Robustifying Out-of-Distribution Detection: A Self-Supervision and Energy Based Approach

Aditya Kumar Akash [1] **Sean Chung** [1] **Shri Shruthi Shridhar** [1] **Wissam Kontar** [1]

Abstract

Out-of-distribution (OOD) detection is essential to deploying machine learning systems in the real world. However, the reliability of the existing OOD detectors is severely hampered when used in an environment with adversarial/natural perturbations. Being such a critical component, this necessitates the study of techniques to robustify it. In this work, we propose using the representation learning power of self-supervision methods with better OOD scoring mechanism based on energy to improve the robustness of OOD detectors. Specifically, we propose a blend of flexible loss function formulations that can effectively learn robust features. Our findings merit the use of a new methodological perspective that focuses on robustifying OOD detection.

1. Introduction

The exciting success of deep machine learning models has made them the de facto choice as the solution to building intelligent systems. The recent progress in their performance has led to an increasing number of real-world applications being powered using deep learning. Some of these areas include autonomous cars, automated facial recognition for security, voice-controlled devices, etc. Many of these applications are critical to influence the lives of people and it is very important that the deployed models are reliable. One important aspect of enforcing reliability (Amodei et al., 2016) is to be able to detect out of distribution (OOD) data and prevent exposing the deep learning models to these. This makes OOD detection very crucial to deploying trustworthy machine learning models.

Even though OOD detection helps in building reliable models, they too are prone to adversarial attacks. Robustness against natural and artificial perturbed inputs is an extremely relevant problem. For instance, an autonomous vehicle could misjudge a road STOP sign and accelerate leading to a fatal car accident at an intersection. A weak facial recognition system could be easily fooled to gain entrance in a secure location. A wealth of recent literature currently exists in the area of building a robust OOD detectors (Hendrycks et al., 2019a; 2020; Chen et al., 2020b; Liang et al., 2017; Lee et al., 2017).

Building robust OOD detector is very challenging since the OOD samples at inference time could be very different from those used during training the model. Hence it is extremely important that the deep learning models learn robust features from the datasets. Recently, self-supervision has been shown to learn meaningful representations (Misra & van der Maaten, 2019) from an unlabelled pool of data and gained popularity for transferring learnt model weights to different datasets (Hendrycks et al., 2019a). Some examples of self-supervised learning tasks that are popular are the prediction of geometric rotations (Gidaris et al., 2018), contrastive learning(Chen et al., 2020c), solving Jigsaw puzzles (Noroozi & Favaro, 2017), etc. Some recent work (Kim et al., 2020) has shown the applicability of self-supervision in learning models that have some degree of robustness against perturbations in the input sample. Overall, self-supervision losses provide an exciting opportunity to learn meaningful features from the input and hence fit in building a robust OOD.

Finally, a good OOD detector should separate the in-distribution samples from out-of-distribution samples. Hence it is important to use a good scoring function. Recent work in OOD detection (Liu et al., 2020a) has shown the usefulness of energy score as a better OOD score. The authors show that energy score is linearly related to the logarithm of the probability density of the samples and propose using energy based regularizer to further improve the OOD detector.

In this work, we propose to improve the robustness of OOD detection by augmenting it with popular self-supervision frameworks and using more meaningful OOD scoring function. We specifically explore the robustness property of con-

[1]Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI, USA. Correspondence to: Aditya Kumar Akash <aka@cs.wisc.edu>, Sean Chung <sean.chung@wisc.edu>, Shri Shruthi Shridhar <shridhar2@wisc.edu>, Wissam Kontar <kontar@wisc.edu>.

trastive losses like SimCLR and other self-supervision tasks like predicting geometric rotations. The idea is to extend the OOD detection learning with self-supervised component and further improve it using energy based scores. Some existing work has proposed solutions for adversarial robustness of self-supervision against perturbations. In the current work, we explore the transferability of robustness to a completely unknown dataset (OOD setting) under a variety of attacks ranging from natural OOD, natural corruptions to compositional attacks which are harder to detect.

