure 1, left) on a network trained with CIFAR10 while lowly activate filters of the penultimate block (i.e., small norm; see Figure 1, right). As a result, OOD detection methods that consider overactivated feature [42] and overconfident output [28] have been successful. However, we find that the norm of the feature map for the OOD and ID is quite separable in the penultimate block compared to the last block.

This motivates a simple and effective OOD detection
framework consists of (1) \textit{FeatureNorm}: the norm of the feature map and (2) \textit{NormRatio}: the ratio of \textit{FeatureNorm} for ID and OOD. In the case of a suitable block for OOD detection, the \textit{FeatureNorm} of ID is large, and the \textit{FeatureNorm} of OOD is small since its filters are trained to be activated on ID [1, 52]. Thus, we use \textit{FeatureNorm} of the block that has the largest \textit{NormRatio} as an indicator to separate ID and OOD. Although \textit{NormRatio} directly represents the OOD detection performance of the block, we cannot access to the OOD before deployment. In order to select the block, that provides the largest difference between \textit{FeatureNorm} of ID and \textit{FeatureNorm} of OOD, we create pseudo OOD from ID training samples by generating jigsaw puzzles [36] and calculate \textit{NormRatio} with them, and the block with the largest value is selected without accessing OOD. Subsequently, we calculate \textit{FeatureNorm} of the given test image for OOD detection. The proposed OOD detection framework is shown in Figure 2.

We provide experimental results and extensive analysis of our framework. We conduct experiments on common OOD detection benchmarks and show that our simple framework outperforms previous OOD detection methods. Below, we summarize our key results and contributions:

- We introduce the \textit{FeatureNorm}, a norm of the feature map, and the \textit{NormRatio}, a ratio of \textit{FeatureNorm} for ID and OOD to select a block for OOD detection. To the best of our knowledge, \textit{FeatureNorm} and \textit{NormRatio} are the first to explore and demonstrate the norm of the feature map can be an indicator of OOD detection.

- We extensively evaluate our proposed framework on common benchmarks and establish state-of-the-art performances among post-hoc OOD detection methods. Our framework outperforms the best baseline by reducing FPR95 by up to 52.77% on CIFAR10 benchmark and by up to 48.53% on ImageNet benchmark.

- We provide ablation and analysis that improve understanding of our framework. Our analysis demonstrates an importance of norm from the suitable block, which can improve previous OOD detection methods.

2. Preliminaries

We first describe the general setting of the supervised learning problem for the image classification network. In the general setting, the classification network is trained with cross-entropy loss for the given training dataset \( D_{in} = \{ (x_i, y_i) \}_{i=1}^{I} \), where \( x_i \in \mathbb{R}^{3 \times W \times H} \) is the input RGB image and \( y_i \in \{ 1, 2, ..., K \} \) is the corresponding label with \( K \) class categories. An OOD detection method is considered as post-hoc method if it does not modify anything during the training stage.

Out-of-distribution detection When deploying a network in the real world, a user can trust the network if the network classify known images correctly and detect the OOD image as "unknown". For an OOD detection problem with image classification networks, the given test image \( x \) is considered as an OOD image when \( x \) semantically (e.g., object identity) or non-semantically (e.g., camera settings or style of the image) differs from the images of the \( D_{in} \). The decision of the OOD detection is a binary classification with a scoring function that produce ID-ness for the given image \( x \). The main object of the OOD detection research is to find the scoring function that can effectively separate the ID samples and OOD samples.
Elements of convolutional neural networks

Convolutional neural networks (CNNs) usually consist of a feature extractor and a classification layer. A feature extractor encodes an RGB image to a feature map with $M$ channel $z \in \mathbb{R}^{M \times W \times H}$ using its block, where $W$ and $H$ refer to the width and height of each feature map. Also, a classification layer encodes a feature map $z$ to a logit vector $v$. There have been various CNN architectures such as AlexNet [25], VGG [41], ResNet [12], and MobileNet [17]. In this paper, we consider a block as a set of a single convolutional layer with an activation function in VGG architecture and a block as a residual block in ResNet and MobileNet.

Note that the output logit and the output probability of a CNN are commonly calculated as follows:

$$v_i = W_i \cdot f = \| f \|_2 \| W_i \|_2 \cos(\theta_i),$$

$$p_i = \frac{\exp(v_i)}{\sum_k \exp(v_k)},$$

where $\| \cdot \|_2$, $v_i$, $f$, and $W_i$ denote the $L_2$-norm, $i$-th element of logit $v$, feature vector, and $i$-th class weight, respectively. Also, $\theta_i$ refers to the angle between the feature vector $f$ and $i$-th class weight vector $W_i$. Since the output probability distribution is calculated by applying a softmax function on the logit, larger $L_2$-norm feature and larger $L_2$-norm class weight produce a harder probability distribution [49]. Because the cross-entropy loss forces the network to produce 1.0 probability for all training data, the norm of the feature vector and the norm of class weight are forced to be large.

3. Method

3.1. Overview of OOD detection framework

Our OOD detection framework is based on the idea that the norm of feature map obtained from a suitable block for OOD detection can be a good indicator and the suitable block can be selected with ratio of ID and pseudo OOD which are generated from ID training samples. We illustrate our framework in Figure 2. After the training is done, we select the block by NormRatio for OOD detection (Figure 2; left). Then, we use the norm of the feature map FeatureNorm obtained from the selected feature map for OOD detection during the inference stage (Figure 2; right). Specifically, we first generate the jigsaw puzzle as pseudo OOD from ID training samples and calculate NormRatio of training samples and corresponding pseudo OOD. Since the jigsaw puzzles have destroyed object information, we argue that these images can be considered as OOD that is semantically shifted. Thus, NormRatio of ID training samples and pseudo OOD (i.e., jigsaw puzzle) is suitable for finding the block that produces FeatureNorm that can separate ID and OOD samples during the inference stage. Note that our proposed OOD detection framework does not modify the training stage of the network and once the input image is detected as in-distribution image during the inference stage, we can always obtain original output without having any disadvantage on classification accuracy.

3.2. FeatureNorm: norm of the feature map

We consider FeatureNorm, a norm of the feature map, to be an indicator of the activation level of the block for the given image. In practice, we consider a pre-trained neural network for $K$-category image classification with a feature extractor and a classification layer. We denote by $z \in \mathbb{R}^{M \times W \times H}$ feature map obtained by the block of the feature extractor. The norm of each channel $a_i$ of feature map $z_i \in \mathbb{R}^{1 \times W \times H}$ is calculated as follows:

$$a_i = \sqrt{\frac{W \cdot H}{\sum_w \sum_h \max(z_i(w, h), 0)^2}},$$

where $z_i(w, h)$ is the $w$-th, $h$-th element of the feature map $z_i$. This equation can be interpreted as Frobenius norm of the rectified $z_i$ by ReLU function. We utilize a ReLU function to eliminate the effect of negative elements of feature map, which can be seemed as deactivation of filters. Thus, the $a_i$ represents the level of activation of $i$-th channel for the obtained feature map $z$.

Subsequently, the channel-wise averaged norm of the feature map for the block $B$ is calculated as follows:

$$f_{\text{FeatureNorm}}(x; B) = \frac{1}{M} \sum_{m=1}^{M} a_m,$$  

where $f_{\text{FeatureNorm}}(x; B) \in \mathbb{R}$ is the level of activation for the given image $x$ and the block $B$. During the inference stage, OOD detection using FeatureNorm with a suitable block $B_s$ can be conducted with a threshold $\gamma$:

$$G(x; \theta) = \begin{cases} 
\text{ID} & \text{if } f_{\text{FeatureNorm}}(x; B_s) \geq \gamma \\
\text{OOD} & \text{else,}
\end{cases}$$

where the threshold $\gamma$ is typically chosen so that 95% of ID data is correctly classified as ID (i.e., true positive rate of 95%) and the $\theta$ refers to the neural network.