2. Related Work

Several works on out-of-distribution (OOD) detection has been proposed. OOD detection refers to models that are able to distinguish OOD samples – which are deviates from in-distribution (ID) samples. From a traditional machine learning methods point of view, it is assumed that training data and testing data are independently identically distributed. Nevertheless, in the real world, it is hardly possible to assure that data fed into deployed models is always in-distribution. That is, it is also likely to be OOD (i.e., outlier). A deep learning model without an outlier detector can easily misrecognize an OOD sample as one of the classes from the ID samples – this is not reasonable. Therefore, it is imperative to build a model that is able to detect OOD data. The propelling motivation behind OOD detection, is that previous work has shown that neural network can produce predictions with large confidence for OOD inputs (Hendrycks & Gimpel, 2016; Lakshminarayanan et al., 2017).

Robust Out of Distribution Detection. (Chen et al., 2020b) propose a novel robust OOD detection method – Adversarial Training with Informative Outlier Mining (ATOM) that carefully samples outliers data for training. ATOM has shown to improve robustness and generalization to adversarial attacks such as clean and perturbed OOD inputs.

Self-Supervised Learning with Rotation. As predicting rotation requires modeling shape, and knowing that a smaller region of an image alone might not be sufficient for deciding if the image is flipped, training with self-supervised auxiliary rotations may improve robustness. (Gidaris et al., 2018) predict image rotations by training models to recognize 2d transformations. (Hendrycks et al., 2019a) show that by applying rotations to inputs, self-supervision can improve robustness to adversarial examples, label corruption, common input corruptions, and out-of-distribution detection on difficult and near-distribution outliers.

Contrastive Learning. Proposed by (Chen et al., 2020d), SimCLR is a framework for contrastive learning of visual representations. It maximizes agreement between 2 augmented (e.g., rotation, crop, resize, etc) versions of the same image. This naturally takes outliers into consideration.

Energy-based Out-of-distribution Detection. (Liu et al., 2020b) focus on energy scores instead of softmax confidence since it is biased and not aligned with a density of inputs. Energy-based OOD detection is useful for information mining and it can improve models by being less susceptible to overconfidence, and superior to softmax confidence score. Energy method gain their

Our Contributions. The main contribution of this work involves (a) We propose using self-supervision based losses to improve robustness of OOD detection, (b) We extend existing SOTA OOD detection using energy based OOD scores and provide further evidence on the efficacy of energy score, (c) We investigate the robustness of self-supervision losses over a variety of OOD attacks, and (d) We empirically show that SimCLR based losses and using energy score to perform outlier mining outperforms existing methods on hard OOD detection tasks.

3. Background: Robust OOD

We first introduce the problem of Robust out-of-distribution detection. OOD detection is defined with respect to an inlier distribution $p(x)$. Consider learning a classifier $f_\theta(x)$ to predict labels $y \in \{1, 2, ..., k\}$. Essentially f_θ learns the conditional distribution $p(y|x)$. The training dataset, sampled from joint inlier distribution $p(x, y)$, is available for learning parameters of f_θ. During inference time, the data could also come from another distribution $q(x)$ which might be perturbed. The problem of learning a robust OOD detector is defined as learning a function $h(x)$ such that

$$h(x) = \begin{cases} -1; & x \in p(x) \\ 1; & \text{otherwise} \end{cases} \tag{1}$$

OOD Perturbations. We consider perturbations of individual input samples x denoted by $\Omega(x)$. At test time, the OOD detection is evaluated on worst case input in $\Omega(x)$ for OOD samples x coming from an unknown distribution $q(x)$ to which we do not have access to during training. However, we assume that access to an auxiliary dataset $\mathcal{D}_{out}^{aux}$ is provided while training $h(x)$.

Following (Chen et al., 2020a) our detector is evaluated on OOD perturbations of the following types

1. **Natual OOD**: $\Omega(x) = \{x\}$, no perturbations
2. L_∞ **attacked OOD (white-box)**: $\Omega(x) = \{y|||x - y||_\infty \leq \epsilon\}$, attack considers worst case perturbations which has low OOD score for OOD samples
3. **Corruption attacked OOD (black-box)**: Realistic type of attacks which could happen naturally. These are based on common corruptions as mentioned in (Chen et al., 2020a).

4. **Compositionally attacked OOD (white-box)**: This attack considers the composition of L_∞ and corruption attack.

4. Robustness using Self-Supervised Losses

In this section, we first describe the out-of-distribution detection using outlier exposure. Next, we show how self-supervised learning losses can be combined with outlier exposure method to improve the robustness of out-of-distribution detection.

4.1. Outlier Exposure using Adversarial Training

We consider a k way deep network f_θ for the classification of inlier distribution images. The standard training objective for learning this classifier is given by

$$\min_{\theta} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{in}^{train}} \left[\ell(x, y; f_\theta)\right] \tag{2}$$

where ℓ is the cross entropy loss. To perform out-of-distribution detection another head is added to this network making it a $k+1$-way classifier. The $(k+1)^{th}$ class label indicates if the input is an outlier. Similar to (Chen et al., 2020a), we define the outlier exposure objective using adversarial training as

$$\begin{aligned}\min_{\theta} \mathcal{L}_{oe} = &\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{in}^{train}} \left[\ell(x, y; f_\theta)\right] \\ &+ \lambda \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{out}^{train}} \left[\max_{x'\sim\Omega_{\infty,\epsilon}(x)} \ell(x', k+1; f_\theta)\right]\end{aligned} \tag{3}$$

where $\mathcal{D}_{out}^{train}$ is the OOD training dataset. This exposes the standard training to outliers. The inner max in (3) is solved using Projected Gradient Descent (PGD) (Madry et al., 2017) and applied to the half of the minibatch while the other half is not perturbed. This is to ensure good performance on both natural OOD and perturbed OOD.

After training the model f_θ, the inlier classification is done using the first k logit of the network using argmax. The OOD detector is constructed using the $(k+1)^{th}$ logit of the network. An input x is labelled as an outlier if $f_\theta(x) < \gamma$ for some threshold γ which can be selected in a way to ensure that a significant fraction of inliers is correctly classified.

Next, we augment this objective with self-supervision losses to improve the robustness of outlier detection. The proposed framework is shown in Figure 1. The training consists of three components (a) Inlier classification objective (b) Outlier exposure component, and (c) Self-supervision component.

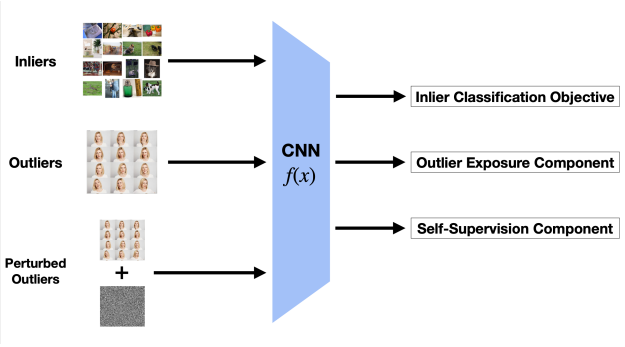

Figure 1. Our proposed framework for training robust OOD detector. During training, it uses inputs from inlier dataset, outliers from an auxiliary dataset, and a perturbed fraction of outliers. The CNN module has three heads for (a) inlier classification, (b) OOD detection, and (c) self-supervised component.

4.2. Robustness using Contrastive Losses

Contrastive losses (Misra & van der Maaten, 2019) have been used in self-supervised learning to learn representations which are useful for downstream tasks. The idea is to maximize the similarity between different augmentations of the same instances and minimize the similarity between representations of different instances. SimCLR (Chen et al., 2020c) is a popular contrastive learning method and has been effectively used to train models in a self-supervised and semi-supervised way. (Kim et al., 2020) show its effectiveness in learning adversarial robust models. We use SimCLR based contrastive losses to improve the robustness of OOD detection under a variety of attack scenarios.

As in SimCLR, we consider a two-layered FC network which outputs the latent representation using the output of the penultimate layer of the classification network f_θ. We use $z = g_\omega(x)$ as the latent representation of input x. Consider $\mathcal{T}$ as the family of augmentations applicable to image x as described in (Chen et al., 2020c). Sampling two random transformations $t, t' \sim \mathcal{T}$ and applying on image x gives us two different views of the data as x_i and x_j. The SimCLR loss is applied on the encoded representation of these two views, i.e $z_i = g_\omega(x_i)$ and $z_j = g_\omega(x_j)$.