3.3. NormRatio: measure of block’s suitability

We consider NormRatio, which is the ratio of ID FeatureNorm and pseudo OOD FeatureNorm, to be an indicator of the suitability of the block for OOD detection. NormRatio directly represents suitability of the block for OOD detection since the suitable block will produce large ID FeatureNorm and the small OOD FeatureNorm. In practice, the main problem of the NormRatio for selecting a block is that we cannot access the OOD before deployment. Thus,
we need to generate the pseudo OOD, that can represent the OOD may seem during the inference stage, to calculate NormRatio. We argue that using the NormRatio for selecting the block with pseudo OOD that can represent the most hard OOD can achieve the best OOD detection results, and since the semantically shifted OOD images are known to be the hardest OOD to detect [18], we generate the 3 × 3 jigsaw puzzle, which is semantically shifted, as done in [36] using training samples. Our selection algorithm using NormRatio is described as Algorithm 1.

**Algorithm 1 Block selection using NormRatio**

<table>
<thead>
<tr>
<th>Input:</th>
<th>Block list {B_1, ..., B_N}, training data {x_i, y_i}_{i=1}</th>
</tr>
</thead>
<tbody>
<tr>
<td>while</td>
<td>(B_n \in {B_1, ..., B_N}) do</td>
</tr>
<tr>
<td>while</td>
<td>(i \in I) do</td>
</tr>
<tr>
<td>(\uparrow\text{Create jigsaw image:} )</td>
<td>(\hat{x}_i \leftarrow \text{Jigsaw}(x_i))</td>
</tr>
<tr>
<td>(\uparrow\text{Compute NormRatio:} )</td>
<td>(R_{n,i} = \frac{\text{FeatureNorm}(\hat{x}_i ; B_n)}{\text{FeatureNorm}(x_i ; B_n)})</td>
</tr>
<tr>
<td>end while</td>
<td></td>
</tr>
<tr>
<td>end while</td>
<td>(R_n \leftarrow \frac{1}{N} \sum_i R_{n,i})</td>
</tr>
<tr>
<td>(s \leftarrow \arg \max_n (R_n))</td>
<td></td>
</tr>
<tr>
<td>Return</td>
<td>(B_s)</td>
</tr>
</tbody>
</table>

**4. Experiments**

**Setup** We use commonly utilized CNN architectures: ResNet18 [12], VGG11 [41] and WideResNet [55] with depth 28 and width 10 (WRN28) for the CIFAR10 [24] benchmark. The ResNet18 and VGG11 are trained with batch size 128 for 100 epochs with weight decay of 0.0005. The WRN28 is trained with batch size 128 for 200 epochs with weight decay of 0.0005. In all training, the SGD optimizer with momentum 0.9 and initial learning rate of 0.1 is utilized, except VGG11 use initial learning rate of 0.05. The learning rate is decreased by a factor of 10 at 50, 75, 90 training epochs for ResNet18 and VGG11, and at 100, 150 training epochs for WRN28. Also, we use pretrained ResNet50 [12], VGG16 [41], and MobileNetV3_large [17] (MobileNetV3) architectures provided by Pytorch for ImageNet [4] benchmark. At test time, all images are resize to 32 × 32 for CIFAR10 networks, and to 256 × 256 and center crop to size of 224 × 224 for ImageNet networks. We use SVHN [34], Textures [2], LSUN-crop [53](LSUN(c)), LSUN-resize [53](LSUN(r)), iSUN [50], and Places365 [56] as OOD datasets for CIFAR10 benchmark. We use iNaturalist [45], SUN [48], PLACES [56], and Textures [2], which are sampled by Huang et al. [19], as OOD dataset for the ImageNet benchmark.

**Comparison with previous methods** We compare our framework with other post-hoc OOD detection method which calculate OOD score from a model trained on ID data using cross-entropy loss. Although, ODIN [28] requires validation set of ID and OOD for hyperparameter setting, we set these hyperparameters without OOD data as in [18] for fair comparison without access to OOD. As a result, we compare our method with Maximum Softmax Probability (MSP; [15]), ODIN [28], Energy [30], Energy+ReAct [42], and Energy+DICE [43].

**Evaluation Metrics** We measure the quality of OOD detection using the two most widely adopted metrics in OOD detection researches: which are (1) area under the receiver operating characteristic curve (AUROC; [3,8]) and (2) false positive rate at 95% true positive rate (FPR95; [28]). AUROC plots the true positive rate of ID data against false positive rate of OOD data by varying a OOD detection threshold. Thus, it can represent the probability that ID samples will have a higher score than OOD samples. FPR95 is a false positive rate at threshold is set to produce 95% true positive rate. Therefore, it can represent OOD detection performance when an application requirement is recall of 95%. In summary, the higher AUROC and the lower FPR95 represent the better quality of the OOD detection method.

**Block selection for OOD detection using NormRatio** we evaluate OOD detection performance of proposed FeatureNorm with the block selection using NormRatio of ID and pseudo OOD. Pseudo OOD are 3 × 3 jigsaw puzzles generated from each ID training sample. Once the training of the network is done, we calculate the NormRatio for every block using Algorithm 1. Since the various architectures are used for the experiments, we summarize results of block selection in Table 1. We find that our method choose the block for each architecture consistently. For instance, the Block 4.1, Block 3.3 and Layer 7 are selected for ResNet18, WRN28 and VGG11 in all five trials.

<table>
<thead>
<tr>
<th>ID</th>
<th>Architecture</th>
<th>Selected Block Name</th>
<th>Output Size</th>
<th>Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet18</td>
<td>Block 4.1</td>
<td>512 × 4 × 4</td>
<td>N-1</td>
<td></td>
</tr>
<tr>
<td>CIFAR10</td>
<td>WRN28</td>
<td>Block 3.3</td>
<td>640 × 8 × 8</td>
<td>N-1</td>
</tr>
<tr>
<td>VGG11</td>
<td>Layer 7</td>
<td>512 × 4 × 4</td>
<td>N-2</td>
<td></td>
</tr>
<tr>
<td>ResNet50</td>
<td>Block 4.2</td>
<td>2048 × 7 × 7</td>
<td>N-1</td>
<td></td>
</tr>
<tr>
<td>ImageNet</td>
<td>VGG16</td>
<td>Layer 13</td>
<td>512 × 14 × 14</td>
<td>N</td>
</tr>
<tr>
<td>MobileNetV3</td>
<td>Block 17</td>
<td>960 × 7 × 7</td>
<td>N</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Summary of the selected blocks for each architecture. Depth N represents the last block, while Depth 1 represents the first block.
5. Results

5.1. Result on CIFAR10 benchmark

In Table 2, we report the performance of OOD detection for ResNet18, WRN28, and VGG11 architectures using various post-hoc detection methods. The performance are calculated using FPR95 and AUROC on six OOD datasets. Our proposed method achieved the best average performance on both ResNet18 and WRN28, and best FPR95 on VGG11. Note that our method reduces the average FPR95 by 13.45%, 52.77%, and 15.33% compared to the second best results on ResNet18, WRN28, and VGG11, respectively.

As shown in Table 2, our method consistently outperforms other methods on three OOD datasets: SVHN, Textures, and LSUN(c). Also, we find that our method is weaker to LSUN(r), iSUN, and Places365. As we expect, the method is stronger to detect images from SVHN, Textures and LSUN(c) since its image has low complexity compared to CIFAR10 [29] and activation of the image cumulatively differs from the early stage of the network to later stage (see 6.2). In contrast, LSUN(r) and iSUN have large complexity [29], which makes its activation large on the shallow layer and to be detected more easily when using a deeper architecture (i.e., WRN28). Finally, Places365 has similar complexity as CIFAR10 [29] which can be interpreted as images from Places365 have similar low-level abstraction information and semantically shifted compared to ID (i.e., semantically shifted OOD [18]).

5.2. Result on ImageNet benchmark

In Table 3, we report the performance of OOD detection for ResNet50, VGG16, and MobileNetV3 architectures. The performance are calculated using FPR95 and AUROC on four datasets. Our proposed method achieved the best averaged performance on VGG16 and MobileNetV3 architectures. Note that our method reduces the FPR95 on ImageNet benchmark by 48.53% and 27.84% compared to the second best results when using VGG16 and MobileNetV3 architectures. In contrast, we find that our method does not effective on ResNet50 architecture compared to the other methods. We argue that the block structure with batch normalization layer of ResNet reduced the separation gap between ID and OOD samples (see 6.4).