For applying SimCLR loss on a minibatch of N examples, pairs of augmented examples are derived from the minibatch producing $2N$ samples. Let $\text{sim}(u, v) = u^T v/||u||||v||$ be the normalized dot product between u, v. The SimCLR loss for positive pair (z_i, z_j) is defined as

$$\ell(i, j) = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{(k\neq i)} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{4}$$

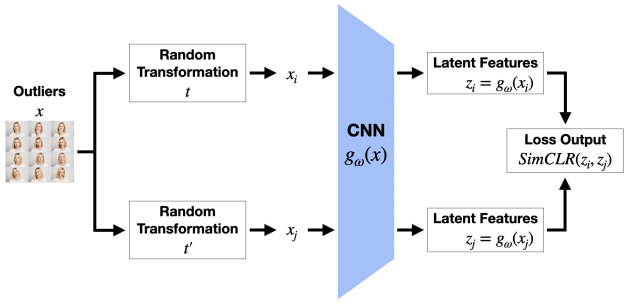

Figure 2. We illustrate the SimCLR head used for self-supervision. This replaces the self-supervision component in Figure 1. SimCLR based constrastive loss is applied on encoded representations z_i, z_j for two different views of the input image x.

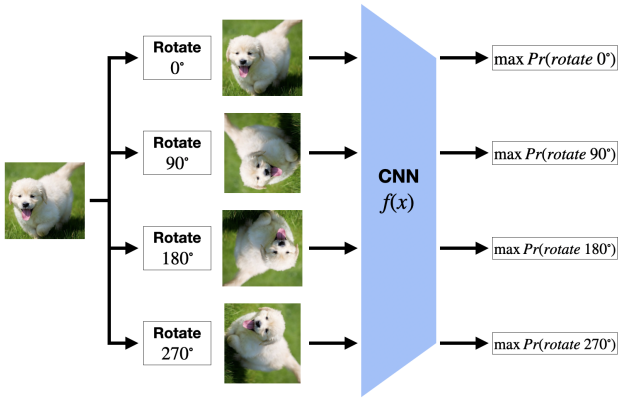

Figure 3. The self-supervised component based on predicting geometric rotation task; Inputs are rotated by $\{0°, 90°, 180°, 270°\}$. The overall loss function is the average of four cross entropy loss for predicting the rotation class.

For a minibatch of size N, the SimCLR loss is given as

$$\mathcal{L}_{simclr} = \frac{1}{2N}\sum_{i=1}^{N}[\ell(2k-1, 2k) + \ell(2k, 2k-1)] \quad (5)$$

Using (3) and (5) gives us the robust training objective as

$$\min_{\theta,\omega} \mathcal{L}_{oe} + \eta\mathcal{L}_{simclr} \quad (6)$$

where the SimCLR is applied on the minibatch samples from the $\mathcal{D}_{out}^{train}$. This setup is depicted in Figure 2.

Next, we describe another self supervised loss that could be used in place of SimCLR for learning robust OOD detector.

4.3. Robustness using Geometric Rotation Prediction

Predicting geometric rotations of images has been successfully used as a self-supervision method to learn useful latent representations. To correctly predict the amount by which the original image has been rotated requires learning global image features. Hence we explore whether learning these features provides robustness from various types of perturbations. (Hendrycks et al., 2019b) provide some evidence in this direction.

Our method consists of adding addition 4 classification head to the original $(k+1)$-way classifier. These additional logits would be used for predicting the degree of rotation of the original image. The images are randomly rotated by $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$. These transformations can be achieved using flips and matrix transpose operations and hence does not introduce any artefacts in the input image. Figure 3 describes the setting.

The loss for predicting rotation is given by

$$\mathcal{L}_{rot} = \sum_{r\in\{0°,90°,180°,270°\}} \ell_{CE}(x, \text{one\_hot}(r); f_\theta^{ss})/4 \quad (7)$$

where f_θ^{ss} is the self-supervised part of the model used for predicting rotations. The robust training objective is given as follows

$$\min_{\theta} \mathcal{L}_{oe} + \eta\mathcal{L}_{rot} \quad (8)$$

where the rotation based loss is applied to all the inliers and outliers samples in the minibatch.

5. Using Energy to improve Robustness

In this section, we describe how we can further improve the robustness of OOD detection using energy based techniques from (Liu et al., 2020a). We first describe the informative outlier mining method (ATOM) introduced in (Chen et al., 2020a) as an improvement over the outlier exposure objective (3). Next, we provide some details on energy based methods which have been shown to perform better for OOD detection and have better theoretical properties. Finally we describe how we can use energy based scores with outlier exposure objective to improve the robustness of OOD detector. The proposed method can be used on top of self-supervision based techniques to improve its robustness.