Note that our method consistently outperforms other methods on detecting Textures dataset. We argue that images of Textures are far-OOD [51] and have a lot of low-level complexity images compared to the other OOD dataset, and the activation cumulatively differs from the early stage of the network to later stage (see 6.2). Also, iNaturalist usually have higher complexity images compared to ImageNet. As a result, deep networks can detect the iNaturalist as OOD unlike VGG11 and ResNet18 on CIFAR10 benchmark. Finally, SUN and PLACES have similar complexity level compared to ImageNet, which means that OOD images is semantically shifted and hard to be detected [18].

6. Discussion

6.1. Effect of NormRatio

We calculate the NormRatio using the ID training samples and the pseudo OOD, which is a jigsaw puzzle generated from the ID image, to select the suitable block for OOD detection. Our insight is that the block that has the largest NormRatio is the suitable for detecting pseudo OOD (i.e., jigsaw puzzle) and the other OOD can be seen during the inference stage as well. To find out that the NormRatio can
represent the OOD detection performance of the block, we calculate the NormRatio and OOD detection performance for each block. In Figure 3, we show that the best OOD detection performance can be achieved by the block that produces the largest NormRatio. Also, in Figure 4, we show that FeatureNorm for the given ID (black), pseudo OOD (gray) and various OOD (SVHN: red, Places: blue, LSUN(r): orange). We see that the FeatureNorm of pseudo OOD (gray) acts as the FeatureNorm of OOD images that have enough low-level abstraction (blue, orange).

### 6.2. Effect of FeatureNorm

To demonstrate the effectiveness of FeatureNorm as an indicator for OOD detection, we show the change of FeatureNorm of various input through blocks in Figure 4. In Figure 4, we demonstrate that norm of the OOD images with low-complexity (SVHN) is low consistently on all blocks except the last one. On the other way, FeatureNorm of the OOD images with high complexity (LSUN(r)) is higher than ID in shallow blocks since the shallow block of network act as edge detector [1, 52] and the high level complexity image tends to have large low-level abstraction. Also, high complexity images obtain large FeatureNorm in shallow blocks, the FeatureNorm is reduced since it cannot activate deeper blocks that act as high-level abstraction detector. We argue that network like VGG11 or ResNet18 is hard to separate ID and OOD with high-level complexity since it has a few deep blocks compared to WRN28. As a result, the low quality ID image obtains low FeatureNorm and the high quality ID image or OOD image with high-level semantic information obtains high FeatureNorm as shown in Figure 5.

### 6.3. Output calibration using the selected block

The output of the network is deteriorated when the last block of the network produces a high norm for both ID and OOD, and we argue that the network output can be calibrated by replacing the norm of the last block with the norm of the selected block. We evaluate the OOD detection method on ResNet18, WRN28, and VGG11, which are suffered by overconfident last block, with and without replacing. In the Table 4, we show that the existing OOD
Figure 4. The FeatureNorm calculated using every block for the given ID (black), Pseudo OOD (gray), and OOD images (SVHN: red, Places: blue, LSUN(r): orange). We show that the high complexity image (LSUN(r)) activates shallow block and low complexity image (SVHN) does not activate the block except the last one. We find that the jigsaw puzzle image represents the high complexity OOD image with no semantic information.

Figure 5. The example results of our proposed OOD detection framework with a VGG16 architecture for ImageNet. We demonstrate that the low-complexity images obtain low FeatureNorm, while high-complexity images obtain high FeatureNorm. Note that, some OOD images actually contain ID objects (e.g., tench, king penguin, honeycomb).

detection methods that utilize the output of the network can be improved by selecting a suitable norm for the OOD detection. This suggests that the current network is underperforming in OOD detection by the overconfident last block, and can be calibrated by the block selection.

6.4. Structure of the block cause the overconfidence

While our method successfully utilizes FeatureNorm and NormRatio for OOD detection, it does not yield the best results for ResNet18 and ResNet50 architectures. We hypothesize that the structure of the residual block is the reason for this. We demonstrate that the order of layers in the block can decrease the OOD detection performance of FeatureNorm in Table 5. The order of the layers in the ResNet [12] block is Conv-BN-ReLU, while the order is BN-ReLU-Conv(a.k.a. PreAct [13]) in WRN [55] for faster training and better accuracy. There have been reports that the Batch Normalization [20] cause the overconfidence [10] and over-

Table 4. An ablation study with and without selected norm for other previous OOD detection methods. We demonstrate that replacing the norm of the last block with the norm of the selected block improve the performance of previous methods. The results are averaged over five runs.

Table 5. OOD detection performance comparison between two ResNet18 architectures with different block orders. Conv-BN-ReLU refers to the basic block order of ResNet18 and BN-ReLU-Conv refers to the block order of WRN28, which we argue that better block order for OOD detection. It demonstrates that the BN-ReLU-Conv block order outperforms the original block order in detecting high complexity OOD images (LSUN(r), iSUN, Places365). Thus, ResNet18 and ResNet50 cannot fully leverage the proposed framework. The results are averaged over five runs.

7. Related Work

7.1. Out-of-distribution detection

To detect OOD samples during the inference stage with pre-trained network, Hendrycks et al. [15] proposed maximum softmax probability (MSP), based on their observa-
tion that the classifier tends to have lower confidence for the OOD sample than for the ID sample. Similarly, the ODIN [28], which is an enhanced version of the MSP, applies two strategies, namely input preprocessing and temperature scaling, to separate the confidence of ID samples and the one of OOD samples, which leads to improved OOD detection performance. In the other hand, Lee et al. [47] proposed a simple modification to the cross-entropy, which enforces a constant vector norm of the logit in training to prevent the norm of the logit keeps increasing during training. This modification helps prevent the network from becoming overconfident and improves the OOD detection performance. In this paper, we did not compare performance with these methods since our method is a post-hoc method.

7.2. Calibration of neural networks

We prefer a well-calibrated network for deploying the neural network in the real world since a well-calibrated network will produce low confidence for the given unseen input. However, there has been a report that the deep neural network, especially modern architecture, is poorly calibrated [10] due to the overfitting and Batch Normalization. The calibration methods can be categorized into post-hoc methods, which utilize the recalibration function for pre-trained model, and pre-hoc methods, which utilize extra training procedure. The very basic way to improve model calibration using recalibration method is temperature scaling of logit [10]. Also, Gupta et al. [11] proposed to utilize a recalibration function based on splines, which maps the logit to calibrated class probabilities. These methods are performed using a held-out calibration set to calculate hyper-parameter of the calibration method. Calibration methods with training procedure use data augmentation [44, 54] or modify training loss [22, 26, 33]. These calibration methods may be helpful for OOD detection researches since the well-calibrated network should produce low probability for OOD.

7.3. Feature of neural networks

It is known that the norm of the feature vector will be lower for unseen images since the feature extractor is forced to produce higher norm in the training stage [46]. Also, there has been reports that the norm of the feature vectors represented the quality of the input image in the face recognition [23, 39, 40], and the norm of the feature vectors for unknown samples is often lower than those of known [5, 46]. However, it is also known that the network can be easily fooled by unrecognizable images [14, 32, 35] or adversarial attack [6, 31] which means that the filter of the network can be activated by unseen image. Also, the feature vector, which is directly used for calculating logit by dot product, is forced to be large during its training stage by cross-entropy loss [49]. Since our work utilizes the norm of the feature map, we believe that our framework can be improved by considering aforementioned works.

8. Conclusion

In this study, we propose a simple OOD detection framework that consists of (1) FeatureNorm, which is a norm of feature map from the block, and (2) NormRatio, which is a ratio of FeatureNorm for given ID and OOD images. We demonstrate that the suitable block for OOD detection can be selected without access to OOD by choosing the block that produce the largest NormRatio with ID and pseudo OOD that is generated from ID, and using FeatureNorm of suitable block for OOD detection outperforms existing approaches. Extensive experiments show our framework can improve other existing methods as well. We hope our research can help to solve the overconfidence issue of neural networks.

Acknowledgement

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


[34] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 427–436, 2015. 8


[43] Yiyao Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In European Conference on Computer Vision, 2022. 4, 5, 6, 8


[45] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The unnatural species classification and detection dataset, 2017. 4

[46] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In International Conference on Learning Representations, 2022. 1, 8

[47] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. 2022. 8


[54] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF international conference on computer vision, pages 6023–6032, 2019. 8