Informative Outlier Mining: (Chen et al., 2020a) show that if the outliers are adaptively chosen for training in (3), the resulting algorithm (ATOM) leads to a much better decision boundary between inliers and outliers. The main idea is to use the scores from the current model to select a

sample set of outliers. Starting with the q^{th} quantile score in the score sorted list of outliers, n outliers are selected for the training. The value of the $(k+1)^{th}$ logit from f_θ is used as the outlier score.

Energy Score: A natural choice of doing OOD detection would be to use density of the samples $p(x)$. Samples with low density are deemed to be outliers. In (Liu et al., 2020a), authors show that logarithm of $p(x)$ is linearly related with the energy score as in

$$\log p(x) = -E(x; f_\theta) - \log Z \tag{9}$$

where Z is a constant and energy function is expressed as

$$E(x; f) = -T. \log \sum_{i=1}^{k} f_i(x)/T \tag{10}$$

where f_i is the i^{th} logit and T is the temperature parameter. The authors show that using energy score as the score for outliers leads to better separability between the score distribution of outliers and inliers making them a superior choice.

5.1. Informative Outlier Mining using Energy Score

From the above discussion we see that energy score is a natural choice for outlier score and informative outlier mining leads to more robust models. We propose combining these two concepts and use energy score as the outlier score in mining informative outliers. The idea is to sort the outliers using the energy score obtained from the first k logits of the model f_θ. n outliers with lowest energy are selected for training the outlier exposure objective (3). We call this method ETOM and is given in Algorithm 1.

Algorithm 1 ETOM - Energy based ATOM

input $\mathcal{D}_{in}^{train}$, $\mathcal{D}_{out}^{aux}$, n, m, N f_θ
output f_θ
for $t = 1, 2, ..., m$ **do**
  Randomly sample N points from $\mathcal{D}_{out}^{aux}$ to get set S
  Compute OOD scores to get $V = \{E(x; f_\theta) | x \in S\}$
  Sort scores in V ascending
  $\mathcal{D}_{out}^{train} \leftarrow V[0:n]$
  Train f_θ for one epoch using objective (3)
end for
Build f_θ

5.2. Energy Regularized Informative Mining

We also explore the use of energy regularizer from (Liu et al., 2020b) on top of Informative Outlier Mining (ATOM). As described earlier the inliers correspond to lower energies, and the outliers correspond to higher energies. The regularizer tries to penalize low energy of outliers and high energy of inliers. This is expected to create more separation between the score distribution of the inliers and outliers, thus making it easy to classify between them. The energy regularizer is given by following

$$\begin{aligned}\mathcal{L}_{\text{energy}} = \mathbb{E}_{(\mathbf{x}_{\text{in}}, y) \in D_{\text{in}}^{\text{train}}} \big[\max(0, E(\mathbf{x}_{\text{in}} - m_{\text{in}})) \big]^2 \\ + \mathbb{E}_{\mathbf{x}_{\text{out}} \in D_{\text{in}}^{\text{train}}} \big[\max(0, m_{\text{out}} - E(\mathbf{x}_{\text{out}})) \big]^2\end{aligned} \tag{11}$$

We combine this with the objective (3) and propose optimizing following

$$\min_\theta \mathcal{L}_{oe} + \eta \mathcal{L}_{\text{energy}} \tag{12}$$

6. Experiment

We now describe our experimental setup and demonstrate the effectiveness of our proposed methods in building a robust OOD detector. We evaluate the proposed OOD methods against a variety of attacks mentioned in Section 3.

6.1. Setup

Datasets. Following (Chen et al., 2020b), we will use CIFAR-10 as in-distribution datasets and SVHN (Netzer et al., 2011), LSUN (Yu et al., 2015), iSUN (van den Oord et al., 2016), Textures (Cimpoi et al., 2014) and Places365 (Zhou et al., 2018) for our OOD inputs.

Auxiliary Out-of-distribution Datasets. We use ImageNet-RC, a variant of ImageNet (Chrabaszcz et al., 2017) as an alternative auxiliary OOD dataset.

Out-of-distribution Datasets. We consider the robust OOD Evaluation tasks described in (Chen et al., 2020b), namely natural OOD input, OOD input with L_∞ perturbations, corruptions bound OOD input and joint L_∞ perturbations and corruptions bound OOD input.

Hyperparameters. (i) Self-supervision with rotation uses $eta = 0.1$, $epoch = 60$, $batch\ size = 64$ (ii) SimCLR robustness uses $eta = 1$, $epoch = 60$, $batch\ size = 64$, $\tau = 1$ (iii) Energy-based approach uses $m_{\text{in}} = -13$, $m_{\text{out}} = -3$

Metrics. We measure the following metrics: Accuracy, False Positive Rate (FPR) at 5% False Negative Rate, False Positive Rate at 80% True Positive Rate, Area Under the Receiver Operating Characteristics (AUROC), Area Under the Precision Recall Curve.

Baselines. We consider the model learned from optimization 3 as the 'baseline' model in our results. We also compare our results with ATOM and NTOM setup from (Chen et al., 2020b) which are based on a similar setup. To provide more prespective on the importance of our results we include the comparisons with various baselines from (Chen et al., 2020b).

D_{in}^{test}	Method	FPR (5% FNR)	AUROC	FPR (5% FNR)	AUROC	FPR (5% FNR)	AUROC	FPR (5% FNR)	Accuracy
		Natural OOD		Corruption OOD		L_∞ OOD		Comp. OOD	
CIFAR-10	Baseline	11.21	97.34	35.58	92.82	65.30	59.82	59.63	64.71
	w/ Rotation	**3.92**	**99.06**	**32.38**	**94.60**	98.91	15.91	99.19	16.80
	w/ SimCLR	10.69	97.70	33.83	93.57	**21.86**	**94.92**	**37.05**	**92.66**

Table 1. Comparison with Comparison of proposed self-supervision based OOD detection method with baselines. We evaluate on four types of OOD inputs: (1) natural OOD, (2) corruption attacked OOD, (3) L_∞ attacked OOD, and (4) compositionally attacked OOD inputs. The smaller value FPR is, the better; the larger AUROC is, the better. All values are percentages and are averaged over six natural OOD test datasets mentioned in 6.1. **Bold** numbers are superior results

D_{in}^{test}	Method	FPR (5% FNR)	AUROC	FPR (5% FNR)	AUROC	FPR (5% FNR)	AUROC	FPR (5% FNR)	AUROC
		Natural OOD		Corruption OOD		L_∞ OOD		Comp. OOD	
CIFAR-10	NTOM	1.87	99.28	30.58	94.67	99.90	1.22	99.99	0.45
	ATOM	**1.69**	**99.20**	**25.26**	95.29	20.55	88.94	38.89	86.71
	w/ Energy Reg. (Ours)	8.08	98.49	50.45	91.65	8.72	98.26	50.50	91.64
	ETOM (Ours)	7.74	98.59	25.46	**95.61**	**8.13**	**98.48**	**25.42**	**95.60**

Table 2. Comparison with proposed energy based OOD detection method with SOTA methods like ATOM and NTOM. We evaluate four types of OOD inputs: (1) natural OOD, (2) corruption attacked OOD, (3) L_∞ attacked OOD, and (4) compositionally attacked OOD inputs. A smaller FPR value is better, while a larger AUROC is better. All values are percentages and are averaged over six natural OOD test datasets mentioned in 6.1. **Bold** numbers are superior results

D_{in}^{test}	Method	FPR (5% FNR)	AUROC	FPR (5% FNR)	AUROC	FPR (5% FNR)	AUROC	FPR (5% FNR)	AUROC
		Natural OOD		Corruption OOD		L_∞ OOD		Comp. OOD	
CIFAR-10	MSP	50.54	91.79	100.00	58.35	100.00	13.82	100.00	13.67
	ODIN	21.65	94.66	99.37	51.44	99.99	0.18	100.00	0.01
	Mahalanobis	26.95	90.30	91.92	43.94	95.07	12.47	99.88	1.58
	SOFL	2.78	99.04	62.07	88.65	99.98	1.01	100.00	0.76
	OE	3.66	98.82	56.25	90.66	99.94	0.34	99.99	0.16
	ACET	12.28	97.67	66.93	88.43	74.45	78.05	96.88	53.71
	CCU	3.39	98.92	56.76	89.38	99.91	0.35	99.97	0.21
	ROWL	25.03	86.96	94.34	52.31	99.98	49.94	100.00	49.48

Table 3. Evaluation results for competitive OOD detection methods on four types of OOD inputs: (1) natural OOD, (2) corruption attacked OOD, (3) L_∞ attacked OOD, and (4) compositionally attacked OOD inputs. The smaller value FPR is, the better; the larger AUROC is, the better. All values are percentages and are averaged over six natural OOD test datasets mentioned in 6.1

6.2. Results

Self-Supervision to improve Robustness. Table 1 provides the comparison of performance of methods based on self-supervision and outlier exposure baseline. We see that the self-supervision based on rotation provides good robustness against normal OOD. However when the detector is attacked using perturbations in L_∞ norm, the method fails to provide any robustness. Hence the behavior for L_∞ OOD and compositional OOD. With SimCLR based robust formulation, we find that it outperforms the baseline on all the OODs. This further provides evidence that the representations learnt using SimCLR are more meaningful in comparison to rotation based self-supervision. For L_∞ OOD and compositional OOD, it performs at par with ATOM which is very encouraging since the model is not trained for these attacks.

Robustness using informative mining with Energy Scores. Table 2 shows the performance of our proposed energy based informative outlier mining training method (ETOM). ETOM performs at par with ATOM on corruption OOD while it significantly outperforms ATOM on L_∞ OOD and compositional OOD. This solidifies the use of energy score as OOD score. It is quite promising since we can now work with quantile based ETOM and analyze its influence on robustness.

Robustness with ATOM + Energy Regularizer. From Table 2, we find that the energy regularized ATOM does better at L_∞ OOD samples. However, it underperforms for other OOD attacks in comparison to ATOM. We are still investigating this behavior.

We also include Table 3 to illustrate a comparison of our proposed method with state of the art OOD detection methods. We find that the proposed method of using SimCLR and energy based informative outlier mining provide promising results and improve the robustness for OOD detection.

7. Conclusion

In this work, we explore improving the robustness of OOD detection using motivations from self-supervision techniques and energy based scores. Our proposed method of using constrastive learning loss (SimCLR) and using Energy based informative outlier gives very promising results and outperforms the existing methods for harder outliers. These seem to be potential directions that can be explored in detail and combined under one framework to build SOTA robust OOD detector.

Future Work. Inspired by the results from our experiments, we would like to investigate the behavior of ETOM under quantile based mining. Another promising direction is understanding properties of Energy function in the current setup which leads to its superior performance. We will also consider how to jointly use all the proposed techniques to build a superior robust OOD detection method.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in ai safety, 2016.

Chen, J., Li, Y., Wu, X., Liang, Y., and Jha, S. Informative outlier matters: Robustifying out-of-distribution detection using outlier mining, 2020a.

Chen, J., Li, Y., Wu, X., Liang, Y., and Jha, S. Robust out-of-distribution detection via informative outlier mining. *arXiv preprint arXiv:2006.15207*, 2020b.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020c.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. A simple framework for contrastive learning of visual representations. In *ICML*, 2020d.

Chrabaszcz, P., Loshchilov, I., and Hutter, F. A downsampled variant of imagenet as an alternative to the cifar datasets. *ArXiv*, abs/1707.08819, 2017.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pp. 3606–3613, USA, 2014. IEEE Computer Society. ISBN 9781479951185. doi: 10.1109/CVPR.2014.461. URL https://doi.org/10.1109/CVPR.2014.461.

Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Hendrycks, D., Mazeika, M., Kadavath, S., and Song, D. Using self-supervised learning can improve model robustness and uncertainty. In *NeurIPS*, 2019a.

Hendrycks, D., Mazeika, M., Kadavath, S., and Song, D. Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems*, pp. 15663–15674, 2019b.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.

Kim, M., Tack, J., and Hwang, S. J. Adversarial self-supervised contrastive learning, 2020.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30:6402–6413, 2017.

Lee, K., Lee, H., Lee, K., and Shin, J. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Liu, W., Wang, X., Owens, J., and Li, Y. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33, 2020a.

Liu, W., Wang, X., Owens, J. D., and Li, Y. Energy-based out-of-distribution detection. *ArXiv*, abs/2010.03759, 2020b.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Misra, I. and van der Maaten, L. Self-supervised learning of pretext-invariant representations, 2019.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. URL http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

Noroozi, M. and Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles, 2017.

van den Oord, A., Kalchbrenner, N., Espeholt, L., kavukcuoglu, k., Vinyals, O., and Graves, A. Conditional image generation with pixelcnn decoders. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 4790–4798. Curran Associates, Inc., 2016.

Yu, F., Zhang, Y., Song, S., Seff, A., and Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. 06 2015.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018.